*Article*

# SRT: A Spectral Reconstruction Network for GF-1 PMS Data Based on Transformer and ResNet

Kai Mu [1,2,3,†], Ziyuan Zhang [1,2,3,†], Yurong Qian [1,2,3,*], Suhong Liu [4], Mengting Sun [5] and Ranran Qi [1,2,3]

1  The School of Software, Xinjiang University, Urumqi 830091, China; mukai@stu.xju.edu.cn (K.M.);
   zhangziyuan@stu.xju.edu.cn (Z.Z.); qiranran@stu.xju.edu.cn (R.Q.)
2  The Key Laboratory of Signal Detection and Processing in Xinjiang Uygur Autonomous Region,
   Xinjiang University, Urumqi 830091, China
3  The Key Laboratory of Software Engineering, Xinjiang University, Urumqi 830091, China
4  The Department of Geography, Beijing Normal University, Beijing 100875, China; liush@bnu.edu.cn
5  The School of Geography and Remote Sensing Science, Xinjiang University, Urumqi 830049, China;
   smt@stu.xju.edu.cn
*  Correspondence: qyr@xju.edu.cn
†  These authors contributed equally to this work.

**Abstract:** The time of acquiring remote sensing data was halved after the joint operation of Gao Fen-6 (GF-6) and Gao Fen-1 (GF-1) satellites. Meanwhile, GF-6 added four bands, including the "red-edge" band that can effectively reflect the unique spectral characteristics of crops. However, GF-1 data do not contain these bands, which greatly limits their application to crop-related joint monitoring. In this paper, we propose a spectral reconstruction network (SRT) based on Transformer and ResNet to reconstruct the missing bands of GF-1. SRT is composed of three modules: (1) The transformer feature extraction module (TFEM) fully extracts the correlation features between spectra. (2) The residual dense module (RDM) reconstructs local features and avoids the vanishing gradient problem. (3) The residual global construction module (RGM) reconstructs global features and preserves texture details. Compared with competing methods, such as AWAN, HRNet, HSCNN-D, and M2HNet, the proposed method proved to have higher accuracy by a margin of the mean relative absolute error (MRAE) and root mean squared error (RMSE) of 0.022 and 0.009, respectively. It also achieved the best accuracy in supervised classification based on support vector machine (SVM) and spectral angle mapper (SAM).

**Keywords:** spectral reconstruction; GF-1; GF-6; Transformer; remote sensing; deep learning

## 1. Introduction

The GF-6 was successfully launched in 2018 as China's first medium-high-resolution agricultural observation satellite, which cooperated with GF-1, China's first high-resolution earth observation satellite that was launched in 2013. It can not only reduce the time of remote sensing data acquisition from 4 days to 2, but also significantly improve the ability to monitor agriculture, forestry, grassland, and other resources, providing remote sensing data support for agricultural and rural development, ecological civilization construction [1], and other significant needs. GF-6 also realized the localization of the 8-band CMOS detector and added the red-edge band that can effectively reflect the unique spectral characteristics of crops [2,3].

However, GF-1 was launched earlier and is mission-oriented differently, so it only contains four multispectral bands. Compared with the GF-6 satellite in Table 1, GF-1 lacks four bands (purple, yellow, red-edge I, and red-edge II bands), which greatly constrains its development for crop-related joint monitoring. So, we try to find a spectral reconstruction method to reconstruct the lacking four bands.

In recent years, spectral reconstruction mainly focused on RGB or multispectral to hyperspectral. Earlier researchers adopted the sparse dictionary method [4–9]. With the

development of deep learning, owing to its excellent feature extraction and reconstruction capabilities, more and more researchers are adopting deep learning methods to gradually replace the traditional sparse dictionary approach [10–16].

In addition, it should be pointed out that most studies on spectral reconstruction focus on visible three bands (red, green, and blue) images, while remote sensing images usually contain at least four bands (red, green, blue, and nir). This results in the lack of one essential nir band as the input, which does not make full use of the original information, thereby leading to a waste of information. There are already some studies of remote sensing spectral reconstruction considering this problem [15,16]. Few studies have been conducted on large-scale and highly complex scenarios such as satellite remote sensing. On the contrary, most of them have only done performed research in a relatively small area [15]. Most deep learning methods adopt a lot of up-sampling, down-sampling, and nonlocal attention structure for ground images. Due to the large-scale, numerous, and complex ground objects of remote sensing images, these structures are difficult to play an excellent effect in the spectral reconstruction of remote sensing images [16].

To better adapt to the spectral reconstruction of remote sensing images, we propose a more suitable spectral reconstruction network (SRT) for GF-1 panchromatic and multi-spectral sensor (PMS) data based on Transformer and ResNet. This network includes a TFEM, the RDM, and the RGM. The first module contributes to the extraction of correlation characteristics between spectra. To avoid the vanishing gradient problem, the second module reconstructs these features nonlinearly at the local features. The third module, mainly used for the global reconstruction of these features, prevents loss of texture details. The main contributions of this article are summarized as follows:

- We propose a spectral reconstruction network. The network trains on GF-6 wide field view (WFV) images to reconstruct the four lacking bands of GF-1 PMS images, which significantly increases the classification capability of GF-1.
- We produce a large-scale dataset that covers a wide area and is rich in land types. It basically meets the ground object information required for spectral reconstruction.
- In order to evaluate the generalization ability of our model, we compare it with other models in image similarity and classification accuracy, and conclude that our model has the best result.

**Table 1.** Band specification of the GF-1 PMS and GF-6 WFV images.

| GF-1 PMS | | | GF-6 WFV | | |
|---|---|---|---|---|---|
| **Band** | **Wavelength (nm)** | **Spatial Resolution (m)** | **Band** | **Wavelength (nm)** | **Spatial Resolution (m)** |
| Blue | 450~520 | 8 | Blue | 450~520 | 16 |
| Green | 520~590 | 8 | Green | 520~590 | 16 |
| Red | 630~690 | 8 | Red | 630~690 | 16 |
| Nir | 730~890 | 8 | Nir | 730~890 | 16 |
| | | | Red edge 1 | 690~730 | 16 |
| | | | Red edge 1 | 730~770 | 16 |
| | | | Purple | 400~450 | 16 |
| | | | Yellow | 590~640 | 16 |
| Pan | 450~900 | 2 | | | |

The remaining part of this article is organized as follows: Section 2 describes the related works of spectral reconstruction methods. We present the network of SRT in Section 3. Section 4 presents our results, including the dataset description, the experimental part, and its analysis. Section 5 is the conclusion.

## 2. Related Works

Due to the limitations of the hardware resources (bandwidth and sensors), researchers have had to make trade-offs in the temporal, spatial, and spectral dimensions of remote sensing images. With the problem of low spectral dimension, researchers mainly used prin-

cipal component analysis (PCA) [17,18], Wiener estimation (WEN) [19], and pseudoinverse (PI) [20,21] to construct a spectral mapping matrix. In recent years, spectral reconstruction methods have been divided into two branches: prior-driven and data-driven methods.

The first type is mainly based on sparse dictionary learning, which aims to extract the most important spectral mapping features. It can represent as much knowledge as possible with as few resources as possible, and this representation has the added benefit of being computationally fast. For example, Arad and Ben-Shahar [4] were the first to apply an overcomplete dictionary to recover hyperspectral images from RGB. Jonas et al. [5] used the A+ algorithm to improve Arad's approach to the sparse dictionary. The A+ algorithm directly constructs the mapping from RGB to hyperspectral at the local anchor point, and the running speed of the algorithm is significantly improved. The sparse dictionary method only considers the sparsity of spectral information and does not use local linearity. The disadvantage is that the reconstruction is inaccurate, and the reconstructed image has metamerism [22]. Li et al. [7] proposed a locally linear embedding sparse dictionary method to improve the representation ability of sparse coding. In order to improve the representation ability of the sparse dictionary, this method only selects the local best samples and introduces texture information in the reconstruction, reducing the metamerism. Geng et al. [8] proposed a spectral reconstruction method that preserves contextual information. Gao et al. [9] performed spectral enhancement of multispectral images by jointly learning low-rank dictionary pairs from overlapping regions.

