A robust and adaptive spatial-spectral fusion model for PlanetScope and Sentinel-2 imagery

Yongquan Zhao and Desheng Liu

Department of Geography, The Ohio State University, Columbus, OH United States

ABSTRACT

Satellite sensors usually compromise between their spatial and spectral resolutions due to the limitations of data volume and signal-to-noise ratio, such as the Very High spatial Resolution (VHR) PlanetScope (PS, four or five 3-m bands) and medium spatial resolution (med-resolution) Sentinel-2 (S2, ten 10-m or 20-m bands) constellations. Concomitant with the growing demand for satellite images with ample spatial details and spectral signatures, Spatial-Spectral Image Fusion (SSF) is important for blending the spatial resolution of PS and the spectral resolution of S2 to produce synthetic 3-m image products in the ten bands for different applications. However, the existing studies conducted for fusing PS and S2 data present limited spatial and spectral fidelities to original PS and S2 bands. Hence, this study presents a new SSF method named Robust and Adaptive Spatial-Spectral image Fusion Model (RASSFM) for that purpose. RASSFM improves the spatial and spectral fidelities of the results through: (1) combining the spectral mapping and the spectral correlation to obtain the high-quality spatial information sources for the S2 bands, (2) utilizing the neighborhood information considering both spatial and spectral constraints to improve the spectral fidelities of the fused bands. We conducted the SSF tests at the degraded and original spatial resolutions to comprehensively evaluate the proposed RASSFM method. Furthermore, we examined the performance of RASSFM in four study sites with highly mixed regular or desert urban, heterogeneous agricultural, and vegetation-dominated landscapes. Moreover, we compared our method with twenty representative SSF methods developed for similar purposes. The results demonstrate that RASSFM not only outperforms the other methods but has robust performance in different landscapes and comparable accuracies in different bands, thereby advancing the fusion of VHR and med-resolution images from PS and S2, respectively.

1. Introduction

Along with the development of Earth Observation (EO) systems, a large number of satellites have been launched with multiple spatial, temporal, and spectral resolutions, including medium spatial resolution (med-resolution) satellites such as Landsat 8/9 and Sentinel-2, and Very High spatial Resolution (VHR, ≤5 m) satellites such as WorldView and RapidEye. In recent years, the PlanetScope (PS) satellite constellation has become a potential game changer in the EO field, which can collect daily imagery of the entire Earth at about 3-m spatial resolution in four (Blue (B), Green (G), Red (R), and Near-InfraRed (NIR)) or five (B, G, R, NIR, and Red Edge (RE)) spectral bands (Planet 2020b). Such a super EO ability is achieved by observations from the 180 + CubeSats in the PS constellation (Planet 2020a), which are operated in the International Space Station (ISS, with a variable equator crossing time) and Sun Synchronous Orbits (SSO, with an equator crossing time between 9:30 and 11:30 am local solar time). To date, there are three generations of PS sensors in operation accompanied by the evolution of sensor technologies, including the first-generation PS2, the second-generation PS2.SD, and the third-generation PSB.SD mounted on the Dove, Dove-R, and SuperDove flocks, respectively (Planet 2020b). Additionally, PS2 and PS2.SD are both with four spectral bands (RGB + NIR), and the spectral ranges of the PS2.SD bands are narrower than PS2. Compared with PS2 and PS2.SD, the latest PSB.SD has an additional RE band (Planet 2020b). Specifically, the spectral and spatial resolutions of the PS sensors are referred to Table 1.

VHR imagery can provide unprecedented spatial details to benefit many remote sensing applications, such as benthic habitat and seagrass species mapping (Wicaksono and Lazuardi 2018), improved leaf area...
Table 1. The wavelengths (nm) and the spatial resolutions (m) of the PS* and S2 sensors.

|        | PS2 | PS2 SD | PSB SD | MSI  |
|--------|-----|--------|--------|------|
| Band   | Spatial resolution | Band   | Spatial resolution | Band   | Spatial resolution | Band   | Spatial resolution |
| B1 – Blue: 455–515 | 3    | B1 – Blue: 465–517 | 3    | B1 – Blue: 465–515 | 3    | B2 – Blue: 459–525 | 10   |
| B2 – Green: 500–590 | 3    | B2 – Green: 547–585 | 3    | B2 – Green: 547–583 | 3    | B3 – Green: 542–578 | 10   |
| B3 – Red: 590–670  | 3    | B3 – Red: 650–682  | 3    | B3 – Red: 650–680  | 3    | B5 – Red: 697–713  | 20   |
| B4 – NIR: 780–860 | 3    | B4 NIR: 840–888 | 3    | B5 NIR: 840–888 | 3    | B6 – RE2: 733–748 | 20   |
|         |      |        |        |      |        | B7 – RE3: 773–793 | 20   |
|         |      |        |        |      |        | B8 – NIR: 780–886 | 10   |
|         |      |        |        |      |        | B8A – NIR: 854–875 | 20   |
|         |      |        |        |      |        | B11 – SWIR1: 1568–1659 | 20   |
|         |      |        |        |      |        | B12 – SWIR2: 2115–2290 | 20   |

* PS CubeSats have different imaging instruments, including the classic Dove sensor (PS2), Dove-R sensor (PS2 SD), and “SuperDove” sensor (PSB SD).

index estimation for agriculture landscapes (Kimm et al. 2020), and fine-scale bathymetry and water quality retrieval (Niroumand-Jadidi et al. 2020). Due to the trade-off between the spatial resolution and the spectral resolution of satellites (Shen, Meng, and Zhang 2016), however, PS is excellent in spatial information collecting ability but unsatisfied in resolving spectral profiles of surface materials. Although PS can acquire daily VHR imagery with global coverage, the spectral bands of PS imagery is limited for remote sensing studies that require more feature bands, such as mangrove species mapping (Li, Wong, and Fung 2019a), forest pest detection (Stych et al. 2019), river discharge estimation (Feng et al. 2019), harmful algal blooms delineation (Zhao, Liu, and Wei 2020), and anomaly target detection (Xie et al. 2021a, 2021b).

On the other hand, med-resolution satellites such as the Sentinel-2 (S2) constellation can acquire global imagery every five days with ample spectral information but at lower spatial resolutions. The spatial resolutions of the S2A/B MultiSpectral Instrument (MSI) sensor are various, including four bands with a 10-m resolution (B, G, R, NIR), six bands with a 20-m resolution (RE1, RE2, RE3, Narrow NIR (NNIR), ShortWave InfraRed 1 (SWIR1), and SWIR2), and three bands with a 60-m resolution. Since this study focuses on land surfaces, the 60-m bands are excluded because they are generally used for atmospheric studies. Specifically, the spectral and spatial resolutions of the adopted ten 10- and 20-m S2 bands are referred to Table 1. The six bands that are missing in PS can improve many remote sensing applications due to the added value of these spectral bands, such as land-use land-cover mapping (Forkuor et al. 2018; Zhao and Zhu 2022), plant species discrimination (Otunga et al. 2019), and cyanobacterial cell density estimation (Page, Kumar, and Mishra 2018). Remote sensing studies relying on either data source would lead to limited accuracy (Feng et al. 2019).

Therefore, combining the merits of both data sources (PS and S2) to provide synthetic data that have ample information content in both spatial and spectral dimensions to broaden their application scopes is of great significance (Zhao and Huang 2017). A feasible and cost-effective solution is adopting Spatial-Spectral image Fusion (SSF) to blend the spatial resolution of PS and the spectral resolution of S2 to generate synthetic VHR (i.e. 3-m) imagery with enhanced spectral resolutions, thereby providing rich spatial and spectral information simultaneously for remote sensing studies. Generally, SSF can be divided into two categories: the classical pansharpening (Zhang 1999) and the relatively new hypersharpening (Selva et al. 2015). For pansharpening, it aims to fuse the high spatial resolution of panchromatic imagery and the high spectral resolution of multispectral or hyperspectral imagery to generate multispectral or hyperspectral imagery with high spatial resolution (Xie et al. 2019). Representative pansharpening methods include the component-substitution-based (Carper, Lillesand, and Kiefer 1990), multiresolution-analysis-based (Liu 2000), variational-optimization-based (Ballester et al. 2006), and machine-learning-based (Masi et al. 2016) methods. For hypersharpening, it aims to fuse the high spatial resolution of multispectral imagery (e.g. SPOT-5, WorldView) and the high spectral resolution of hyperspectral imagery (e.g. Hyperion, AVIRIS) to generate high-spatial-resolution hyperspectral imagery. To date, several approaches have been developed for hypersharpening or similar purposes, such as the component-substitution-based (Winter et al. 2007), unmixing-based (Zurita-Milla, Clevers, and Schaepman 2008),
sparse-representation-based (Song et al. 2014; Wei et al. 2015), and integrated (Vivone and Chanussot 2020) methods. However, the SSF fusion algorithms designed for specific satellite sensors may perform less satisfactorily when applied to another sensor with different spatial and spectral characteristics or scales (Huang et al. 2013).

Therefore, increasing attention has been placed on the fusion between 3-m PS imagery and other satellite data in recent years due to the growing popularity of PS imagery. For instance, Gašparović et al. (2018) adopted pansharpening methods to improve the spatial resolution of the eight 10-m/20-m S2 bands (excluding the two SWIR bands) using the spatial details from the four PS bands. Specifically, the PS band with the highest correlation to a S2 band is directly used as the pansharmonic image to sharpen the corresponding S2 band. Houborg and McCabe (2018) proposed a multi-scale cross-sensor radiometric calibration method to transform the spectral characteristics of PS to Landsat 8 to produce temporally consistent surface reflectance images in their equivalent bands (B, G, R, NIR), which did not consider the spectral bands that overshoot the spectra of PS, i.e. the SWIR bands of Landsat 8. Furthermore, Li et al. (2019b) devised a hybrid fusion scheme that includes a spatial unmixing process and an existing image superresolution method to downscale the 10-m/20-m S2 bands using the spatial frequencies from the four PS bands. The results demonstrate limited image sharpness effects in the fused results (see Figures 5 and 6 in (Li et al. 2019b)), which show the spatial characteristics of the 10-m/20-m S2 bands. In operational use on continental or global scales, the combination of both spatial unmixing and superresolution techniques may lead to massive computing burdens (Zhao et al. 2021). Li et al. (2020b) evaluated the performance of two effective pansharpening approaches (high pass modulation and Model 3) in improving S2 10-m/20-m imagery to the PS 3-m resolution. The straightforward algebraic calculation in the pansharpening methods works well for homogeneous landscapes, but it may have limitations for highly mixed urban areas with complex spectral signatures in VHR satellite imagery (Alparone et al. 2008). In summary, the spectral fidelities of the fused results in different landscapes have not been deeply investigated, particularly for the SWIR bands that overshoot the PS spectra. Moreover, the spatial fidelities of the results to original 3-m PS imagery still have room for improvement. Hence, further efforts on the SSF method development for fusing the 10-m/20-m S2 and 3-m PS imagery are necessary.