The second type is mainly based on deep learning. With the development of deep learning, a large number of excellent models have gradually replaced the first method owing to its powerful generalization ability. However, compared to the first one, deep learning usually requires enormous amounts of data, and the training process takes a lot of computational time. However, with the increase in computing power, deep learning becomes much more effective, and the related methods are used by more and more researchers. Xiong et al. [10] proposed a deep learning framework for recovering spectral information from spectrally undersampled images. Koundinya et al. [12] compared 2D and 3D kernel-based CNN for spectral reconstruction. Alvarez-Gila et al. [11] posed spectral reconstruction as an image-to-image mapping problem and proposed a generative adversarial networks for spatial context-aware spectral image reconstruction. In the NTIRE 2018 [23] first spectral reconstruction challenge, the entries of Shi et al. [13] ranked in first (HSCNN-D) and second (HSCNN-R) place on both the "Clean" and "Real World" tracks. The main difference between the two networks is that the former adopts a series method for feature fusion, while the latter is an addition method. The series method can learn the mapping relationship between spectra very well. Respectively considering shallow feature extraction and deep feature extraction, Li et al. [24] proposed an adaptive weighted attention network, which obtained the first rank on the "Clean" track. Zhao et al. [14] proposed a hierarchical regression network (HRNet) that obtained first place on the "Real World" track; it is a 4-level multi-scale structure that uses down-sampling and up-sampling to extract spectral features. In the processing of remote sensing images, Deng et al. [15] proposed a more suitable network (M2H-Net) for remote sensing to meet the needs of multiple bands and complex scenes. Li and Gu [16] proposed a progressive spatial–spectral joint network for hyperspectral image reconstruction.

## 3. Proposed Method

### 3.1. SRT Architecture

Figure 1 shows the architecture of SRT. In training, the model inputs red, blue, green, and nir bands of GF-6 WFV, and the remaining purple, yellow, red-edge I, and red-edge II bands are used as labels. The overall structure includes TFEM, RDM, RGM, convolution operations, and other related operations.
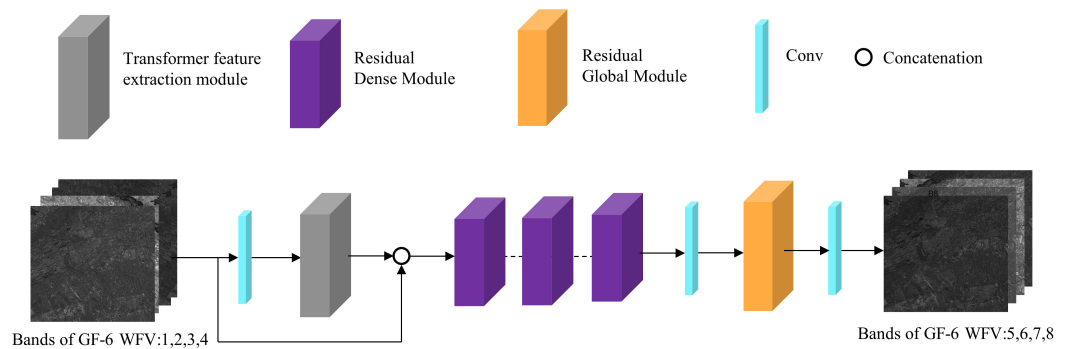
**Figure 1.** Network architecture of our SRT.

The whole SRT is an end-to-end structure, which can be divided into three parts:

1. The TFEM is used to extract correlation between spectra by self-attention mechanism.
2. The RDM, which can fully learn and reconstruct these local features to prevent gradient vanishing in training.
3. The RGM is able to reconstruct these global features. Considering the model is ultimately used for GF-1 PMS (8 m) images, it doubles the spatial resolution compared to the trained GF-6 WFV (16 m) images. This module can prevent losing the texture details in the training or inference process.

### 3.2. TFEM

Google first proposed the Transformer architecture in June 2017 [25]. The impact on the whole natural language processing (NLP) field has been tremendous. In just four years since it was proposed, Transformer has become the dominant model in NLP [26]. Since 2020, it has started to shine in the field of computer vision (CV): image classification (ViT [27], DeiT [28]), object detection (DETR [29], Deformable DETR [30]), semantic segmentation (SETR [31], MedT [32]), image generation (GANsformer [33]) and so on. He et al. [34] showed scalable self-supervised learners for CV (masked autoencoders, MAE). Once again, Transformer shined in the CV. Inspired by the development of Transformer, we try to use Transformer as the backbone of feature extraction for SRT to fully extract relevant features between spectra with the help of its effective attention mechanism. The architecture of TFEM is shown in Figure 2.
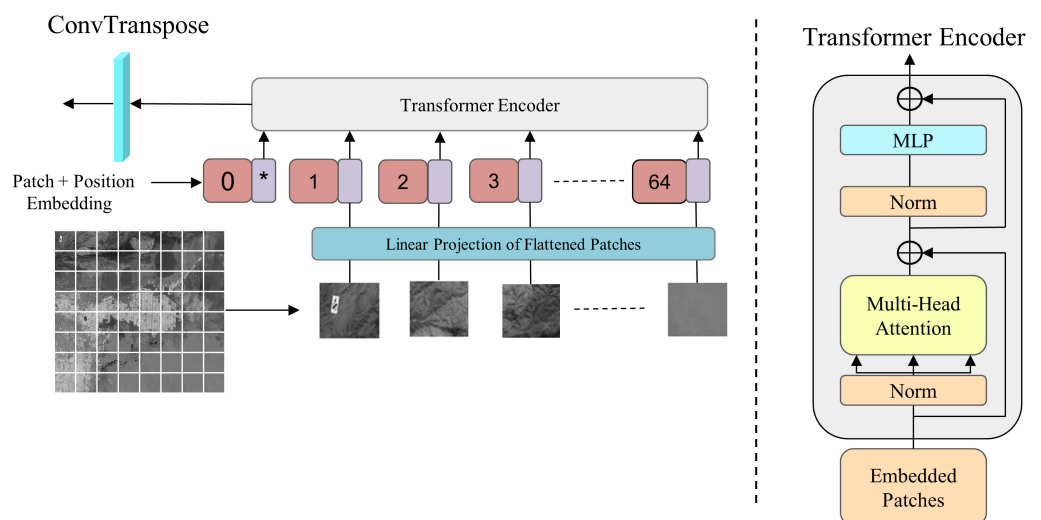


**Figure 2.** The architecture of TFEM, where the purple blocks are obtained vectors from each linear projection of the flattened patch, and the red blocks are the learnable positional encodings of the corresponding vectors.

Following ViT [27], we divide the remote sensing images into multiple small patches and serialize each patch through a linear projection of flattened patches so that a vision problem turns into a NLP problem. The module needs to add learnable position embedding parameters to maintain the spatial location information between the input patches. The Transformer encoder extracts spectral features from input sequences with the help of its multi-attention mechanism. In our experiment, considering Transformer is only used for feature extraction; we remove the learnable classification embedded in the ViT and use ConvTranspose to replace the MLP head to ensure that the model maps to the same dimension.

### 3.3. RDM

He et al. [35] proposed a residual learning framework (ResNet) to ease the training of networks that are substantially deeper than those used previously. Based on ResNet, DenseNet makes each layer connect to all previous layers, it [36] is a new network framework that enriches the CNN network system from LeNet [37] to the present ones. It connects all layers to ensure maximum exchange of spectral information flow in the network. In addition, DenseNet also has the advantage that it requires fewer parameters for the same performance or the same number of layers. This is because it has a direct connection to all the previous layers, so it does not have to relearn some of the features that have already been learned.

The RDM contains four residual dense blocks which is shown in Figure 3, and a long skip connection is added in front of the module to prevent the vanishing gradient problem in the network. The spectral reconstruction model of the residual network and dense network can alleviate the vanishing gradient problem during training and ensure more accurate results.
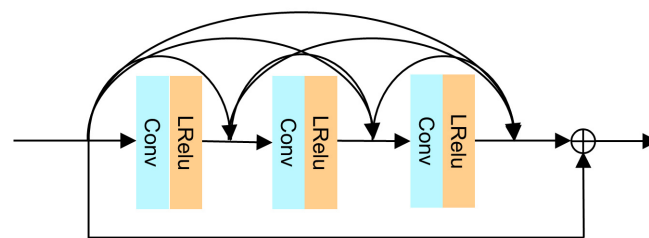


**Figure 3.** The architecture of residual dense block, where LRelu is leakyRelu.

### 3.4. RGM

The RGM references SE-ResNet [38] and HRNet [14] which is shown in Figure 4. Average pooling can bias the features of the image toward the overall characteristics and prevent the loss of too much high-dimensional information. The final convolution layer is used for channel number mapping, and the global residual is used to preserve spatial details in the image of different spatial resolutions.
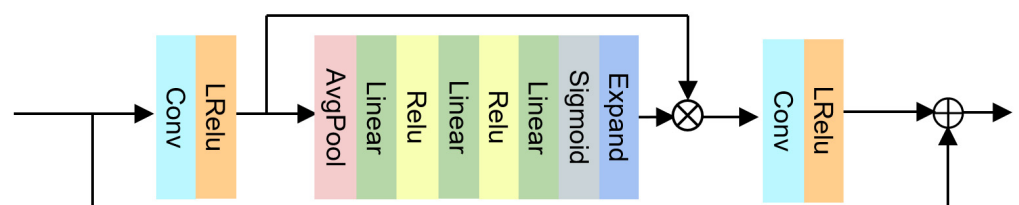


**Figure 4.** The architecture of RGM.

### 3.5. Loss Function

We use the mean relative absolute error (MRAE, Equation (1)) as the loss function, due to the reflectance of the same object on the ground, varies greatly in different bands.

It replaces the absolute difference of the mean square error (MSE, Equation (2)), with the mean relative absolute error to achieve adaptive error adjustment according to each band. In a way, it can effectively reduce the high errors caused by different reflectance and demonstrate the accuracy of the reconstructed network more visually. In the validation set, we measure the metric of the models by peak signal-to-noise ratio (PSNR [39], Equation (3)), and save the best model.

$$\text{MRAE} = \frac{1}{n} \sum_{i=1}^{n} \frac{\left| P_{\text{gt}_i} - P_{\text{rec}_i} \right|}{P_{\text{gt}_i}} \tag{1}$$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left( P_{\text{gt}_i} - P_{\text{rec}_i} \right)^2 \tag{2}$$

where $P_{\text{gt}_i}$ is the gray-scale value of the $i$th pixel in the reference image, $P_{\text{rec}_i}$ is the reconstructed gray-scale value of the $i$th pixel, and $n$ is the number of pixels in the image.

$$\text{PSNR} = 20 \cdot \log_{10} \left( \frac{MAX_I}{\sqrt{\text{MSE}}} \right) \tag{3}$$

where $MAX_I$ is the maximum value of the gray-scale value. All data in this experiment is normalized, $MAX_I$ is 1.

### 3.6. Network Training and Parameter Settings

The parameters of the Transformer encoder are set by default, and the network hyper-parameters are set according to Table 2. The size of each convolution kernel in the network is $3 \times 3$. For the optimizer, we choose Adam.

**Table 2.** Hyper-parameters setting of SRT network.

| Parameter Name | Parameter Setting |
| --- | --- |
| Batch size | 32 |
| Initial learning rate | 0.01 |
| Optimizer | Adam |
| Decay rate | 0.1 |
| Learning rate decay steps | 2000 steps |
| Epochs | 200 |
| | Relu |
| Activation function | Sigmod |
| | Leaky-Relu |

The computer configuration in this study: CPU is Intel (R) Xeon (R) Gold 6148, GPU is Telsa V100 16 G, and RAM is 16 G. Paddle2.2 was chosen as the development environment.

## 4. Experiments

The experiment evaluates the quality of the spectral reconstruction by accuracy and classification. Furthermore, AWAN, HRNet, HSCNN-D, and remote sensing image reconstruction (M2H-Net) are the four outstanding methods that are selected to compare with our model SRT, SRT*, and the former three are spectral reconstruction champion methods in the NTIRE challenge. SRT* removes the RGM compared to SRT to test the effect of the module.

### 4.1. Dataset Description

We use image scenes from GF-1 PMS and GF-6 WFV. The data acquisition for the study areas is shown in Figure 5.
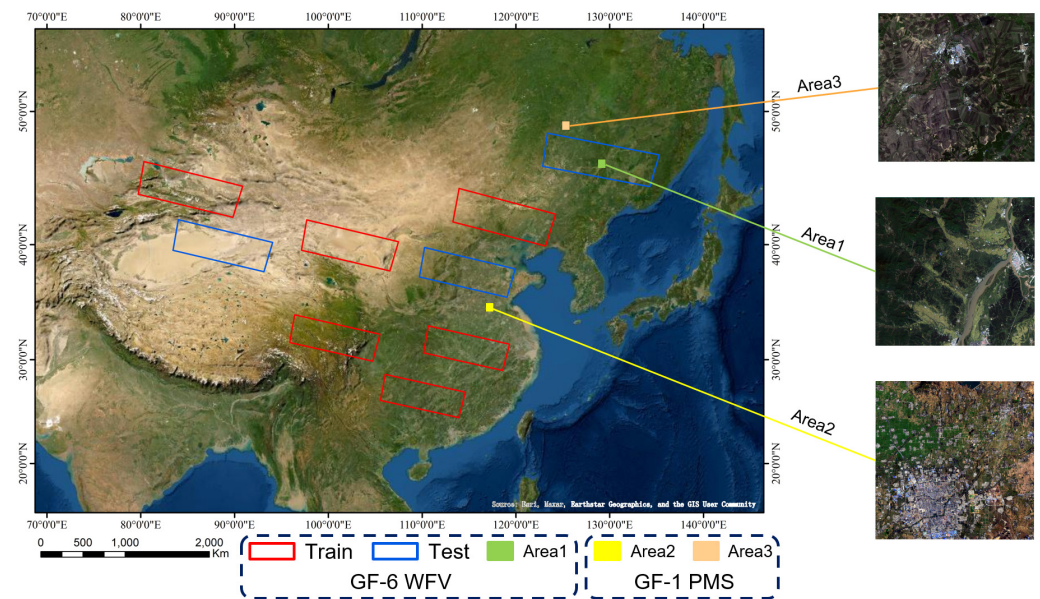
**Figure 5.** Image regions used for training and testing from China Center For Resources Satellite Data and Application website [40].

We select nine GF-6 WFV images to form the dataset, six for training and three for testing. The dataset covers a wide range of land types and provides sufficient feature information for the spectral reconstruction of GF-1 PMS. We randomly divide the training images into 13,500 overlapping patches of 128 × 128 pixels, 90% of them for training and the rest for validation. The testing ones are divided into 2000 overlapping patches of 128 × 128 pixels.

The image shown in Area1 is the Songhua River, located in Yilan, Heilongjiang. It is a cropped GF-6 WFV test image that contains abundant information on water, vegetation, tree, and so on. The size of it is 2275 × 2174.

Area2, imaged by GF-1 on 11 Aprill 2016, is located in Tengzhou, Shandong, and contains ample information on building, vegetation, and road. The size of its image is 2500 × 2322.

Area3, imaged by GF-1 on 21 June 2018, is located in Nenjiang, Heilongjiang, and contains rich vegetation, bare land, and tree. The size of its image is 3254 × 3145.

The preprocessing of GF-1 PMS and GF-6 WFV images includes radiometric correction and atmospheric correction in ENVI 5.3. The parameters for the correlation correction are obtained from China Resource Satellite Application Center [40].

Table 3 lists the detailed number of pixels of the training and testing samples for classification in the three areas. Each of them is manually annotated into six classes in ENVI 5.3 software (Exelis Inc., Boulder, CO, USA) to test the classification ability of the reconstructed images, as is shown in Figure 6.

**Table 3.** Details of the ground truth in Area1–3.

|  | Area1 | | Area2 | | Area3 | |
|---|---|---|---|---|---|---|
|  | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| Water | 6381 | 2451 | 3992 | 2661 | 1921 | 3487 |
| Build | 738 | 392 | 3008 | 2005 | 1219 | 1598 |
| Bare land | 101 | 99 | 4306 | 2870 | 1397 | 1491 |
| Plant | 2431 | 1273 | 4169 | 2780 | 7204 | 8551 |
| Tree | 10371 | 8361 | 1691 | 1127 | 6518 | 3131 |
| Road | 87 | 146 | 1592 | 1062 | 300 | 259 |
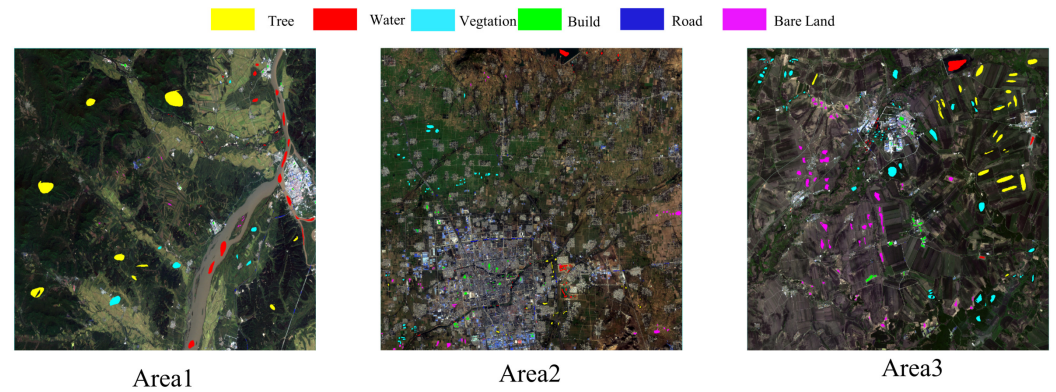
**Figure 6.** Distribution of the selected sample objects in Area1–3.