In light of the limitations described above, a new SSF method called Robust Adaptive Spatial-Spectral image Fusion Model (RASSFM) is proposed and validated in this study, which fuses observations from PS sensors with VHR but low spectral resolution and S2 MSI with med-resolution but high spectral resolution to produce synthetic image products with VHR and high spectral resolution. Basically, RASSFM differs in two aspects from previous hypersharpening studies. First, the pre-fusion step adopts a combination of spectral mapping and spectral correlation to generate ten simulated 3-m bands to act as the spatial information sources for the posterior fusion steps rather than using only one component of the two (Vivone and Chanussot 2020). Second, the spatial neighborhood information considering both spatial and spectral constraints is introduced into the fusion model to improve the spectral fidelities of the results. We evaluated RASSFM qualitatively and quantitatively via different fusion schemes in different landscapes and compared it with twenty representative fusion methods developed for similar purposes. In the remainder of this paper, we will demonstrate the study areas and datasets in Section 2, describe the methodology and evaluation schemes in Section 3, present the results in Section 4, discuss and conclude the study in Sections 5 and 6, respectively.

2. Study areas and experimental datasets

2.1. Study areas

We examined the fusion methods in four study sites with different land covers and climate types to demonstrate their effectiveness (see Figure 1). The first site is in Columbus, Ohio, United States (centered at 40.0056°N, 83.0292°W). This site is an urban area with various land cover types that are highly fragmented, such as built-up areas (e.g. buildings/grounds/roads/parking lots), vegetation (e.g. woodlands/lawns), waters (e.g. rivers/lakes), wetlands, bare lands. The rich spectral characteristics and the complex spatial patterns of the surface materials pose more challenges on resolving the fine-resolution spectral profiles. The second site is in Las Vegas, Nevada, United States (centered at 36.2179°N,
115.1864°W), which is another challenging study area in desert urban areas. This site is mainly covered by buildings, airport features, bare lands, roads, and a small proportion of vegetation. Involving this site is to examine the fusion performance in desert cities that are different from regular cities like Site I. The third study site is dominated by agricultural land covers in the Coleambally Irrigation Area in southern New South Wales, Australia (centered at 34.7849°S, 145.9903°E). This site is an irrigated field of various summer crops, such as rice, maize, sorghum, and soybean (Emelyanova et al. 2013), presenting a heterogeneous landscape. In this study, the Coleambally site mainly contains irrigation areas, dry cropland, trees, and other land cover types. The spatial heterogeneity of this site leads to high spectral variances, which can be used to examine the robustness of the fusion methods. The fourth site in northeastern Ohio, United States (centered at 41.6002°N, 81.3709°W) is mainly covered by vegetation (e.g. forest and lawns) and small parcels of built-up areas (e.g. roads and houses). Introducing this site is to enrich the land cover conditions of our experiments. On the other hand, vegetation is one of the most important variables for remote sensing applications (Tian et al. 2021).

Furthermore, the area of the four study sites is 3.6 × 3.6 km², which corresponds to an image size of 1200 × 1200 pixels for 3-m PS images, 360 × 360 pixels for 10-m S2 bands, and 180 × 180 pixels for 20-m S2 bands, respectively. The spatial coverage of the study sites can clearly demonstrate the fusion effects at a local scale as achieving good fusion results at a local scale is the basis for obtaining good fusion at a regional or global scale (Zhao, Huang, and Song 2018).
2.2. PlanetScope imagery

The PS image of Site I was acquired by PSB.SD on 1 October 2021; the PS images of Sites II&IV were observed by PS2.SD on 8 August 2020 and 1 July 2019 respectively; the PS image covering Site III was collected by PS2 on 10 August 2020. Additionally, all the PS images were acquired by the PS sensors in the SSO orbit. We adopted the surface reflectance product (Level 3B) (Planet 2017) from the Planet Explorer (https://www.planet.com/explorer/) as the high spatial but low spectral resolution data. The PS illustrations are displayed in true- and false-color (NIR-R-G as RGB composites) to present the spectral characteristics.

2.3. Sentinel-2 imagery

The S2 images of Sites I&II&IV were acquired by S2A and on October 01, 10 August 2021 2020, and 1 July 2019 respectively; the S2 image of Site II was acquired by S2B on 8 August 2020. We used the S2 surface reflectance product (Level 2A) as the low spatial but high spectral resolution data, which were downloaded via Google Earth Engine (GEE; https://code.earthengine.google.com/). The S2 bands were up-sampled to 3-m using the bi-cubic interpolation method to match the grids of PS. Moreover, all clipped PS and S2 images with the same geographic coverage were projected onto the Universal Transverse Mercator (UTM) coordinate system under the WGS-84 datum. Additionally, the S2 images were geometrically matched with the PS images by applying optimal offsets (Gevaert and García-Haro 2015) within the same map coordinates. The S2 illustrations are displayed with different color models to demonstrate the spectral information explicitly.

3. Methodology and Evaluation Settings

For the convenience and clarity of describing the image fusion method, the symbols of the main variables used in this study are defined in Table 2.

Figure 2 demonstrates the workflow of the proposed RASSFM method for fusing PS 3-m and S2 10-m/20-m imagery, which comprises two important modules: (1) the spectral mapping and correlation based on $S^{(H,A)}$ and $P^{(H,A)}$ to obtain $P^{(H,A)}$; (2) the fusion of the spatial details of $P^{(H,A)}$ and the spectral components of $S^{(H,A)}$ to obtain $SP^{(H,A)}$. Note that although PS has 3-m observations in the four bands (BRG + NIR), we still downscale the corresponding S2 10-m bands to 3-m because the radiometric characteristics and cross-satellite consistencies of PS CubeSats are not as good as rigorously calibrated scientific satellites such as Landsat8 or Sentinel-2 (Houborg and McCabe 2018).

| Table 2. The main variables and their symbols used in RASSFM. |
| Symbol | Variable | Description |
|--------|----------|-------------|
| $\lambda$ | The four spectral bands of PS | B, G, R, NIR |
| $\Lambda$ | The ten spectral bands of S2 | B, G, R, RE1, RE2, RE3, NIR, NNIR, SWIR1, SWIR2 |
| $\tilde{\lambda}$ | The ten transformed bands similar to $\Lambda$ | Obtained by Section 3.1.1 |
| $\tilde{\Lambda}$ | The ten correlated bands similar to $\Lambda$ | Obtained by Section 3.1.2 |
| $\Lambda'$ | The ten simulated bands similar to $\Lambda$ | The combination of $\tilde{\lambda}$ and $\Lambda$ with the better spectral similarity to $\Lambda$ |
| $h$ | The spatial resolution of original S2 images, i.e. 10-m/20-m | Actual low spatial resolution |
| $H$ | The spatial resolution of original PS images, i.e. 3-m | Actual high spatial resolution |
| $H'$ | The spatial resolution of up-sampled S2 images | Up-sampled 3-m grids from 10-m/20-m grids using bicubic interpolation |
| $p^{(H,A)}$ | Input high-spatial-resolution (3-m) but low-spectral-resolution (H) PS images | The source of spatial information |
| $S^{(H,A)}$ | Spatially up-sampled S2 images at 3-m grids with the actual ten spectral bands ($\lambda$) | The source of spectral information, which was resampled from 10-m/20-m to 3-m grids to match the girds of $P^{(H,A)}$ |
| $P^{(H,A)}$ | Actual 3-m resolution images with the ten simulated bands ($\Lambda'$) | The band amount of $P^{(H,A)}$ was transformed from four to ten to match the spectral bands of $S^{(H,A)}$ for the band-to-band fusion process |
| $p^{(H,A)}$ | Up-sampled images at 3-m grids with the ten simulated bands | Obtained by down-sampling $P^{(H,A)}$ to 10-m/20-m using image aggregation and then up-sampling to 3-m to match the girds of $P^{(H,A)}$ |
| $SP^{(H,A)}$ | The synthetic image with the spatial details of PS in a 3-m resolution and the spectral characteristics of S2 in the ten bands | Final fusion result |
hypersharpness, spectral transformation and spectral correlation have been used for similar purposes (Vivone and Chanussot 2020). However, our tests showed that either approach generates 3-m bands with different levels of spectral similarities to the S2 bands in different study sites, particularly for the six 20-m bands. Hence, we combined the 3-m bands generated by the two approaches based on their spectral similarities to the S2 bands.

Note that the spectral characteristics of the obtained 3-m simulated image would be similar to the original S2 bands as a whole, but they are not the same. However, the simulated image has actual 3-m spatial resolution in each band, which would act as the spatial information source rather than the spectral information source for the posterior fusion steps. The spectral information source of the final fusion result is still the original S2 bands.

3.1.1. Spectral mapping

The datasets used for the spectral mapping should have the same spatial coverage and small temporal gaps (Winter et al. 2007), which is easy to satisfy in this study due to the daily and global coverage of PS imagery. For simplicity and robustness, a linear spectral transformation between PS and S2 imagery is conducted to transform the spectral bands of PS imagery from four to ten to match the spectral bands of S2 imagery:

\[ P^{(H \lambda)} = M \times P^{(H \lambda)} + \epsilon \]  

where \( \epsilon \) is the error term, \( M \) is the spectral mapping matrix that can be estimated by a least square regression as:

\[ M = S^{(H \lambda)} \times P^{(H \lambda)}^T \times \left( p^{(H \lambda)} \times p^{(H \lambda)}^T \right)^{-1} \]  

For Eq. (1)-(2), each band of \( S^{(H \lambda)} \) and \( P^{(H \lambda)} \) were reshaped to vector forms. Hence, the row and column numbers of \( M \) correspond with the band numbers of \( S^{(H \lambda)} \) and \( P^{(H \lambda)} \), respectively. Note that although \( p^{(H \lambda)} \) has the same number of bands with \( S^{(H \lambda)} \), it has the same rank with the \( p^{(H \lambda)} \) because it is generated via a global linear transformation between PS and S2 images (Winter et al. 2007). Therefore, \( p^{(H \lambda)} \) has the same number of statistically separable spectral components as \( p^{(H \lambda)} \) (Winter et al. 2007).

3.1.2. Spectral correlation

In similar studies, \( p^{(H \lambda)} \) was directly used as the spatial information source for fusion (Li et al. 2020b; Winter et al. 2007). However, we found that the linear transformation above was less satisfying for the RE1&2&3 and NNIR bands in some study sites. Therefore, we added another option for the four bands based on the spectral correlation (Selva et al. 2015) between PS and S2 bands:

\[ \bar{I} = \arg \max_j \text{corr} \left( P^j_{(H \lambda)} , S^k_{(H \lambda)} \right) \]
\[ p_k^{(H,x)} = p_i^{(H,x)} \]

where \( l \) and \( k \) are the band indices of PS and S2, respectively. \( \text{corr} \) means the Pearson's correlation coefficient. \( p_i^{(H,x)} \) is the resampled PS band that has the largest correlation coefficient with the \( k \)th S2 band. Then, we combined \( p^{(H,A)} \) and \( p^{(H,x)} \) based on their spectral similarities to the original S2 bands to obtain \( p^{(H,A)} \) for the following fusion process:

\[ p^{(H,A)} = \text{BandCombination}(p^{(H,A)}, p^{(H,x)}) \]

where the BandCombination operator means a band of \( p^{(H,A)} \) or \( p^{(H,x)} \) that has the largest correlation coefficient with the corresponding S2 band will be given to their counterpart in \( p^{(H,A)} \).