*4.2. Evaluation Metrics*

We use five indicators to evaluate the different methods, including RMSE, MRAE (Equation (1)), PSNR (Equation (3)), spectral angle mapper (SAM [41]), and structural similarity (SSIM [42]). The formulas of RMSE, SAM, and SSIM are given as follows:

$$\text{RMSE} = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n}\left(P_{\text{gt}_i} - P_{\text{rec}_i}\right)^2}{N}} \tag{4}$$

$$\text{SAM} = \cos^{-1}\frac{P_{\text{gt}}^T P_{\text{rec}}}{\left(P_{\text{gt}}^T P_{\text{gt}}\right)^{1/2}\left(P_{\text{rec}}^T P_{\text{rec}}\right)^{1/2}} \tag{5}$$

where $P_{\text{gt}_i}$ is the gray-scale value of the $i$th pixel in the reference image, $P_{\text{rec}_c}$ is the reconstructed gray-scale value of the $i$th pixel, and $n$ is the number of pixels in the image.

$$\text{SSIM}(\text{gt}, \text{rec}) = \frac{\left(2\mu_{\text{gt}}\mu_{\text{rec}} + C_1\right)\left(2\sigma_{\text{gtrec}} + C_2\right)}{\left(\mu_{\text{gt}}^2 + \mu_{\text{rec}}^2 + C_1\right)\left(\sigma_{\text{gt}}^2 + \sigma_{\text{rec}}^2 + C_2\right)} \tag{6}$$

where $\mu_{\text{gt}}$ is the average value of the reference image, $\mu_{\text{rec}}$ is the average value of the reconstructed image, $\sigma_{\text{gtrec}}$ is the covariance of the reference image and the reconstructed image, $\sigma_{\text{gt}}$ is the standard deviation of the reference image, $\sigma_{\text{rec}}$ is the standard deviation of the reconstructed image, and $C_1 = (k_1 L)^2$ and $C_2 = (k_2 L)^2$ are constants used to maintain stability. $L$ is the dynamic range of the pixel values and $k_1$ is set to 0.01 and $k_2$ to 0.03.

Classification is an essential application of remote sensing images, and we use SVM and SAM classification to test the classification performance of images. SVM can solve linear and non-linear classification problems well, with fewer support vectors to determine the classification surface, and is not sensitive to the number of samples and spectral dimensionality. SAM measures the similarity between spectra by treating both spectra as vectors and calculating the spectral angle between them. Therefore, it is sensitive to samples and spectral dimensionality.

For the testing of GF-1 PMS images, we cannot use the above indicators to evaluate the four generated bands, except for the classification accuracy. The assessment steps include the following: First, input the original image to the model after radiometric calibration and atmospheric correction. Then, classify the outputs by SVM and SAM methods. Finally, compare the overall accuracy (OA), kappa coefficient (Kappa), and accuracy for every class of all the methods with each other.

*4.3. Similarity-Based Evaluation*

Table 4 shows the accuracy assessment of the reconstructed GF-6 WFV images on the dataset. Overall, the PSNR and SSIM of the four bands are all high, not less than 38.92 and

0.970, respectively. Similarly, MARE, SAM, and RMSE are all relatively low, indicating that the overall accuracy of the reconstruction is high.

**Table 4.** Quantitative assessment of different spectral reconstruction methods for the dataset. The best results are shown in bold.

| Evaluation | Band | Method | | | | | |
|---|---|---|---|---|---|---|---|
| | | AWAN | HSCNN-D | HRNet | M2Hnet | SRT | SRT* |
| PSNR | band5 | 40.06 | 40.10 | 43.92 | 40.80 | **45.23** | 44.91 |
| | band6 | 39.37 | 39.47 | 43.00 | 38.92 | **44.00** | 43.39 |
| | band7 | 42.51 | 41.71 | 46.56 | 42.50 | **48.29** | 47.81 |
| | band8 | 41.23 | 42.10 | 44.71 | 40.95 | **45.87** | 45.63 |
| | avg | 40.79 | 40.85 | 44.55 | 40.79 | **45.85** | 45.43 |
| SSIM | band5 | 0.980 | 0.972 | **0.991** | 0.979 | **0.991** | **0.991** |
| | band6 | 0.970 | 0.983 | **0.992** | 0.976 | 0.991 | 0.990 |
| | band7 | 0.980 | 0.985 | 0.990 | 0.973 | **0.992** | **0.992** |
| | band8 | 0.970 | 0.981 | 0.990 | 0.978 | **0.992** | **0.992** |
| | avg | 0.975 | 0.980 | 0.991 | 0.976 | **0.992** | 0.991 |
| MRAE | band5 | 0.038 | 0.041 | 0.024 | 0.038 | **0.021** | 0.024 |
| | band6 | 0.026 | 0.024 | **0.019** | 0.037 | 0.020 | 0.020 |
| | band7 | 0.031 | 0.032 | 0.022 | 0.036 | **0.019** | 0.025 |
| | band8 | 0.032 | 0.029 | 0.021 | 0.038 | **0.020** | 0.022 |
| | avg | 0.032 | 0.032 | 0.022 | 0.037 | **0.020** | 0.023 |
| SAM | band5 | 1.66 | 1.59 | 1.17 | 1.68 | **1.06** | 1.07 |
| | band6 | 1.21 | 1.23 | 0.87 | 1.36 | **0.81** | 0.84 |
| | band7 | 1.42 | 1.43 | 0.88 | 1.51 | **0.80** | 0.81 |
| | band8 | 1.66 | 1.50 | 1.08 | 1.72 | **1.00** | 1.02 |
| | avg | 1.49 | 1.44 | 1.00 | 1.57 | **0.92** | 0.93 |
| RMSE | band5 | 0.010 | 0.012 | 0.008 | 0.016 | **0.007** | 0.010 |
| | band6 | 0.015 | 0.021 | **0.007** | 0.011 | 0.009 | 0.014 |
| | band7 | 0.009 | 0.014 | 0.009 | **0.004** | 0.006 | 0.008 |
| | band8 | 0.010 | 0.013 | **0.007** | 0.016 | **0.007** | 0.010 |
| | avg | 0.011 | 0.015 | **0.008** | 0.012 | **0.008** | 0.011 |

Among the six methods, the results of the AWAN, HSCNN-D, and M2HNet methods are similar. HRNet, SRT, and SRT* are much better than the other three methods in PSNR, MRAE, and SAM. The SRT outperforms HRNet on the dataset, demonstrating that our TFEM outperforms the multi-scale feature extraction of HRNet. In addition, SRT* lacks the RGM compared to SRT and is slightly worse than SRT in some indicators, but still has some advantages compared to other methods.

Compared with the scatter plot in Figure 7, it turns out that the inference results of bands 5 and 6 have larger areas of scattering compared to bands 7 and 8, which indicates that the reconstruction is less relevant. It is also reflected by the PSNR metric on Table 4. The larger the PSNR is, the smaller the scattering region and the strongest the correlation between the predicted band and the original one. The PSNR of band 7 in Table 4 is the highest, and the scatter region of band 7 in Figure 7 is the smallest. Therefore, we can conclude that the reconstruction accuracy of band 7 is the best. It can be seen from the scatter plot that the reconstruction accuracy of each band is different. Using MRAE as the loss function compared to RMSE can well avoid the training of the band-dominant model with large errors.
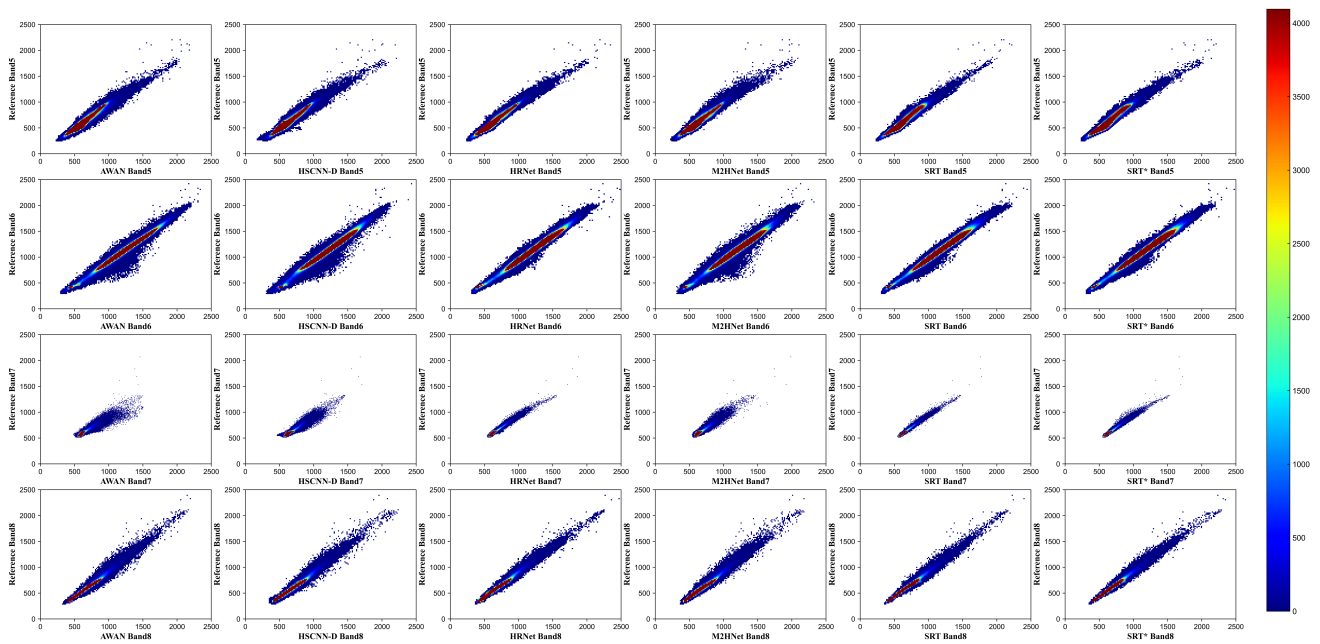
**Figure 7.** The scatter plot shows the predicted bands of each method compared with the original bands.

### 4.4. Classification-Based Evaluation

For GF-6 WFV images, we evaluate the confusion matrix by the classification results of the original image and the predicted one. Table 5 shows the evaluation results of the SVM classification. Among them, both the OA and KAPPA coefficients of SRT are the highest, 3.3% and 4.2% higher than AWAN, respectively. In the classification result of vegetation, the SRT classification result is 6.3% higher than the second-highest M2HNet. In Figure 8, we can see that the water classification result of M2HNet is significantly different from the reference image.

**Table 5.** Accuracy of classification result of Area1 with SVM. The best results are shown in bold.

| SVM | AWAN | HSCNN-D | HRNet | M2HNet | SRT | SRT* | GF-6 |
|---|---|---|---|---|---|---|---|
| OA | 0.8909 | 0.9015 | 0.9030 | 0.9038 | **0.9237** | 0.9118 | 0.9291 |
| Kappa | 0.7900 | 0.8071 | 0.8106 | 0.8133 | **0.8321** | 0.8215 | 0.8357 |
| Water | 0.9560 | 0.8272 | 0.9733 | 0.9786 | 0.8324 | **0.9813** | 0.9847 |
| Build | 0.9217 | 0.8918 | 0.9849 | 0.9295 | 0.9777 | **0.9894** | 0.9817 |
| Bare Land | 0.5719 | 0.7455 | 0.5855 | 0.5804 | 0.5035 | **0.6646** | 0.6654 |
| Vegetation | 0.8803 | 0.8836 | 0.8836 | 0.8506 | **0.9140** | 0.8871 | 0.9262 |
| Tree | 0.8812 | 0.8196 | 0.8909 | 0.8983 | 0.8657 | **0.9001** | 0.9200 |
| Road | 0.5756 | 0.5857 | 0.5823 | 0.5785 | 0.4357 | **0.5872** | 0.5768 |
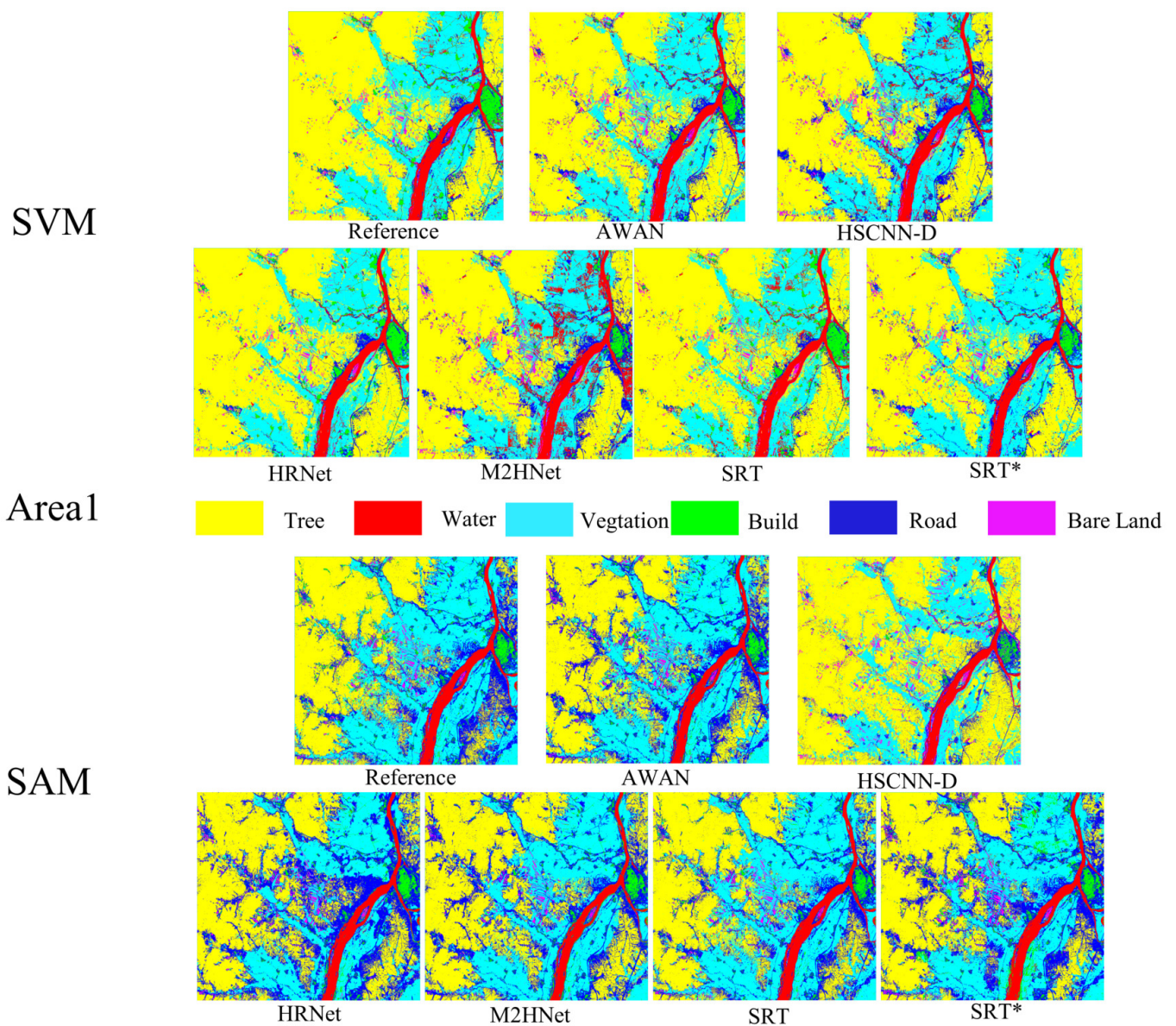
**Figure 8.** Result of SVM and SAM classification on Area1.

Table 6 shows the evaluation results of SAM classification, and the SRT results are still the best. Its errors in the OA and Kappa coefficients with the original image classification are only 0.5% and 0.24%. It indicates that the spectral reconstruction capability of SRT is optimal among other methods.