### 3.2. Blending PlanetScope and Sentinel-2 imagery

#### 3.2.1. Spatial-spectral fusion model

Theoretically, there is a constant systematic bias between homogenous pixels observed by multiple satellites on the same date due to the different observation conditions, e.g., satellite orbit parameters, bandwidth, imaging time (Gao et al. 2006; Huang, Zhang, and Yu 2012). In this study, the simulated \( p^{(H,A)} \) and the target \( S^{(H,A)} \) can be regarded as the observations of two virtual satellite sensors with a 3-m spatial resolution. Correspondingly, \( p^{(H,A)} \) and \( S^{(H,A)} \) are the spatially degraded counterparts of \( p^{(H,A)} \) and \( S^{(H,A)} \) from 3-m to 10-m/20-m resolutions to model the spatial degradation of the S2 imaging process. Hence, there is a systematic bias between the two data sources across different scales (3-m and 10-m/20-m) based on the above assumption:

\[ S^{(H,A)}_k = p^{(H,A)}_k + \epsilon \]

\[ S^{(H,A)} = p^{(H,A)} + \epsilon \]

where \( \epsilon \) is the systematic bias. Therefore, we can derive the following spatial-spectral fusion model based on Eq. (6) and (7):

\[ S^{(H,A)}_k = S^{(H,A)} + (p^{(H,A)}_k - p^{(H,A)}_k) \]

The subtraction between \( p^{(H,A)}_k \) and \( p^{(H,A)}_k \) is used to extract the spatial information of the band \( k \). Then, the spatial information is added to \( S^{(H,A)}_k \) to obtain \( S^{(H,A)}_k \), which has the spectral information of \( S^{(H,A)}_k \) and the spatial information of \( p^{(H,A)}_k \). This process is conducted in a band-by-band form for the ten bands of \( S^{(H,A)}_k \).

However, the above model that processes the image band as a whole may lead to significant biases for heterogeneous or mixed pixels (Gao et al. 2006). Hence, we introduce the neighborhood information of a target pixel to mitigate the deficiency via the weighted summation model below:

\[ S^{(H,A)}_{(c,c,k)} = \sum_{i=0}^{2c} \sum_{j=0}^{2c} W_{(i,j)} \left( S^{(H,A)}_{(i,j,k)} + p^{(H,A)}_{(i,j,k)} - p^{(H,A)}_{(i,j,k)} \right) \]

where \( c \) is the central pixel coordinate of a local moving window, \( i \) and \( j \) represent the pixel coordinates within the moving window. \( W_{(i,j)} \) is the weight of each pixel within the moving window. Note that the moving window size can be adjusted based on the spatial resolution gap between the input PS and S2 images. We only adopt the neighboring pixels that are similar to the central pixel for the fusion model in Eq. (9) to ensure high spectral fidelity of the fusion results, which contains two important steps, i.e., similar neighbor searching and neighbors’ weight calculation.

#### 3.2.2. Similar neighbor searching and weight determination

We use the image-patch-based non-local searching scheme to obtain its similar neighbors within a moving window (Zhao, Huang, and Song 2018) for each pixel in \( p^{(H,A)} \), which is expanded as a central image patch of the moving window. The similarities between the neighboring patches (including the central patch itself) and the central patch are measured by the following normalized exponential function,

\[ R = \frac{1}{Z(c)} e^{-\frac{\| V_c p^{(H,A)} - V_{(i,j)} p^{(H,A)} \|_2^2}{h^2}} \]

where \( V_c p^{(H,A)} \) and \( V_{(i,j)} p^{(H,A)} \) are the central patch and its similar neighborhoods in \( p^{(H,A)} \), respectively. \( R \) represents their similarity measurement. \( \| V_c p^{(H,A)} - V_{(i,j)} p^{(H,A)} \|_2^2 \) is a Gaussian kernel (with standard deviation \( a \)) weighted Euclidean distance, \( Z(c) \) is a normalizing
constant $Z(c) = \sum_{h} e^{-\frac{\|V_{c} - V_{n(h)}\|^2}{2\sigma^2}}$, and $h$ is the degree of decay. In this study, the image patch size is set as $3 \times 3$ ($9 \times 9$ m$^2$) for computing efficiency and searching accuracy. Additionally, we adopt the first $N$ most similar neighbors based on the similarity measurements. Note that the similar neighbor’s amount $N$ can be adjusted based on the spatial resolution gap between PS and S2 images.

After obtaining the similar neighbors, the weight of each neighbor is determined based on four factors: 1) similar neighbors with higher homogeneity ($Q$) should have larger weights, 2) similar neighbors with smaller cross-sensor spectral differences ($F$) should have larger weights, 3) similar neighbors with smaller spatial distances to the central patches ($D$) should have larger weights, and 4) similar neighbors with smaller spectral differences to the central patches ($SD$) should have larger weights. These four weight factors are combined as:

$$X = Q \ast F \ast (D + \rho) \ast (SD + \rho)$$

(11)

$$Q = \ln \left( \left| V_{c(i,j)}^{(i',j')} - V_{n(i,j)}^{(i',j')} \right| \right)$$

(12)

$$F = \ln \left( \left| V_{c(i,j)}^{(i',j')} - V_{n(i,j)}^{(i',j')} \right| \right)$$

(13)

$$D = \sqrt{(i - c)^2 - (j - c)^2 / \omega}$$

(14)

$$SD = \ln \left( \left| V_{c(i,j)}^{(i',j')} - V_{n(i,j)}^{(i',j')} \right| \right)$$

(15)

where $\rho$ is a small constant to prevent $D$ and $SD$ from becoming zero at the central image patch, $\omega$ is the radius of the moving window. Then, the final weights that consider the four weight factors are calculated by minimizing the locally linear reconstruction error of the following cost function:

$$\epsilon(W) = \sum_{c} \left| X_c - \sum_{n=1}^{N} W_n X_n \right|^2, s.t. W_n \geq 0; \sum_{n} W_n = 1$$

(16)

$X_c$ and $X_n$ mean the combined weight factors for the central patch and its similar neighbors, respectively. $W_n$ is the vector form of $W_{(i,j)}$ in Eq. (9) that contains the optimal weights. The diagrams of the similar neighbor searching and weight calculation steps are shown in Figure 3.

Based on the similar neighbors and their weights obtained from Eq. (10) and (16), $SP_{H,A}$ can be constructed by the weighting function in Eq. (9) with a pixel-wise manner, which assigns the weights of $W_n$ and zero to the similar and non-similar neighbors, respectively. Note that the central pixel value of each reconstructed patch is assigned to the target pixel (the center pixel of the moving window) to ensure image sharpness.

Furthermore, remote sensing images usually have noises, outliers, or abnormal observations due to various interferences, such as cloud cover, atmospheric effects, image calibration (Zhao, Huang, and Song 2018), which would negatively affect the similar neighbor searching and weight calculation steps. For instance, there may be extreme cases that no similar neighbors can be found within a predefined neighborhood. In that case, the target pixel will be assigned a weight of 1. Additionally, if the reconstructed values go beyond the normal range of surface reflectance (0–1), the pixel values generated by Eq. (8) will be employed as alternatives. In short, image noises, mis-identified clouds, uncertainties of atmospheric correction would affect the similar neighbor searching and weight calculation. These factors are tricky for remote sensing image fusion because more uncertainty sources are involved, so we used the simple fusion model in Eq. (8) to reduce error sources for such cases.

### 3.4. Fusion schemes and evaluation

#### 3.4.1. Metrics for fusion at degraded spatial resolutions

We adopted a commonly used validation strategy for spatial-spectral fusion with reference data (Shao et al. 2019; Wald, Ranchin, and Mangolini 1997), which uses image aggregation to down-sample the PS and S2 images to lower spatial resolutions for fusion and takes the original S2 images as references. Since there are two spatial resolutions (10-m and 20-m) for the S2 bands, two branches of fusion were conducted accordingly. First, the four PS and S2 bands (RGB + NIR) were downsampled to 10-m and 30-m resolutions, respectively, based on the original spatial resolution gap (~3X). Second, the four PS bands and the six S2 bands (REs, NNIR, SWIRs) were downsampled to 20-m and 120-m resolutions, respectively, based on
the original spatial resolution gap (~6X). The image fusion evaluation was conducted qualitatively and quantitatively by comparing fused results with their ground truth counterparts (Pohl, Moellmann, and Fries 2017).

The qualitative evaluation was based on the similarities between the fused 10-m/20-m and reference S2 10-m/20-m bands in terms of their clarity, structure similarity, spectral fidelity. Moreover, the image sharpening effects of the fused 10-m/20-m images over their 30-m/120-m counterparts were also examined. The quantitative evaluation was based on statistical metrics between fused and reference images from spectral and structural similarity perspectives. The widely used Correlation Coefficient (CC), Root Mean Square Error (RMSE), Average Absolute Difference (AAD), Erreur Relative Global Adimensionnelle de Synthèse (ERGAS; Wald 2002), and Spectral Angle Mapper (SAM; Yuhas, Goetz, and Boardman 1992) between fused and reference images were used to evaluate their spectral similarities. The Structure SIMilarity (SSIM) index (SSIM; Wang et al. 2004) and Universal Image Quality Index (UIQI; Wang and Bovik 2002) were adopted to assess their overall image structure and radiometric similarities. Additionally, the averaged AAD maps between the fused and reference bands were used to demonstrate the spatial distribution and magnitude of fusion errors. The scatter plots of fused bands against reference bands were also plotted to provide an intuitive comparison between fused and reference values in describing the approximate extent of their correlation (Zhao, Huang, and Song 2018).

The evaluation criteria for the fusion at degraded spatial resolutions are: (1) the closer CC, SSIM, and UIQI are to one, the more similar the predictions are to the references; (2) the closer RMSE, AAD, ERGAS, and SAM are to zero, the more similar the predictions are to the references; (3) the closer the scatter plots are to the 1:1 line, the more similar the predictions are to the references.

3.4.2. Metrics for fusion at original spatial resolutions

In addition to the fusion at the degraded resolution level with direct references, the fusion at the full resolution level (3-m PS and 10-m/20-m S2) was performed as well because (1) the practical target is to generate 3-m images in the ten bands and (2) the fusion strategy in Section 3.4.1 may not be applicable to VHR imagery, especially for heterogeneous landscapes (Alparone et al. 2008). We evaluated the spectral and spatial fidélities of results to the original S2 and PS bands, respectively, based on the criteria devised for spatial-spectral fusion without high spatial-spectral resolution references (Alparone et al. 2008; Song et al. 2014).