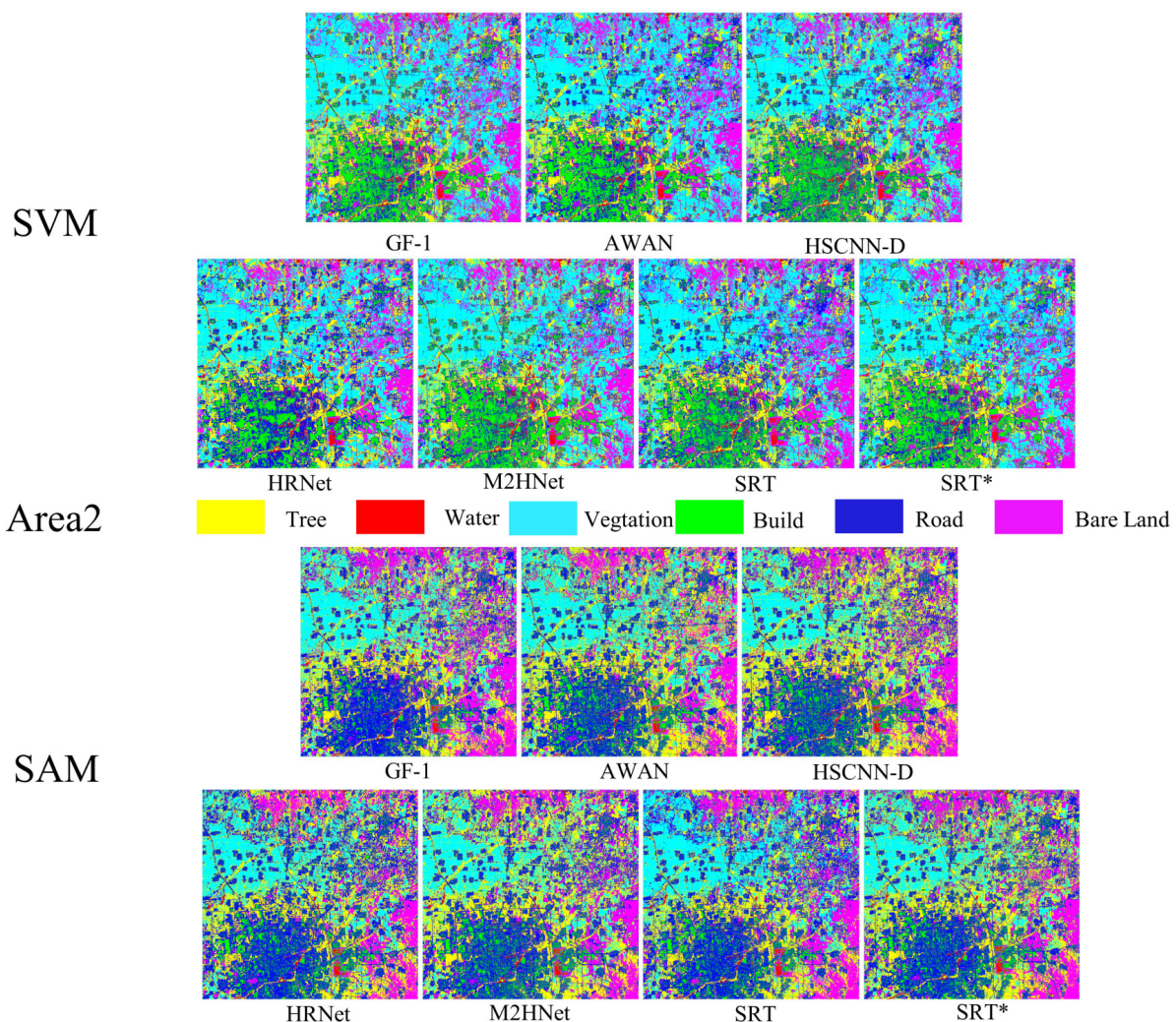**Table 6.** Accuracy of classification result of Area1 with SAM. The best results are shown in bold.

| SAM | AWAN | HSCNN-D | HRNet | M2HNet | SRT | SRT* | GF-6 |
|---|---|---|---|---|---|---|---|
| OA | 0.8743 | 0.8709 | 0.8782 | 0.8701 | **0.8939** | 0.8854 | 0.8992 |
| Kappa | 0.7435 | 0.7409 | 0.7513 | 0.7401 | **0.8012** | 0.7821 | 0.8036 |
| Water | **0.8970** | 0.8683 | 0.8324 | 0.8498 | 0.8823 | 0.8849 | 0.8500 |
| Build | 0.5148 | **0.6386** | 0.5077 | 0.5339 | 0.4636 | 0.4851 | 0.4715 |
| Bare Land | 0.7121 | **0.7490** | 0.5035 | 0.6606 | 0.5539 | 0.5746 | 0.8823 |
| Vegetation | **0.9518** | 0.9429 | 0.914 | 0.9492 | 0.9202 | 0.9411 | 0.9278 |
| Tree | 0.8799 | 0.8774 | 0.9157 | 0.8862 | **0.9258** | 0.9028 | 0.9331 |
| Road | 0.5703 | 0.6183 | 0.4357 | 0.6449 | 0.4196 | **0.6977** | 0.6976 |

For GF-1 PMS images, our classification results should be higher than GF-1 (8 m spatial resolution, four bands). Table 7 shows the accuracy metrics for SVM classification in Area2. Most methods improve the classification evaluation metrics, with SRT improving OA and KAPPA by 2.1% and 4.3%, respectively. Except in the two classes of tree and road, the classification accuracy of SRT is higher than the original GF-1 PMS for other classes.

**Table 7.** Accuracy of classification result of Area2 with SVM. The best results are shown in bold.

| SVM | AWAN | HSCNN-D | HRNet | M2Hnet | SRT | SRT* | GF-1 |
|---|---|---|---|---|---|---|---|
| OA | 0.8662 | 0.8749 | 0.8839 | 0.8788 | **0.8862** | 0.8853 | 0.8648 |
| Kappa | 0.8309 | 0.8359 | 0.8507 | 0.8399 | **0.8655** | 0.8548 | 0.8223 |
| Water | 0.9754 | 0.9808 | 0.9854 | 0.9880 | **0.9881** | 0.9844 | 0.9775 |
| Build | 0.7527 | 0.7638 | **0.8299** | 0.8045 | 0.7590 | 0.7514 | 0.7519 |
| Bare Land | 0.8378 | 0.8879 | 0.8492 | 0.8600 | **0.9398** | 0.9395 | 0.8527 |
| Vegetation | **0.9614** | 0.9505 | 0.9535 | 0.9555 | 0.9543 | 0.9573 | 0.9531 |
| Tree | 0.8279 | 0.7876 | **0.8358** | 0.8227 | 0.7940 | 0.7907 | 0.8370 |
| Road | 0.6650 | 0.6784 | **0.6947** | 0.6547 | 0.6458 | 0.6558 | 0.6264 |

Table 8 shows the evaluation results for the SAM classification, where all the methods are still higher than the original results, and the SRT method is the best. Additionally, the results in Figure 9 show that the accuracy of the SVM is higher than the SAM, especially for urban scenes.



**Figure 9.** Result of SVM and SAM classification on Area2.

**Table 8.** Accuracy of classification result of Area2 with SAM. The best results are shown in bold.

| SVM | AWAN | HSCNN-D | HRNet | M2Hnet | SRT | SRT* | GF-1 |
|---|---|---|---|---|---|---|---|
| OA | 0.8046 | 0.7954 | 0.8047 | 0.8058 | **0.8196** | 0.8066 | 0.7923 |
| Kappa | 0.7920 | 0.7836 | 0.7921 | 0.7947 | **0.8048** | 0.7956 | 0.7822 |
| Water | 0.9973 | 0.9972 | 0.9907 | 0.9972 | **0.9997** | 0.9990 | 0.9988 |
| Build | 0.9087 | 0.9104 | 0.9087 | **0.9381** | 0.8822 | 0.9140 | 0.9015 |
| Bare Land | 0.4623 | 0.4403 | 0.4625 | 0.4288 | **0.4812** | 0.4611 | 0.4233 |
| Vegetation | 0.8059 | 0.7872 | 0.8059 | 0.7959 | **0.8418** | 0.8062 | 0.7775 |
| Tree | 0.8610 | 0.8726 | 0.8610 | 0.9264 | **0.9429** | 0.8737 | 0.9160 |
| Road | 0.9874 | **0.9886** | 0.9874 | 0.9875 | 0.9779 | 0.9852 | 0.9776 |

Table 9 shows the classification accuracy of Area3. Compared to the GF-1 image classification results, it can improve the OA and Kappa of SRT by 2.41% and 2.0%, respectively. Most classes' accuracies are better than before. Except for water and bare land, the classification accuracy of SRT is higher than that of other methods for other classes. As shown in Table 10, SRT remains the highest. However, the SAM classification accuracy of all methods in Area3 is much lower than that of SVM. The original image's OA and Kappa coefficients of the SAM classification are lower than the SVM, with differences as high as 8.8% and 16.7%, respectively. From Figure 10, it also can be seen that the difference between SVM and SAM results classification. SAM classification does not classify the build area well, it divides a small part of bare land into water and divides bare land into two lots of tree. This vast difference may result from the lower spectral dimensions, while the SAM method is more sensitive to the spectrum, so the classification accuracy of SAM is lower than before.