First, the spectral CC and spatial CC were used to evaluate the spectral and spatial distortions, respectively (Zhou, Civco, and Silander 1998). Specifically, spectral CCs were calculated between downsampled fusion results and original S2 images in band-by-band forms. Spatial CCs were calculated between high-frequency components (spatial details) of the fused and original PS bands (RGB + NIR). Note the high-frequency components were extracted by
a Laplacian high pass filter. Second, the spectral distortion index \(D_\lambda\), the spatial distortion index \(D_s\), and the joint spectral and spatial quality index, i.e. Quality with No Reference (QNR) (Alparone et al. 2008) were employed to evaluate the spectral or spatial distortions. Note \(D_\lambda\) was calculated based on all the ten bands before and after fusion, and \(D_s\) was calculated based on the four PS and fused bands. For more details about the calculation of the three indices, we refer readers to (Alparone et al. 2008).

The evaluation criteria for the fusion at original spatial resolutions are: (1) the spectral CC, spatial CC, and QNR are to one, the better the fusion results; (2) the closer \(D_\lambda\) and \(D_s\) are to zero, the better the fusion results.

### 3.5. Comparison Methods and Parameterization

We compared the proposed RASSFM method with twenty representative SSF algorithms that were developed for similar purposes, including the Color Resolution Improvement Software Package (CRISP; Winter et al. 2007), Unmixing-Based Data Fusion algorithm (UBDF; Zurita-Milla, Clevers, and Schaepman 2008), Dictionary-Pair Learning Method (DPLM; Song et al. 2014), High Pass Modulation (HPM; Li et al. 2020b), and sixteen well-established hypersharpening (HySharp) approaches presented by Vivone and Chanussot (2020). CRISP, UBDF, and DPLM utilize component substitution, spectral unmixing, and compressed sensing techniques, respectively, for spatial-spectral fusion, which have been widely used as benchmarks for image fusion algorithm development (Huang et al. 2013; Sun et al. 2014; Zhu et al. 2016). HPM has been recommended for the fusion of PS and S2 imagery because it introduces small spatial and spectral distortions in the sharpened S2 bands (Li et al. 2020b). The sixteen HySharp approaches (Vivone and Chanussot 2020) are eight well-developed SSF algorithms, i.e. the Gram-Schmidt (GS) approach, GS adaptive (GSA) method, Gauss High-Pass Modulation (Gauss-HPM), Gauss High-Pass Filtering (Gauss-HPF), Gauss Context-Based Decision (Gauss-CBD), Bilinear filter based HPM (Bil-HPM), Bil-HPF, and Bil-CBD, based on two high-spatial-resolution band generation schemes, i.e. band synthesizing and band selection (Selva et al. 2015). More details about the sixteen approaches can be found in (Vivone and Chanussot 2020). Two save space, we show the best results among the sixteen HySharp approaches (Table S1).

The parameterizations were different for the two types of fusion:(1) fusion at the degraded spatial resolution level with direct references and (2) fusion at the original spatial resolution level with indirect references. For the fusion at the degraded spatial resolutions, the low- and high-pass Butterworth filters in CRISP were both parameterized with the cutoff frequency of 15 and the order of 3 via fine-tuning; the moving window size and the class number in UBDF were set as 25 and 30, respectively; the number of spectral dictionary atoms and the sparseness degree of representation coefficients in DPLM were fine-tuned to be 10 and 0.8, respectively; the moving window size and the number of similar neighbors for RASSFM were set as 7 and 10, respectively. For the fusion at the original spatial resolutions, low- and high-pass Butterworth filters in CRISP were both parameterized with the cutoff frequency of 35 and the order of 3 via fine-tuning; the moving window size and the class number in UBDF were set as 31 and 40, respectively; the number of spectral dictionary atoms and the sparseness degree of the representation coefficients in DPLM were fine-tuned to be 12 and 0.8, respectively; the moving window size and the number of similar neighbors for RASSFM were set as 11 and 25, respectively.

### 4. Experimental results

#### 4.1. Fusion at degraded spatial resolutions

##### 4.1.1. Results of Site 1

This dataset is used to examine the performance of the SSF methods in terms of highly complex land covers with rich spectral characteristics. Showing the prior low-resolution (30-m/120-m) S2 images is to provide a benchmark to demonstrate the downscaling effects of the SSF methods (Figure 4 and S1). All methods improved the spatial resolutions of the 30-m/120-m S2 images significantly, showing comparable sharpness to the reference 10-m/20-m S2 images. However, the spatial distortions of UBDF are substantial compared to the other methods, which are prone to patch effects and lose local fine-resolution textures, particularly for the 20-m fusion scenario (see Figure 4&l).
As to visual spectral fidelities of the fusion at the 10-m resolution level (Figure S1), the proposed RASSFM method performs better than the other algorithms, such as the marked stadium in Figures S1e-j. Quantitatively, it can be clearly observed from the AAD maps (Figs. S1q-v) that RASSFM has the lowest fusion biases than the others. Furthermore, the statistical metrics in Table S2 and the scatter plots in Figure S9 explicitly demonstrate that RASSFM has lower fusion errors and higher spectral and structural similarities to the reference 10-m image.

For visual evaluations of the fusion at the 20-m resolution level (see Figure 4), UBDF and DPLM have larger spectral disparities to the references than the other methods. For instance, the vegetation in the UBDF and DPLM results (Figure 4&m) tend to be brighter and bluer, respectively, than the reference images. The DPLM results look more natural than the results of UBDF due to its superiority in preserving the spatial information of the prior PS image. The spectral fidelity of RASSFM is better than CRISP regarding the image saturation degrees compared to the reference images, especially for the RE bands (Figure 4&j). Moreover, RASSFM performs better than HPM regarding the bare land in the northwest of the study area (Figure 4&p). For HySharp, the results look darker than the reference images, such as the vegetation and bare lands (Figure 4&o). Quantitatively, the 20-m average AAD maps explicitly demonstrate that the RASSFM has lower fusion errors than the other methods. Furthermore, the statistical metrics in Table 3 and the scatter plots in Figure S10 indicate that RASSFM has the highest spectral and image structure similarities and lowest fusion biases to the reference 20-m image compared to the others.

### 4.1.2. Results of Site II

Site II is adopted to examine the fusion performance of the involved methods in desert urban areas. For spatial information enhancement, the input 30-m/120-m S2 images were obviously sharpened to 10-m/20-m (Figure 5 and S2). In addition, the spatial fidelity of UBDF is worse than the other methods. For spectral fidelity, the results indicate that DPLM performs less satisfying than the others, particularly for the SWIR bands. This is probably caused by the limited spectral constraints when resolving the representative coefficients of the PS spectral atoms for each PS pixel because PS only has four visual and NIR bands (no constraints for the SWIR bands),

---

**Figure 4.** Visual evaluation of the fusion at degraded spatial resolutions (120-m to 20-m) in Site I. The images (a, b) and (e) – (j) are displayed with RE3-RE2-RE1 bands as RGB composites, and the images (c, d) and (k) – (p) are displayed with SWIR2-SWIR1-NNIR bands as RGB composites. The images (q) – (v) represent the averaged AAD maps of the six fused and reference bands.
which may lead to biases by applying the coefficients to the corresponding S2 spectral atoms. CRISP, HPM, and HySharp have comparable visual performance for the 10-m fusion results, and RASSFM demonstrates the highest spectral similarity to the reference image (see Figures S2q-v). Moreover, the 20-m fusion results show the better spectral fidelity of RASSFM than the other methods obviously (Figure 5). Quantitatively, the evaluation metrics in Table 4 and S3, the scatter plots in Figures S11-S12, and the AAD maps in Figure 5 and S2 explicitly demonstrate that RASSFM has the lowest fusion errors compared to the other methods.

4.1.3. Results of Site III
This dataset is used to investigate the fusion effects of the SSF methods in heterogeneous agricultural landscapes. Qualitatively, the image sharpness effects of the SSF methods are obvious at both 10-m and 20-m fusion scenarios (see Figure 5 and 6 and S3). For instance, the blurry boundaries and textures of the croplands in the prior 30-m/120-m S2 images become much clearer in the fused 10-m/20-m images. Nevertheless, the spatial fidelity of UBDF is lower than the other methods. In addition to the sharpness enhancement, the spectral fidelities of the 10-m/20-m results fused by RASSFM is slightly better than CRISP, HPM, and HySharp by visual evaluations, and these four methods are significantly better than UBDF and DPLM, e.g. the SWIR bands (Figure 6&m). Quantitatively, the AAD maps explicitly demonstrate that the fusion biases of RASSFM are the smallest among all methods. Moreover, the statistical metrics in Table 5 and S4 and the scatter plots in Figures S13-S14 indicate that RASSFM has higher image spectral and spatial similarities and lower fusion errors with the reference bands than the compared methods. Note that since the spectral variances within each cropland are smaller than the variances in Sites I and II, the fusion performance of each method is generally better (except SWIRs).

4.1.4. Results of Site IV
This group of experiments further demonstrates the performance of our proposed and the compared SSF methods in vegetation-dominated landscapes, e.g. forest and lawns. For image sharpness, the spatial details of the input 30-m/120-m images were greatly improved by all methods (Figure 7 and S4), while differences still exist for different methods. For the 20-m fusion scenario, for example, CRISP- and UBDF-fused results have blurry (see Figure 7&m) and block (see Figure 7&l) effects, respectively. DPLM preserves the spatial details of the prior 10-m/20 m images better than CRISP and UBDF, but DPLM loses some textures regarding the vegetated areas (see Figure 7&m). HPM, HySharp, and RASSFM have comparable spatial textures, but their spectral fidelities to the references are quite different. For visual spectral fidelity, the performance of all methods at the 10-m fusion level is comparable because this is a relatively homogeneous landscape, but the AAD maps explicitly demonstrate that RASSFM has the smallest fusion biases (Figure S4). Moreover, the fusion at the 20-m level shows the superiority of RASSFM explicitly because it has the closest spectral characteristics to the reference images, such as the built-up and vegetation areas (Figure 7 -p). Quantitatively, the image fusion metrics in Table 6 and S5 and the scatter plots in Figures S15-S16 show that RASSFM has the largest spectral and spatial fidelities and the smallest fusion errors regarding the reference images.

4.2. Fusion at original spatial resolutions
4.2.1. Results of site I
To complement the image fusion validation in Section 4.1 at the degraded scales, the fusion at the original scales, i.e. 3-m PS and 10-m/20-m S2, is conducted to evaluate the fused high spatial-spectral resolution images at user ends. Note that the input 3-m PS and 20-m S2 images are adopted as spatial and spectral references, respectively, because the spatial and spectral contents of the fused bands come from the two data sources, respectively. In Figure 8 and S5, the spatial details of the PS images were injected into the S2 images well, and the spatial resolution improvement of the fused 3-m images are substantial compared with the original 10-m/20-m S2 images for all fusion methods, such as the playgrounds and buildings, which are obvious in both global and local zoomed-in views.

Furthermore, the spectral fidelity of RASSFM is better than the other methods (Figure 8), such as the distorted color of the outfield (fake grass) of the baseball field (location A) for the other methods. Moreover, the quantitative evaluation metrics demonstrate distinct differences. In Table 7, spectral
Table 3. The quantitative evaluation of the six fused and reference 20-m bands in Site I (bold numbers represent the best results).

| Method       | CC          | RMSE        | AAD          |
|--------------|-------------|-------------|--------------|
|              | RE1 | RE2 | RE3 | NNIR | SWIR1 | SWIR2 | RE1 | RE2 | RE3 | NNIR | SWIR1 | SWIR2 | RE1 | RE2 | RE3 | NNIR | SWIR1 | SWIR2 |
| CRISP        | 0.9481 | 0.9569 | 0.9622 | 0.9666 | 0.9368 | 0.9051 | 0.0221 | 0.0201 | 0.0210 | 0.0211 | 0.0224 | 0.0236 | 0.0137 | 0.0139 | 0.0151 | 0.0154 | 0.0157 | 0.0156 |
| UBDF         | 0.9445 | 0.9464 | 0.9479 | 0.9500 | 0.8285 | 0.7671 | 0.0214 | 0.0216 | 0.0237 | 0.0248 | 0.0369 | 0.0373 | 0.0149 | 0.0157 | 0.0176 | 0.0183 | 0.0262 | 0.0246 |
| DPLM         | 0.9726 | 0.9346 | 0.9535 | 0.9698 | 0.9107 | 0.8400 | 0.0254 | 0.0314 | 0.0306 | 0.0301 | 0.0320 | 0.0341 | 0.0151 | 0.0221 | 0.0221 | 0.0222 | 0.0228 | 0.0230 |
| HySharp (best of 16) | 0.9782 | 0.9750 | 0.9772 | 0.9807 | 0.9496 | 0.9200 | 0.0196 | 0.0168 | 0.0181 | 0.0181 | 0.0259 | 0.0276 | 0.0120 | 0.0114 | 0.0128 | 0.0131 | 0.0161 | 0.0163 |
| HPM          | 0.9318 | 0.9474 | 0.9549 | 0.9604 | 0.9045 | 0.8587 | 0.0252 | 0.0218 | 0.0226 | 0.0227 | 0.0274 | 0.0284 | 0.0163 | 0.0142 | 0.0157 | 0.0161 | 0.0190 | 0.0195 |
| RASSFM       | 0.9853 | 0.9871 | 0.9887 | 0.9889 | 0.9683 | 0.9321 | 0.0115 | 0.0112 | 0.0115 | 0.0136 | 0.0160 | 0.0203 | 0.0068 | 0.0081 | 0.0086 | 0.0099 | 0.0102 | 0.0116 |

| Method       | SSIM         | UIQI         | ERGAS        | SWIR2       |
|--------------|--------------|--------------|--------------|-------------|
|              | RE1 | RE2 | RE3 | NNIR | SWIR1 | SWIR2 | RE1 | RE2 | RE3 | NNIR | SWIR1 | SWIR2 |
| CRISP        | 0.9033 | 0.9213 | 0.9225 | 0.9303 | 0.8727 | 0.8563 | 0.9269 | 0.9460 | 0.9535 | 0.9600 | 0.9340 | 0.9011 | 2.1601 | 3.0003 |
| UBDF         | 0.8664 | 0.8878 | 0.8824 | 0.8859 | 0.7172 | 0.7093 | 0.9354 | 0.9383 | 0.9421 | 0.9460 | 0.8278 | 0.7668 | 2.8458 | 4.4366 |
| DPLM         | 0.9160 | 0.8610 | 0.8867 | 0.9080 | 0.8467 | 0.8356 | 0.8911 | 0.8343 | 0.8792 | 0.9018 | 0.8223 | 0.8047 | 2.9701 | 4.4402 |
| HySharp (best of 16) | 0.9422 | 0.9529 | 0.9527 | 0.9579 | 0.8914 | 0.8755 | 0.9010 | 0.9704 | 0.9727 | 0.9761 | 0.9333 | 0.9012 | 2.1817 | 2.9615 |
| HPM          | 0.8563 | 0.9073 | 0.9081 | 0.9157 | 0.8238 | 0.7995 | 0.9010 | 0.9365 | 0.9461 | 0.9536 | 0.9004 | 0.8501 | 2.4972 | 3.6329 |
| RASSFM       | 0.9698 | 0.9661 | 0.9678 | 0.9629 | 0.9348 | 0.9125 | 0.9824 | 0.9843 | 0.9869 | 0.9836 | 0.9679 | 0.9320 | 1.4694 | 2.4882 |
Figure 5. Visual evaluation of the fusion at degraded spatial resolutions (120-m to 20-m) in Site II. The images (a, b) and (e) – (j) are displayed with RE3-RE2-RE1 bands as RGB composites, and the images (c, d) and (k) – (p) are displayed with SWIR2-SWIR1-NNIR bands as RGB composites. The images (q) – (v) represent the averaged AAD maps of the six fused and reference bands.

Figure 6. Visual evaluation of the fusion at degraded spatial resolutions (120-m to 20-m) in Site III. The images (a, b) and (e) – (j) are displayed with RE3-RE2-RE1 bands as RGB composites, and the images (c, d) and (k) – (p) are displayed with SWIR2-SWIR1-NNIR bands as RGB composites. The images (q) – (v) represent the averaged AAD maps of the six fused and reference bands.
Table 4. The quantitative evaluation of the six fused and reference 20-m bands in Site II (bold numbers represent the best results).

| Method      | CC       | RMSE     | AAD       |
|-------------|----------|----------|-----------|
|             | RE1 RE2 RE3 NNIR SWIR1 SWIR2 | RE1 RE2 RE3 NNIR SWIR1 SWIR2 | RE1 RE2 RE3 NNIR SWIR1 SWIR2 |
| CRISP       | 0.9729 0.9771 0.9793 0.9814 0.9665 0.9400 | 0.0270 0.0241 0.0230 0.0217 0.0280 0.0320 | 0.0196 0.0177 0.0169 0.0161 0.0213 0.0232 |
| UBDF        | 0.9563 0.9550 0.9551 0.9560 0.8831 0.8231 | 0.0323 0.0316 0.0316 0.0314 0.0540 0.0588 | 0.0238 0.0235 0.0236 0.0232 0.0393 0.0413 |
| DPLM        | 0.9446 0.9818 0.9735 0.9536 0.8976 0.7845 | 0.0463 0.0341 0.0403 0.0481 0.0481 0.0668 | 0.0347 0.0287 0.0350 0.0406 0.0391 0.0500 |
| HySharp (best of 16) | 0.9765 0.9732 0.9765 0.9794 0.9752 0.9534 | 0.0242 0.0245 0.0231 0.0218 0.0253 0.0310 | 0.0176 0.0179 0.0169 0.0159 0.0186 0.0212 |
| HPM         | 0.9652 0.9617 0.9617 0.9629 0.9642 0.9451 | 0.0349 0.0364 0.0365 0.0361 0.0318 0.0313 | 0.0257 0.0267 0.0268 0.0265 0.0235 0.0224 |
| RASSFM      | 0.9885 0.9891 0.9911 0.9928 0.9831 0.9624 | 0.0169 0.0158 0.0144 0.0144 0.0201 0.0254 | 0.0127 0.0115 0.0106 0.0105 0.0147 0.0174 |
|             | SSIM     | UIQI     | ERGAS     | SAM       |
|             | RE1 RE2 RE3 NNIR SWIR1 SWIR2 | RE1 RE2 RE3 NNIR SWIR1 SWIR2 | RE1 RE2 RE3 NNIR SWIR1 SWIR2 | RE1 RE2 RE3 NNIR SWIR1 SWIR2 |
| CRISP       | 0.9298 0.9442 0.9494 0.9540 0.8688 0.8110 | 0.9665 0.9716 0.9745 0.9776 0.9660 0.9395 | 1.3682 | 2.0687 |
| UBDF        | 0.8798 0.8787 0.8787 0.8785 0.6572 0.5901 | 0.9546 0.9531 0.9535 0.9549 0.8823 0.8198 | 2.1535 | 3.9579 |
| DPLM        | 0.8417 0.9534 0.9324 0.8781 0.8274 0.7191 | 0.8914 0.9672 0.9642 0.9476 0.8886 0.7439 | 2.5250 | 5.9094 |
| HySharp (best of 16) | 0.9407 0.9342 0.9426 0.9490 0.8977 0.8507 | 0.9764 0.9730 0.9764 0.9794 0.9743 0.9498 | 1.3142 | 2.1856 |
| HPM         | 0.8790 0.8669 0.8648 0.8674 0.8677 0.8310 | 0.9384 0.9258 0.9256 0.9282 0.9503 0.9365 | 1.8045 | 2.4396 |
| RASSFM      | 0.9581 0.9646 0.9692 0.9714 0.9206 0.8640 | 0.9878 0.9887 0.9906 0.9903 0.9823 0.9617 | 0.9516 | 1.8830 |
Table 5. The quantitative evaluation of the six fused and reference 20-m bands in site III (bold numbers represent the best results).

| Method            | CC       | RMSE      | AAD       |
|-------------------|----------|-----------|-----------|
|                   | RE1      | RE2       | RE3       | NNIR     | SWIR1    | SWIR2    | RE1      | RE2       | RE3       | NNIR     | SWIR1    | SWIR2    | RE1      | RE2       | RE3       | NNIR     | SWIR1    | SWIR2    |
| CRISP             | 0.9731   | 0.9928    | 0.9937    | 0.9936    | 0.8859   | 0.9246   | 0.0051   | 0.0115    | 0.0147    | 0.0147   | 0.0082   | 0.0074   | 0.0037   | 0.0088    | 0.0112    | 0.0113    | 0.0049    | 0.0047   |
| UBDF              | 0.8892   | 0.9857    | 0.9875    | 0.9864    | 0.6891   | 0.8332   | 0.0100   | 0.0152    | 0.0194    | 0.0201   | 0.0149   | 0.0113   | 0.0068   | 0.0105    | 0.0136    | 0.0140    | 0.0088    | 0.0070   |
| DPLM (best of 16) | 0.9319   | 0.9930    | 0.9951    | 0.9938    | 0.7039   | 0.8046   | 0.0079   | 0.0115    | 0.0145    | 0.0160   | 0.0131   | 0.0123   | 0.0061   | 0.0092    | 0.0117    | 0.0130    | 0.0095    | 0.0089   |
| HySharp           | 0.9689   | 0.9932    | 0.9930    | 0.9928    | 0.6787   | 0.9191   | 0.0060   | 0.0131    | 0.0181    | 0.0183   | 0.0091   | 0.0088   | 0.0039   | 0.0098    | 0.0137    | 0.0139    | 0.0052    | 0.0055   |
| HPM               | 0.9721   | 0.9947    | 0.9957    | 0.9955    | 0.8769   | 0.9173   | 0.0052   | 0.0101    | 0.0124    | 0.0125   | 0.0085   | 0.0077   | 0.0037   | 0.0076    | 0.0095    | 0.0096    | 0.0053    | 0.0050   |
| RASSFM (best of 16) | 0.9837   | 0.9969    | 0.9976    | 0.9979    | 0.9201   | 0.9527   | 0.0039   | 0.0071    | 0.0084    | 0.0080   | 0.0069   | 0.0059   | 0.0027   | 0.0048    | 0.0058    | 0.0056    | 0.0039    | 0.0036   |