**Table 9.** Accuracy of classification result of Area3 with SVM. The best results are shown in bold.

| SVM | AWAN | HSCNN-D | HRNet | M2Hnet | SRT | SRT* | GF-1 |
|---|---|---|---|---|---|---|---|
| OA | 0.9303 | 0.9212 | 0.9327 | 0.9322 | **0.9487** | 0.9406 | 0.9246 |
| Kappa | 0.9258 | 0.9127 | 0.9302 | 0.9190 | **0.9357** | 0.9348 | 0.9157 |
| Water | 0.991 | 0.9854 | **0.9897** | 0.9888 | 0.9880 | 0.9853 | 0.9931 |
| Build | 0.6304 | 0.6299 | 0.5710 | 0.5933 | **0.6650** | 0.5944 | 0.4950 |
| Bare Land | 0.9047 | 0.9347 | 0.9545 | **0.9564** | 0.9248 | 0.9525 | 0.9149 |
| Vegetation | 0.9571 | 0.9358 | 0.9624 | 0.9608 | **0.9798** | 0.9725 | 0.9691 |
| Tree | 0.9520 | 0.9492 | 0.9598 | 0.9503 | **0.9728** | 0.9725 | 0.9472 |
| Road | 0.9620 | 0.9535 | 0.9638 | 0.9613 | **0.9894** | 0.9682 | 0.9660 |

**Table 10.** Accuracy of classification result of Area3 with SAM. The best results are shown in bold.

| SVM | AWAN | HSCNN-D | HRNet | M2Hnet | SRT | SRT* | GF-1 |
|---|---|---|---|---|---|---|---|
| OA | 0.8298 | 0.8344 | 0.8439 | 0.8301 | **0.8490** | 0.8438 | 0.8367 |
| Kappa | 0.7397 | 0.7445 | 0.7563 | 0.7420 | **0.7599** | 0.7509 | 0.7484 |
| Water | 0.8583 | 0.8583 | 0.8714 | 0.8485 | 0.8856 | **0.8892** | 0.8574 |
| Build | 0.3822 | 0.4195 | 0.4016 | 0.4119 | **0.4345** | 0.4338 | 0.3973 |
| Bare Land | 0.8599 | 0.8617 | **0.8804** | 0.8596 | 0.8800 | 0.8779 | 0.877 |
| Vegetation | 0.6271 | 0.6101 | **0.6786** | 0.6732 | 0.6758 | 0.6205 | 0.6083 |
| Tree | 0.9309 | 0.9289 | **0.9396** | 0.9237 | 0.9375 | 0.9276 | 0.9363 |
| Road | 0.6356 | 0.6634 | 0.6574 | 0.6634 | **0.6647** | 0.6634 | 0.6436 |
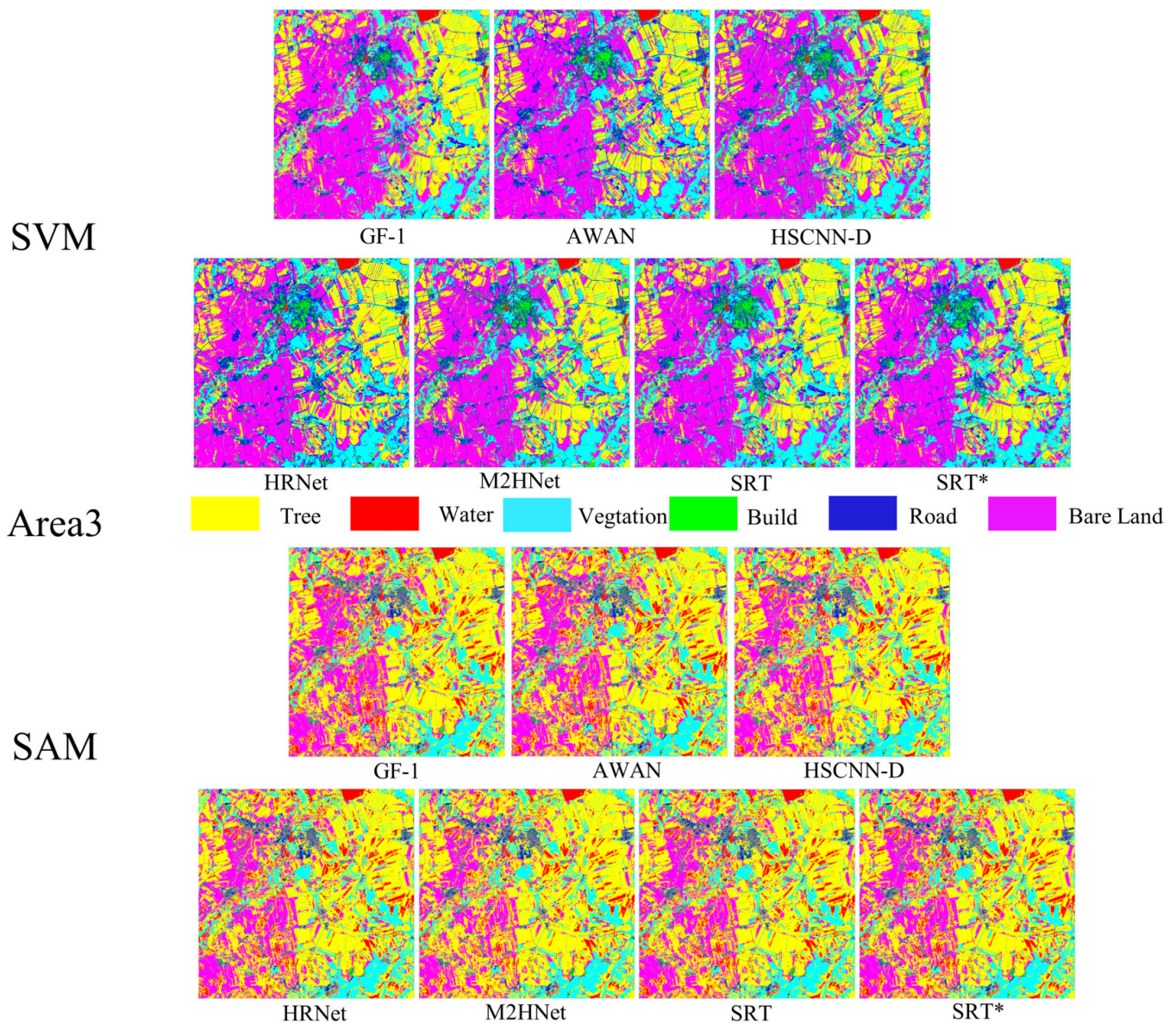
**Figure 10.** Result of SVM classification on Area3.

Tables 4–10 show that both the SRT and SRT* outperform other methods in terms of overall accuracy, which indicates that the TFEM has a significant advantage in performing spectral feature extraction. The SRT results are still the best in terms of SVM and SAM. By comparing the results of SRT and SRT*, we find that SRT needs to use RGM to prevent the model from losing some details during GF-1 PMS image inference. In addition, under the condition of the same samples, the classification result of SAM is lower than that of SVM. We think that the main reason is that the number of image bands used for classification is too small compared to hyperspectral images, which cannot exert the performance of SAM.

Our method has a robust spectral reconstruction capability, and the reconstructed bands can improve the classification capability of GF-1 PMS images.

*4.5. Comparison of Computational Cost*

Table 11 shows the parameters, GFLOPs (giga floating-point operations per second), and the running time of all test methods on an input image of $4 \times 128 \times 128$ pixels. Comparing the parameter quantities of SRT and SRT*, it can be found that the parameter quantity of RGM is only 0.08 M, and the GFLOPs and running time increase by 1.21 and

0.02 s, respectively. In addition, the SRT method is only higher than HSCNN-D in the number of parameters and lower than the other three methods. Although the parameter quantity of HSCNN-D is small, the running time is very long, much higher than 0.27 s of SRT, mainly due to the series structure of HSCNN-D, the number of network layers is deepened, and the network operation takes a lot of time.

**Table 11.** The complexity comparison of different models

|  | **AWAN** | **HSCNN-D** | **HRNet** | **M2Hnet** | **SRT** | **SRT*** |
|---|---|---|---|---|---|---|
| Params (M) | 21.58 | 4.62 | 32.04 | 22.73 | 17.62 | 17.54 |
| GFLOPs | 352.85 | 75.64 | 40.89 | 245.86 | 121.66 | 120.45 |
| Time (S) | 0.21 | 1.21 | 0.41 | 0.24 | 0.27 | 0.25 |

**5. Conclusions**

This article proposes a Transformer- and ResNet-based network (SRT) to reconstruct GF-1 PMS images from GF-6 WFV. SRT consists of three parts: the TFEM, the RDM, and the RGM. The TFEM learns correlation between spectra by the attention mechanism. We use the RDM to reconstruct these relevant features locally and apply the RGM to globally reconstruct.

To ensure the model's generalization, we produce a wide-range, land-type-rich band mapping dataset and test the accuracy in similarity and classification. Meanwhile, to verify whether the knowledge learned from the GF-6 WFV images can be applied to the GF-1 PMS images with inconsistent spatial resolution, we refer to the method of Deng [15] and Li [16]. We believe that the reconstructed band can improve the classification ability of the original image and test it on the Area2 (city is the main scene) and Area3 (farmland is the main scene) GF-1 PMS images. The results show that SRT performs well on both the testing set and the classification accuracy of Area1, Area2, and Area3 compared to other spectral reconstruction methods. The classification accuracy of the reconstructed 8-band images is significantly higher than that of the original 4-band GF-1 PMS images.