| Method            | SSIM     | UIQI      | ERGAS     | SAM       |
|-------------------|----------|-----------|-----------|-----------|
|                   | RE1      | RE2       | RE3       | NNIR     | SWIR1    | SWIR2    | RE1      | RE2       | RE3       | NNIR     | SWIR1    | SWIR2    |
| CRISP             | 0.9305   | 0.9461    | 0.9476    | 0.9453    | 0.8419   | 0.8229   | 0.9685   | 0.9913    | 0.9924    | 0.9924   | 0.8828   | 0.9242   | 1.0170   | 1.8278    |
| UBDF              | 0.7910   | 0.8844    | 0.8822    | 0.8740    | 0.6815   | 0.7183   | 0.8892   | 0.9853    | 0.9872    | 0.9862   | 0.6852   | 0.8327   | 1.5818   | 2.2301    |
| DPLM              | 0.9249   | 0.9678    | 0.9706    | 0.9677    | 0.7918   | 0.7677   | 0.9263   | 0.9913    | 0.9925    | 0.9907   | 0.6901   | 0.8038   | 1.4418   | 2.5628    |
| HySharp (best of 16) | 0.9287   | 0.9497    | 0.9424    | 0.9417    | 0.8302   | 0.8175   | 0.9644   | 0.9902    | 0.9899    | 0.9866   | 0.8767   | 0.9098   | 1.2078   | 2.2457    |
| HPM               | 0.9318   | 0.9577    | 0.9561    | 0.9537    | 0.8342   | 0.8198   | 0.9674   | 0.9934    | 0.9947    | 0.9945   | 0.8702   | 0.9165   | 0.9852   | 1.8294    |
| RASSFM (best of 16) | 0.9461   | 0.9693    | 0.9703    | 0.9751    | 0.8744   | 0.8734   | 0.9830   | 0.9969    | 0.9976    | 0.9978   | 0.9185   | 0.9526   | 0.7318   | 1.2331    |
CCs and $D_s$ with the S2 bands show that the spectral fidelity of RASSFM is higher than the others. In addition, the spectral CCs reveal that the spectral fusion accuracies of the other methods decrease significantly when it comes to the SWIR bands (beyond the PS spectra), but RASSFM keeps comparable accuracies. Furthermore, the spatial CCs and $D_s$ with the four PS bands indicate that the spatial information fidelity of RASSFM is better than the compared methods. Moreover, the joint spatial and spectral quality index, i.e. QNR, shows that RASSFM has the best fusion effects regarding the spatial and spectral fidelities.

**4.2.2. Results of Site II**

As shown in Figure 9 and S6, the urban objects are sharpened significantly from 10-m/20-m to 3-m resolution, but CRISP and RASSFM have higher spatial fidelities than UBDF, DPLM, HPM, and HySharp (see the zoom-in comparisons in Figure 9). For their visual spectral fidelities, the 10-m fusion results in the four bands are generally satisfying (Figure S6), but the 20-m fusion results in the six S2 bands show larger differences, indicating the better performance of RASSFM in preserving the spectral information of S2, such as the cylindrical buildings (Figure 9 -g). Quantitatively, the spectral (spectral CC and $D_s$) and spatial (spatial CC and $D_s$) fidelity evaluation metrics show that the RASSFM-based results have higher spectral and spatial fusion accuracies than the other methods (Table 8). The QNRs also indicate the balanced performance of RASSFM in preserving the spectral and spatial information from S2 and PS images, respectively.

**4.2.3. Results of Site III**

For the heterogeneous agricultural landscapes, all the SSF methods substantially improve the spatial resolutions of the original 10-m/20-m S2 bands, such as the boundaries of the cropland parcels and crowns (see Figure 10 and S7). Visually, the spectral fidelities of the results obtained from the 10-m S2 bands are better than the ones obtained from the 20-m S2 bands because finer spatial information can be provided from the 10-m S2 bands. The results fused by CRISP, UBDF, DPLM, and HPM have lower degrees of spectral similarity to the original S2 bands than the results produced by HySharp and RASSFM, which is more
Table 6. The quantitative evaluation of the six fused and reference 20-m bands in Site IV (bold numbers represent the best results).

| Method       | SSIM          | ERGAS         | SAM          |
|--------------|---------------|---------------|--------------|
|              | RE1 RE2 RE3   | RE1 RE3 NNIR SWIR1 SWIR2 | RE1 RE3 NNIR SWIR1 SWIR2 |
| CRISP        | 0.9139 0.8928 0.9001 0.7937 0.8411 | 0.9387 0.9187 0.9311 0.9301 0.8577 0.8915 | 1.0403 1.4667 |
| UBDF         | 0.6403 0.7552 0.7815 0.7917 0.5445 0.6844 | 0.7609 0.8213 0.8624 0.8671 0.5985 0.7854 | 1.8680 2.3971 |
| DPLM         | 0.9102 0.8176 0.8268 0.8494 0.6837 0.8515 | 0.8954 0.7809 0.8122 0.8307 0.5669 0.8448 | 1.4464 1.8568 |
| HySharp (best of 16) | 0.9253 0.9115 0.9162 0.9163 0.8240 0.8734 | 0.9501 0.9454 0.9525 0.9536 0.8879 0.9122 | 1.0951 1.4342 |
| RASSFM       | 0.9336 0.9203 0.9374 0.9435 0.8584 0.8761 | 0.9614 0.9608 0.9688 0.9674 0.9089 0.9254 | 0.8256 1.3830 |
significant for fusion at the 20-m level (Figure 10-g). Moreover, the spatial fidelity of RASSFM is better than HySharp, particularly for the boundaries of the croplands (see Figure 10g). Quantitatively, the spectral (spectral CC and $D_s$), spatial (spatial CC and $D_s$), and joint spatial-spectral (QNR) distortion evaluation metrics in Table 9 explicitly demonstrate that the results fused by RASSFM have the most balanced performance in preserving the spectral and spatial information from the S2 and PS images, respectively.

4.2.4. Results of Site IV

This dataset demonstrates the performance of the involved SSF methods in homogeneous landscapes dominated by vegetation. The spatial resolution improvements are obvious for all methods at both 10-m and 20-m fusion levels (see Figure 11 and S8). Additionally, the spatial evaluation metrics in Table 10, i.e. spatial CCs and $D_s$, indicate that the bands fused by RASSFM have better spatial consistency with the input 3-m PS image bands. For the
**Table 7.** The quantitative evaluation of the ten fused 3-m bands in Site I (bold numbers represent the best results).

| Method      | Spectral CC | Spatial CC |
|-------------|-------------|------------|
|             | B | G | R | RE1 | RE2 | RE3 | NIR | NNIR | SWIR1 | SWIR2 |
| CRISP       | 0.9918 | 0.9921 | 0.9919 | 0.9893 | 0.9897 | 0.9895 | 0.9899 | 0.9898 | 0.9791 | 0.9741 | 0.9837 | 0.9850 | 0.9840 | 0.9765 | 0.0275 | 0.0106 | 0.9622 |
| UBDF        | 0.9929 | 0.9948 | 0.9943 | 0.9883 | 0.9870 | 0.9877 | 0.9931 | 0.9890 | 0.9642 | 0.9506 | 0.9208 | 0.9354 | 0.9265 | 0.9233 | 0.0120 | 0.0156 | 0.9726 |
| DPLM        | 0.9924 | 0.9943 | 0.9934 | 0.9446 | 0.8953 | 0.9392 | 0.9980 | 0.9371 | 0.8407 | 0.6279 | 0.9791 | 0.9848 | 0.9819 | 0.9702 | 0.0834 | 0.0085 | 0.9088 |
| HySharp (best of 16) | 0.9971 | 0.9973 | 0.9977 | 0.9908 | 0.9914 | 0.9917 | 0.9980 | 0.9918 | 0.9907 | 0.9899 | 0.9700 | 0.9709 | 0.9725 | 0.9714 | 0.0146 | 0.0114 | 0.9742 |
| HPM         | 0.9925 | 0.9928 | 0.9933 | 0.9852 | 0.9849 | 0.9854 | 0.9939 | 0.9851 | 0.9797 | 0.9814 | 0.9632 | 0.9648 | 0.9671 | 0.9699 | 0.0229 | 0.0241 | 0.9535 |
| RASSFM      | **0.9990** | **0.9992** | **0.9992** | **0.9979** | **0.9979** | **0.9979** | **0.9987** | **0.9981** | **0.9968** | **0.9964** | **0.9944** | **0.9940** | **0.9950** | **0.9937** | **0.0095** | **0.0002** | **0.9903** |

**Table 8.** The quantitative evaluation of the ten fused 3-m bands in Site II (bold numbers represent the best results).

| Method      | Spectral CC | Spatial CC |
|-------------|-------------|------------|
|             | B | G | R | RE1 | RE2 | RE3 | NIR | NNIR | SWIR1 | SWIR2 |
| CRISP       | 0.9918 | 0.9907 | 0.9925 | 0.9931 | 0.9936 | 0.9937 | 0.9933 | 0.9936 | 0.9836 | 0.9759 | 0.9432 | 0.9522 | 0.9666 | 0.9645 | 0.0183 | 0.0105 | 0.9714 |
| UBDF        | 0.9933 | 0.9947 | 0.9948 | 0.9891 | 0.9896 | 0.9900 | 0.9957 | 0.9904 | 0.9830 | 0.9773 | 0.9240 | 0.9373 | 0.9426 | 0.9367 | 0.0200 | 0.0139 | 0.9663 |
| DPLM        | 0.9895 | 0.9913 | 0.9930 | 0.9033 | 0.9187 | 0.9211 | 0.9943 | 0.9193 | 0.9097 | 0.8381 | 0.9594 | 0.9661 | 0.9782 | 0.9768 | 0.0664 | 0.0136 | 0.9209 |
| HySharp (best of 16) | 0.9982 | 0.9984 | 0.9981 | 0.9935 | 0.9954 | 0.9952 | 0.9987 | 0.9952 | 0.9947 | 0.9910 | 0.9727 | 0.9747 | 0.9787 | 0.9798 | 0.0122 | 0.0077 | 0.9802 |
| HPM         | 0.9903 | 0.9915 | 0.9908 | 0.9885 | 0.8980 | 0.9892 | 0.9943 | 0.9892 | 0.9847 | 0.9835 | 0.9014 | 0.9434 | 0.9457 | 0.9658 | 0.0182 | 0.0315 | 0.9509 |
| RASSFM      | **0.9986** | **0.9988** | **0.9900** | **0.9982** | **0.9985** | **0.9986** | **0.9993** | **0.9987** | **0.9972** | **0.9962** | **0.9858** | **0.9887** | **0.9916** | **0.9885** | **0.0084** | **0.0011** | **0.9905** |
spectral fidelity, their visual differences at the 10-m fusion level are not as obvious as the previous three sites due to the homogeneity of this site (see Figures S8a-g), but the spectral metrics in Table 10, i.e. spectral CCs and $D_\lambda$, show the differences explicitly. Moreover, the fused results at the 20-m fusion level demonstrate larger visual differences (see Figure 11 -g), and the RASSFM-fused results have larger similarities to the original S2 images, which is also indicated by the spectral CCs and $D_\lambda$ in Table 10. For the joint spatial and spectral fidelities, the QNRs in Table 10 reveal that RASSFM performs better in balancing the two aspects than the others.

5. Discussion

The utilization of spatial neighborhood information to improve the accuracy of image fusion has been proved effective in Spatial-Temporal image Fusion (STF; Gao et al. 2021; Li et al. 2020a; Liu et al. 2019).
### Table 9. The quantitative evaluation of the ten fused 3-m bands in Site III (bold numbers represent the best results).

| Method         | Spectral CC | Spatial CC |
|----------------|-------------|------------|
|                | B           | G          | R          | RE1         | RE2         | RE3         | NIR         | NNIR        | SWIR1       | SWIR2       |
| CRISP          | 0.9906      | 0.9862     | 0.9898     | 0.9814      | 0.9963      | 0.9960      | 0.9957      | 0.9960      | 0.9014      | 0.9953      |
| UBDF           | 0.9869      | 0.9854     | 0.9937     | 0.9819      | 0.9974      | 0.9980      | 0.9980      | 0.9979      | 0.9406      | 0.9942      |
| DPLM           | 0.9785      | 0.9689     | 0.9744     | 0.9420      | 0.9891      | 0.9942      | 0.9953      | 0.9941      | 0.4820      | 0.5510      |
| HySharp (best of 16) | 0.9942  | 0.9957     | 0.9967     | 0.9918      | 0.9980      | 0.9992      | 0.9980      | 0.9945      | 0.9999      |            |
| HPM            | 0.9917      | 0.9902     | 0.9928     | 0.9826      | 0.9960      | 0.9959      | 0.9976      | 0.9959      | 0.9608      | 0.9807      |
| RASSFM         | **0.9982**  | **0.9978** | **0.9990** | **0.9972**  | **0.9996**  | **0.9966**  | **0.9997**  | **0.9996**  | **0.9957**  | **0.9950**  |

### Table 10. The quantitative evaluation of the ten fused 3-m bands in Site IV (bold numbers represent the best results).

| Method         | Spectral CC | Spatial CC |
|----------------|-------------|------------|
|                | B           | G          | R          | RE1         | RE2         | RE3         | NIR         | NNIR        | SWIR1       | SWIR2       |
| CRISP          | 0.9845      | 0.9858     | 0.9874     | 0.9830      | 0.9778      | 0.9800      | 0.9820      | 0.9781      | 0.9530      | 0.9764      |
| UBDF           | 0.9824      | 0.9739     | 0.9845     | 0.9563      | 0.9511      | 0.9649      | 0.9751      | 0.9635      | 0.9024      | 0.9527      |
| DPLM           | 0.9719      | 0.9713     | 0.9715     | 0.8769      | 0.8835      | 0.9187      | 0.9535      | 0.9131      | 0.4481      | 0.7723      |
| HySharp (best of 16) | 0.9907  | 0.9918     | 0.9934     | 0.9840      | 0.9817      | 0.9806      | 0.9911      | 0.9804      | 0.9921      | 0.9866      |
| HPM            | 0.9843      | 0.9848     | 0.9854     | 0.9794      | 0.9760      | 0.9782      | 0.9838      | 0.9760      | 0.9607      | 0.9760      |
| RASSFM         | **0.9972**  | **0.9973** | **0.9980** | **0.9972**  | **0.9962**  | **0.9969**  | **0.9975**  | **0.9967**  | **0.9942**  | **0.9969**  |

| B              | 0.9786      | 0.9850     | 0.9878      | 0.9855      | 0.9021      | 0.0010      | 0.9779      |            |            |            |
|----------------|-------------|------------|------------|-------------|-------------|-------------|------------|-------------|-------------|-------------|
| G              |             |            |            |             |             |             |            |            |             |             |
| R              | 0.9786      | 0.9850     | 0.9878      | 0.9855      | 0.9021      | 0.0010      | 0.9779      |            |            |             |
| RE1            |             |            |            |             |             |             |            |            |             |             |
| RE2            |             |            |            |             |             |             |            |            |             |             |
| RE3            |             |            |            |             |             |             |            |            |             |             |
| NIR            |             |            |            |             |             |             |            |            |             |             |
| NNIR           |             |            |            |             |             |             |            |            |             |             |
| SWIR1          |             |            |            |             |             |             |            |            |             |             |
| SWIR2          |             |            |            |             |             |             |            |            |             |             |
Obtaining similar neighbors and determining their weights accurately are the keys. Compared with the traditional neighborhood utilization scheme in STF (Gao et al. 2006), the non-local constrained linear regression strategy adopted in this study has several advantages (Zhao, Huang, and Song 2018). Firstly, the image patch based searching scheme considers both spatial and spectral similarities between the target image patch and its similar neighbors to guarantee the reliability of the searched candidates, which is more robust than traditional pixel-based searching schemes that are prone to be affected by noises or outliers in satellite images (Gao et al. 2006). Secondly, the local constrained linear regression can maintain the local spatial structure of the prior VHR PS imagery, and the non-negative weights are less sensitive to image noises (Saul and Roweis 2003). Hence, the local constrained linear regression is more rigorous than arbitrary weight schemes such as inverse distance weighting (Chang 2006).

Generally, there is a tradeoff between the image sharpness and spectral fidelity of the fusion results. The moving window size and the similar neighbor number of RASSFM are responsible for the tradeoff, which would affect the similar neighbor searching and weight determination, respectively. The smaller the moving window size, the blurrier but the higher the spectral fidelity of the results, and vice versa. The fewer the similar neighbors, the blurrier but the higher spectral fidelity of the results, and vice versa. The parameters used in this study are determined through many trial-and-error experiments to achieve a good balance between image sharpness and spectral fidelity.

Furthermore, all fusion algorithms perform better in the four 10-m S2 bands than the six 20-m S2 bands. The different spectral and spatial disparities between PS and S2 account for that: (1) the spectral characteristics of the four PS bands are more similar to the four 10-m S2 bands than the six 20-m S2 bands, (2) the spatial resolution of the four S2 bands (10-m) is higher than the six S2 bands (20-m). Additionally, all methods have larger fusion errors in the SWIR bands than in the RE and NNIR bands in all study areas. The reason is that the SWIR bands overshoot the spectra of PS sensors, which have larger spectral differences to the four PS bands than the RE and NNIR bands, thereby leading to more challenges in extrapolating the spectral information of the SWIR bands (Winter et al. 2007). Moreover, the spectral fidelities of the results in Sites III and IV are better than the ones in Sites I and II (except SWIRs) as the land covers of Sites III and IV are less heterogeneous than Sites I and II. Therefore, small spatial resolution gaps and large spectral correlation between PS and S2, and high homogeneity of land covers will benefit the fusion accuracy.

As to efficiency, the computing time of each method in each study area was examined based on MATLAB® 2021b in a ThinkPad X1 carbon laptop equipped with an Intel® Core i7-1185G7 CPU @ 3.00 GHz and 32 GB of installed RAM (Table S6). It indicates that UBDF is the most time-consuming due to the unmixing process for all S2 pixels in each band. RASSFM and DPLM are the second and third time-consuming ones because of the similar neighbor searching and weight calculation, and the spectral dictionary learning and sparse coding. CRISP needs more time than HySharp and HPM due to the Fourier and inverse Fourier transformations. HySharp and HPM have similar efficiency due to their comparable computation burden.

Remote sensing image fusion is a broad topic, which mainly includes spatial-temporal fusion (Gao et al. 2006), spatial-spectral fusion (Selva et al. 2015), spatial-angular fusion (Huang, Zhang, and Yu 2012), spectral resolution enhancement (Sun et al. 2014), and spatial resolution enhancement (Zhao et al. 2021). Additionally, SSF covers the traditional pansharpening and the relatively new hypersharpening. We define this study as a type of SSF because we aim to fuse VHR images with limited spectral bands (PS) and med-resolution images with more bands (S2), and there are no temporal changes between the two data sources. Additionally, this study goes beyond the scope of pansharpening as no high-spatial-resolution panchromatic bands were involved. It is also slightly different from hypersharpening due to the absence of hyperspectral imagery. On the other hand, although utilizing spatial neighborhood information in RASSFM was motivated from previous STF methods (Gao et al. 2006; Zhao, Huang, and Song 2018) to improve the accuracy and robustness of SSF, their fusion purposes are quite different. STF aims to blend coarse-resolution (e.g. 250-m) satellite imagery with a frequent revisiting cycle (e.g. daily) and fine-resolution (e.g. 30-m) satellite imagery with a sparse revisiting cycle (e.g. 16 days) to produce fine-resolution images with the frequent revisiting cycle. The fine-but-sparse and coarse-but-frequent inputs of STF have the same band configurations. The trickiest thing of STF is downscaling the coarse-resolution change information to fine-resolution,
which basically does not exist in SSF because its data sources are acquired at the same or very close time. SSF deals with spectra downscaling rather than change downscaling. Note that although UBDF was originally proposed for SSF purposes (Zurita-Milla, Clevers, and Schaepman 2008), afterward it was revised for STF because the unmixing theory is applicable to both fusion scenarios (Li et al. 2020a; Zhu et al. 2016).

6. Conclusions

This study presents a new SSF method, i.e. RASSFM, for fusing the spatial details of PS and the spectral signatures of S2 to produce synthetic VHR (3-m) images in the ten bands (B, G, R, RE1&2&3, NIR, NNIR, SWIR1&2). Specifically, we proposed to combine the spectral mapping and the spectral correlation to generate VHR bands to sharpen their S2 counterparts. Moreover, the spatial neighborhood information of a target pixel considering spatial and spectral constraints was introduced to spatial-spectral fusion to improve the spectral fidelities of the results. We evaluated RASSFM qualitatively and quantitatively in different landscapes based on the fusion at the degraded and original spatial resolutions. The results indicate that RASSFM has robust and adaptive performance across different landscapes and outperforms the other methods in preserving the spatial details of PS and the spectral information of S2. This study blends the advantages of existing VHR and med-resolution satellite observations to resolve scientific questions that cannot be addressed by either data source. Additionally, RASSFM enriches the SSF community for blending PS and S2 imagery. Although RASSFM is proposed based on PS and S2 imagery, it can be used to fuse images from other satellites with similar spatial characteristics and spectral responses to produce synthetic images with high spatial-spectral resolutions.