In future work, our method still has the following aspects worth expanding on and improving: (1) The structure of the model needs to be improved. Although the parameter quantity of SRT decreases, the detection time does increase slightly. (2) Can it be extended to other satellites, such as GaoFen-2 and GaoFen-4?

## Appendix A

**Table A1.** Data acquisition for the study areas.

| Application | Satellite | Sensor | Acquisition Data | Location |
|---|---|---|---|---|
| | GF-6 | WFV | 10 October 2018 | 85.8°E 44.6°N |
| | GF-6 | WFV | 4 September 2018 | 100.5°E 31.3°N |
| | GF-6 | WFV | 11 October 2018 | 102.5°E 40.2°N |
| Train | GF-6 | WFV | 5 October 2018 | 110.1°E 26.9°N |
| | GF-6 | WFV | 29 October 2018 | 114.8°E 31.3°N |
| | GF-6 | WFV | 18 September 2018 | 118.6°E 42.4°N |
| | GF-6 | WFV | 1 October 2018 | 88.8°E 40.2°N |
| Test | GF-6 | WFV | 17 October 2018 | 114.9°E 38.0°N |
| | GF-6 | WFV | 16 September 2018 | 129.9°E 46.8°N |
| Area1 | GF-6 | WFV | 16 September 2018 | 129.9°E 46.8°N |
| Area2 | GF-1 | PMS1 | 4 November 2016 | 125.3°E 48.8°N |
| Area3 | GF-1 | PMS2 | 21 June 2018 | 117.2°E 35.2°N |

## References

1. Wu, Z.; Zhang, J.; Deng, F.; Zhang, S.; Zhang, D.; Xun, L.; Javed, T.; Liu, G.; Liu, D.; Ji, M. Fusion of GF and MODIS Data for Regional-Scale Grassland Community Classification with EVI2 Time-Series and Phenological Features. *Remote Sens.* **2021**, *13*, 835. doi: 10.3390/rs13050835. [CrossRef]
2. Jiang, X.; Fang, S.; Huang, X.; Liu, Y.; Guo, L. Rice Mapping and Growth Monitoring Based on Time Series GF-6 Images and Red-Edge Bands. *Remote Sens.* **2021**, *13*, 579. [CrossRef]
3. Kang, Y.; Hu, X.; Meng, Q.; Zou, Y.; Zhang, L.; Liu, M.; Zhao, M. Land Cover and Crop Classification Based on Red Edge Indices Features of GF-6 WFV Time Series Data. *Remote Sens.* **2021**, *13*, 4522. [CrossRef]
4. Arad, B.; Ben-Shahar, O. Sparse recovery of hyperspectral signal from natural RGB images. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 19–34.
5. Aeschbacher, J.; Wu, J.; Timofte, R. In defense of shallow learned spectral reconstruction from RGB images. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 471–479.
6. Fu, Y.; Zheng, Y.; Zhang, L.; Huang, H. Spectral Reflectance Recovery From a Single RGB Image. *IEEE Trans. Comput. Imaging* **2018**, *4*, 382–394. [CrossRef]
7. Li, Y.; Wang, C.; Zhao, J. Locally Linear Embedded Sparse Coding for Spectral Reconstruction From RGB Images. *IEEE Signal Process. Lett.* **2018**, *25*, 363–367. [CrossRef]
8. Geng, Y.; Mei, S.; Tian, J.; Zhang, Y.; Du, Q. Spatial Constrained Hyperspectral Reconstruction from RGB Inputs Using Dictionary Representation. In Proceedings of the IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 3169–3172. [CrossRef]
9. Gao, L.; Hong, D.; Yao, J.; Zhang, B.; Gamba, P.; Chanussot, J. Spectral superresolution of multispectral imagery with joint sparse and low-rank learning. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 2269–2280. [CrossRef]
10. Xiong, Z.; Shi, Z.; Li, H.; Wang, L.; Liu, D.; Wu, F. Hscnn: Cnn-based hyperspectral image recovery from spectrally undersampled projections. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 518–525.
11. Alvarez-Gila, A.; Van De Weijer, J.; Garrote, E. Adversarial networks for spatial context-aware spectral image reconstruction from rgb. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 480–490.
12. Koundinya, S.; Sharma, H.; Sharma, M.; Upadhyay, A.; Manekar, R.; Mukhopadhyay, R.; Karmakar, A.; Chaudhury, S. 2D-3D cnn based architectures for spectral reconstruction from RGB images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 844–851.
13. Shi, Z.; Chen, C.; Xiong, Z.; Liu, D.; Wu, F. Hscnn+: Advanced cnn-based hyperspectral recovery from rgb images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 939–947.
14. Zhao, Y.; Po, L.M.; Yan, Q.; Liu, W.; Lin, T. Hierarchical regression network for spectral reconstruction from RGB images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 422–423.
15. Deng, L.; Sun, J.; Chen, Y.; Lu, H.; Duan, F.; Zhu, L.; Fan, T. M2H-Net: A Reconstruction Method For Hyperspectral Remotely Sensed Imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 323–348.: 10.1016/j.isprsjprs.2021.01.019. [CrossRef]
16. Li, T.; Gu, Y. Progressive Spatial–Spectral Joint Network for Hyperspectral Image Reconstruction. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]

17. Zhang, X.; Xu, H. Reconstructing spectral reflectance by dividing spectral space and extending the principal components in principal component analysis. *J. Opt. Soc. Am. A* **2008**, *25*, 371–378. [CrossRef] [PubMed]

18. Liu, X.; Liu, L. Improving chlorophyll fluorescence retrieval using reflectance reconstruction based on principal components analysis. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1645–1649.

19. Haneishi, H.; Hasegawa, T.; Hosoi, A.; Yokoyama, Y.; Tsumura, N.; Miyake, Y. System design for accurately estimating the spectral reflectance of art paintings. *Appl. Opt.* **2000**, *39*, 6621–6632. [CrossRef] [PubMed]

20. Imai, F.H.; Berns, R.S. Spectral estimation using trichromatic digital cameras. In Proceedings of the International Symposium on Multispectral Imaging and Color Reproduction for Digital Archives, Chiba, Japan, 21–22 October 1999; Volume 42, pp. 1–8.

21. Cheung, V.; Westland, S.; Li, C.; Hardeberg, J.; Connah, D. Characterization of trichromatic color cameras by using a new multispectral imaging technique. *JOSA A* **2005**, *22*, 1231–1240. [CrossRef] [PubMed]

22. Zhang, J.; Su, R.; Ren, W.; Fu, Q.; Nie, Y. Learnable Reconstruction Methods from RGB Images to Hyperspectral Imaging: A Survey. *arXiv* **2021**, arXiv:2106.15944.

23. Arad, B.; Ben-Shahar, O.; Timofte, R.; Gool, L.V.; Yang, M.H. NTIRE 2018 Challenge on Spectral Reconstruction from RGB Images. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–23 June 2018.

24. Li, J.; Wu, C.; Song, R.; Li, Y.; Liu, F. Adaptive weighted attention network with camera spectral sensitivity prior for spectral reconstruction from RGB images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 462–463.

25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017.

26. Lin, T.; Wang, Y.; Liu, X.; Qiu, X. A survey of transformers. *arXiv* **2021**, arXiv:2106.04554.

27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.

28. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning (PMLR), Virtual Event, 13–14 August 2021; pp. 10347–10357.

29. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Seattle, WA, USA, 14–19 June 2020; pp. 213–229.

30. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.

31. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.

32. Valanarasu, J.M.J.; Oza, P.; Hacihaliloglu, I.; Patel, V.M. Medical transformer: Gated axial-attention for medical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; pp. 36–46.

33. Arad Hudson, D.; Zitnick, L. Compositional Transformers for Scene Generation. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Online, 6–12 December 2020; pp. 9506–9520.

34. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. *arXiv* **2021**, arXiv:2111.06377.

35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

36. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 4700–4708.

37. LeCun, Y.; Boser, B.; Denker, J.; Henderson, D.; Howard, R.; Hubbard, W.; Jackel, L. Handwritten digit recognition with a back-propagation network. In Proceedings of the Advances in Neural Information Processing Systems 2, Denver, CO, USA, 27–30 November 1989.

38. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.

39. Huynh-Thu, Q.; Ghanbari, M. Scope of validity of PSNR in image/video quality assessment. *Electron. Lett.* **2008**, *44*, 800–801. [CrossRef]

40. CRESDA. China Centre For Resources Satellite Data and Application. 2021. Available online: http://www.cresda.com/CN/index.shtml (accessed on 2 June 2022).

41. Yuhas, R.H.; Goetz, A.F.; Boardman, J.W. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. In *Summaries of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop*; Jet Propulsion Laboratory: La Cañada Flintridge, CA, USA, 1992.

42. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]