The temporal resolution of the outputs of RASSFM is the same as S2 (five days) because PS and S2 collect data in daily and five-day intervals, respectively. Moreover, unlike some VHR satellites with limited observing areas per users’ request, both PS and S2 regularly observe the entire Earth. Therefore, fusing the advantages of the two sensors can generate 3-m image products of the ten S2 bands every five days in a global scope. Along with the development of the PS constellation and the accumulation of PS and S2 data archives, the utility of RASSFM will be highlighted in the context that increasing remote sensing applications require ample spatial and spectral information, such as precision agriculture, forest structure monitoring, and artificial surface change tracking.

Although the temporal resolution of the fusion results is quite frequent, not all daily PS images are fully utilized. In fact, for PS observations that do not have corresponding S2 imagery, the six 3-m bands (i.e. RE1&2&3, NNIR, SWIR1&2) cannot be generated by RASSFM because the temporal change information in the six bands is not available. Our future study will be pursuing a spectral enhancement algorithm that can augment the spectral resolution of PS images without S2 counterparts.

Acknowledgements

This study was supported by the United States Department of Agriculture under the National Institute of Food and Agriculture Capacity Building Grant # 2017-38821-26419.

Disclosure statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported by the United States Department of Agriculture.

ORCID

Yongquan Zhao http://orcid.org/0000-0002-8360-6731
Desheng Liu http://orcid.org/0000-0002-6088-5985

Data availability statement

The data that support the findings of this study are openly available in Planet Explorer at https://www.planet.com/explorer/.

References

Alparone, L., B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva. 2008. “Multispectral and Panchromatic Data Fusion Assessment without Reference.” Photogrammetric Engineering and Remote Sensing 74 (2): 193–200. doi:10.14358/PERS.74.2.193.
Ballester, C., V. Caselles, L. Igual, J. Verdera, and B. Rougé. 2006. “A Variational Model for P+ X Image Fusion.” *International Journal of Computer Vision* 69 (1): 43–58. doi:10.1007/s11263-006-6852-x.

Carper, W., T. Lillesand, and R. Klefer. 1990. “The Use of Intensity-hue-saturation Transformations for Merging SPOT Panchromatic and Multispectral Image Data.” *Photogrammetric Engineering and Remote Sensing* 56: 459–467.

Chang, K. T. 2006. “Spatial Interpolation.” In *Introduction to Geographic Information Systems*, edited by K. T. Chang. McGraw-Hill Higher Education Boston. pp. 315–342.

Emelyanova, I. V., T. R. McVicar, T. G. Van Niel, L. T. Li, and A. I. van Dijk. 2013. “Assessing the Accuracy of Blending Landsat–MODIS Surface Reflectances in Two Landscapes with Contrasting Spatial and Temporal Dynamics: A Framework for Algorithm Selection.” *Remote Sensing of Environment* 133: 193–209. doi:10.1016/j.rse.2013.02.007.

Feng, D., C. J. Gleason, X. Yang, and T. M. Pavelsky. 2019. “Comparing Discharge Estimates Made via the BAM Algorithm in High-order Arctic Rivers Derived Solely from Optical CubeSat, Landsat, and Sentinel-2 Data.” *Water Resources Research* 55 (9): 7753–7771. doi:10.1029/2019WR025599.

Forkuor, G., K. Dimobe, I. Serme, and J. E. Tondoh. 2018. “Landsat-8 Vs. Sentinel-2: Examining the Added Value of Sentinel-2’s Red-edge Bands to Land-use and Land-cover Mapping in Burkina Faso.” *IScience & Remote Sensing* 55 (3): 331–354. doi:10.1080/15481603.2017.1370169.

Gao, F., J. Masek, M. Schwaller, and F. Hall. 2006. “On the Blending of the Landsat and MODIS Surface Reflectance: Predicting Daily Landsat Surface Reflectance.” *IEEE Transactions on Geoscience and Remote Sensing* 44 (8): 2207–2218. doi:10.1109/TGRS.2006.872081.

Gao, H., X. Zhu, Q. Guan, X. Yang, Y. Yao, W. Zeng, and X. Peng. 2021. “cuFSDAF: An Enhanced Flexible Spatiotemporal Data Fusion Algorithm Parallelized Using Graphics Processing Units.” *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–16. doi:10.1109/TGRS.2021.3080384.

Gašparović, M., D. Medak, I. Pilaš, L. Jurjević, and I. Balenović. 2018. “Fusion of Sentinel-2 and PlanetScope Imagery for Vegetation Detection and Monitoring.” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLI-1: 155–160. doi:10.5194/isprs-archives-XLII-1-155-2018.

Gevaert, C. M., and F. J. Garcia-Haro. 2015. “A Comparison of STARFM and an Unmixing-based Algorithm for Landsat and MODIS Data Fusion.” *Remote Sensing of Environment* 156: 34–44. doi:10.1016/j.rse.2014.09.012.

Houborg, R., and M. F. McCabe. 2018. “A Cubesat Enabled Spatio-temporal Enhancement Method (Cestem) Utilizing Planet, Landsat and Modis Data.” *Remote Sensing of Environment* 209: 211–226. doi:10.1016/j.rse.2018.02.067.

Huang, B., H. Song, H. Cui, J. Peng, and Z. Xu. 2013. “Spatial and Spectral Image Fusion Using Sparse Matrix Factorization.” *IEEE Transactions on Geoscience and Remote Sensing* 52 (3): 1693–1704. doi:10.1109/TGRS.2013.2253612.

Huang, B., H. Zhang, and L. Yu. 2012. “Improving Landsat ETM+ Urban Area Mapping via Spatial and Angular Fusion with MISR Multi-angle Observations.” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5 (1): 101–109. doi:10.1109/JSTARS.2011.2177247.

Kimm, H., K. Guan, C. Jiang, B. Peng, L. F. Gentry, S. C. Wilkin, S. Wang, Y. Cai, C. J. Bernacchi, and J. Peng. 2020. “Deriving High-spatiotemporal-resolution Leaf Area Index for Agroecosystems in the US Corn Belt Using Planet Labs CubeSat and STAIR Fusion Data.” *Remote Sensing of Environment* 239: 111615. doi:10.1016/j.rse.2019.111615.

Li, X., G. M. Foody, D. S. Boyd, Y. Ge, Y. Zhang, Y. Du, and F. Ling. 2020a. “SFSDAF: An Enhanced FSDF that Incorporates Sub-pixel Class Fraction Change Information for Spatio-temporal Image Fusion.” *Remote Sensing of Environment* 237: 111537. doi:10.1016/j.rse.2019.111537.

Li, W., J. Jiang, T. Guo, M. Zhou, Y. Tang, Y. Wang, Y. Zhang, T. Cheng, Y. Zhu, and W. Cao. 2019b. “Generating Red-Edge Images at 3 M Spatial Resolution by Fusing Sentinel-2 and Planet Satellite Products.” *Remote Sensing* 11 (12): 1422. doi:10.3390/rs11121422.

Li, Q., F. K. K. Wong, and T. Fung. 2019a. “Classification of Mangrove Species Using Combined WordView-3 and LiDAR Data in Mai Po Nature Reserve, Hong Kong.” *Remote Sensing* 11 (18): 2114. doi:10.3390/rs11182114.

Li, Z., H. K. Zhang, D. P. Roy, L. Yan, and H. Huang. 2020b. “Sharpening the Sentinel-2 10 and 20 M Bands to PlanetScope-0 3 M Resolution.” *Remote Sensing* 12 (15): 2406. doi:10.3390/rs12152406.

Liu, J. 2000. “Smoothing Filter-based Intensity Modulation: A Spectral Preserve Image Fusion Technique for Improving Spatial Details.” *International Journal of Remote Sensing* 21 (18): 3461–3472. doi:10.1080/014311600750037499.

Liu, M., W. Yang, X. Zhu, J. Chen, X. Chen, L. Yang, and E. H. Helmer. 2019. “An Improved Flexible Spatiotemporal Data Fusion (IFSDAF) Method for Producing High Spatiotemporal Resolution Normalized Difference Vegetation Index Time Series.” *Remote Sensing of Environment* 227: 74–89. doi:10.1016/j.rse.2019.03.012.

Masi, G., D. Cozzolino, L. Verdoliva, and G. Scarpa. 2016. “Pan-sharpening by Convolutional Neural Networks.” *Remote Sensing* 8 (7): 594. doi:10.3390/rs8070594.

Niroumand-Jadidi, M., F. Bovolo, L. Bruzzone, and P. Gege. 2020. “Physics-based Bathymetry and Water Quality Retrieval Using PlanetScope Imagery: Impacts of 2020 COVID-19 Lockdown and 2019 Extreme Flood in the Venice Lagoon.” *Remote Sensing* 12 (15): 2381. doi:10.3390/rs12152381.

Otunga, C., J. Odindi, O. Mutanga, and C. Adjorlolo. 2019. “Evaluating the Potential of the Red Edge Channel for C3 (Festuca Spp.) Grass Discrimination Using Sentinel-2 and Rapid Eye Satellite Image Data.” *Geocarto International* 34 (10): 1123–1143. doi:10.1080/10106049.2018.1474274.

Page, B. P., A. Kumar, and D. R. Mishra. 2018. “A Novel Cross-satellite Based Assessment of the Spatio-temporal Development of A Cyanobacterial Harmful Algal Bloom.” *International Journal of Applied Earth Observation and Geoinformation* 69: 69–81. doi:10.1016/j.jag.2017.11.003.
Zhao, Y., B. Huang, and H. Song. 2018. “A Robust Adaptive Spatial and Temporal Image Fusion Model for Complex Land Surface Changes.” Remote Sensing of Environment 208: 42–62. doi:10.1016/j.rse.2018.02.009.

Zhao, Y., D. Liu, and X. Wei. 2020. “Monitoring Cyanobacterial Harmful Algal Blooms at High Spatiotemporal Resolution by Fusing Landsat and MODIS Imagery.” Environmental Advances 2: 100008. doi:10.1016/j.envadv.2020.100008.

Zhao Y., and Zhu Z. 2022. “ASI: An artificial surface Index for Landsat 8 imagery.” International Journal of Applied Earth Observation and Geoinformation 107: 102703. doi:10.1016/j.jag.2022.102703.

Zhou, J., D. Civco, and J. Silander. 1998. “A Wavelet Transform Method to Merge Landsat TM and SPOT Panchromatic Data.” International Journal of Remote Sensing 19 (4): 743–757. doi:10.1080/014311698215973.

Zhu, X., E. H. Helmer, F. Gao, D. Liu, J. Chen, and M. A. Lefsky. 2016. “A Flexible Spatiotemporal Method for Fusing Satellite Images with Different Resolutions.” Remote Sensing of Environment 172: 165–177. doi:10.1016/j.rse.2015.11.016.

Zurita-Milla, R., J. G. Clevers, and M. E. Schaepman. 2008. “Unmixing-based Landsat TM and MERIS FR Data Fusion.” IEEE Geoscience and Remote Sensing Letters 5 (3): 453–457. doi:10.1109/LGRS.2008.919685.