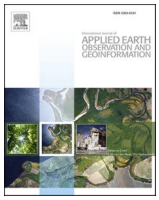
International Journal of Applied Earth Observations and Geoinformation 108 (2022) 102738

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Fusing Landsat 8 and Sentinel-2 data for 10-m dense time-series imagery using a degradation-term constrained deep network

Jingan Wu [a], Liupeng Lin [b], Tongwen Li [a], Qing Cheng [c], Chi Zhang [d], Huanfeng Shen [b,e,*]

[a] School of Geospatial Engineering and Science, Sun Yat-sen University, Zhuhai 519082, China
[b] School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China
[c] School of Computer Science, China University of Geosciences, Wuhan 430079, China
[d] Guangzhou Urban Planning & Design Survey Research Institute, Guangzhou 510060, China
[e] Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China

ARTICLE INFO

Keywords:
10-m time-series imagery
Spatiotemporal fusion
Landsat 8
Sentinel-2
Image degradation

ABSTRACT

Dense medium-resolution imagery is essential for fine-scale time-series applications. The combined use of Landsat 8 and Sentinel-2 can derive 10-m time-series imagery at a nominal temporal resolution of ~ 2.9 days. Specifically, Landsat images can be downscaled to a 10-m resolution by fusing them with temporally adjacent Sentinel-2 images. Current approaches simply use a linear model or a shallow network that is insufficient to obtain the complex mapping between inputs and outputs, and they rarely consider the temporal variation issue, especially for scenes experiencing land cover changes. Facing these limitations, we proposed a degradation-term constrained spatiotemporal fusion network (DSTFN). Technically, a deep network architecture incorporating residual dense blocks and attention mechanism modules is adopted to enhance the feature representation and extraction. A degradation constraint term is embedded into the loss function to maximize the use of the input coarse-resolution image and improve the capability of predicting change. A series of experiments based on two new datasets indicate that DSTFN achieves the best quantitative scores in every test and is thus effective and robust. In 20 resolution-degraded tests, on average, DSTFN decreases the mean relative error by 0.85%–5.35% and increases the peak signal-to-noise ratio 0.97–6.23 relative to baseline approaches. The tests featuring diverse temporal dynamics also confirm the strong generalization ability of DSTFN to deal with land cover change. The proposed network can be used to produce 10-m dense time-series imagery and shows great promise for a variety of time-series analyses and applications. The test materials are expected to be employed as standard datasets for future model assessment.

1. Introduction

Medium-resolution satellites usually observe the Earth's surface at a spatial resolution of tens of meters and show a superior capability to characterize the spatial structures of ground features relative to low-resolution satellites; thus, they are important for global and regional remote sensing applications (Gutman et al., 2008). Among the various medium-resolution satellites, Landsat 8 and Sentinel-2 are flagship missions because of their outstanding quality, global coverage, and free access to data archives (Woodcock et al., 2008). The Landsat family has been acquiring Earth observation data since 1972 (Irons et al., 2012), and the current Landsat 8 carries the Operational Land Imager (OLI) that records reflective signals from visible to shortwave infrared ranges at a 30-m resolution (Roy et al., 2014). As a part of the European Copernicus Project, Sentinel-2 is composed of two on-orbit satellites and collects multispectral imagery at 10/20/60-m resolutions depending on specific bands (Drusch et al., 2012). The medium-resolution imagery acquired by Landsat 8 and Sentinel-2 has been employed in a variety of applications, such as mapping land covers (Sánchez-Espinosa and Schröder, 2019), estimating biophysical properties (Korhonen et al., 2017), and evaluating disaster risks (Roy et al., 2019). Although the two representative missions are widely employed, a single instrument usually observes at a limited frequency that is inadequate to detect rapid changes (Pan et al., 2021), especially over dynamic landscapes. For example, Sentinel-2 revisits the same region every 5 days with twin satellites, but the authentic time gap of usable images is lengthened

* Corresponding author at: School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China.
E-mail address: shenhf@whu.edu.cn (H. Shen).

https://doi.org/10.1016/j.jag.2022.102738
Received 30 December 2021; Received in revised form 18 February 2022; Accepted 5 March 2022
Available online 18 March 2022

Table 1
Comparison of specifications between Sentinel-2A MSI and Landsat 8 OLI.

		Sentinel-2A MSI	Landsat 8 OLI
Swath width		290 km	185 km
Revisit cycle		10 days	16 days
Field of view		20.6°	15°
Equator crossing time		10:30 a.m.	10:13 a.m.
Band configuration and spatial resolution	Coastal	B01: 433–453 nm (10 m)	b1: 430–450 nm (30 m)
	Blue	B02: 458–523 nm (10 m)	b2: 450–515 nm (30 m)
	Green	B03: 543–578 nm (10 m)	b3: 525–600 nm (30 m)
	Red	B04: 650–680 nm (10 m)	b4: 630–680 nm (30 m)
	Red Edge 1	B05: 698–713 nm (20 m)	——
	Red Edge 2	B06: 733–748 nm (20 m)	——
	Red Edge 3	B07: 773–793 nm (20 m)	——
	NIR	B08: 785–900 nm (10 m)	——
	NIR-narrow	B8A: 855–875 nm (20 m)	b5: 845–885 nm (30 m)
	Water vapor	B09: 935–955 nm (60 m)	——
	Cirrus	B10: 1360–1390 nm (60 m)	b9: 1360–1390 nm (30 m)
	SWIR-1	B11: 1565–1655 nm (20 m)	b6: 1560–1660 nm (30 m)
	SWIR-2	B12: 2100–2280 nm (20 m)	b7: 2100–2300 nm (30 m)
	PAN	——	b8: 503–676 nm (15 m)

because of frequent cloud covers (Ju and Roy, 2008; Shen et al., 2019); hence, the capability of Sentinel-2 to reveal land surface dynamics is reduced. Besides, the 5-day revisit frequency hinders the monitoring of rapid changes that occur on a daily basis. Given the specification similarities of Landsat 8 and Sentinel-2 described in Table 1, an increasing number of studies have started to cooperatively use Earth observations from the two missions to densify the imagery (Chen et al., 2021b; Pan et al., 2021) and improve the ability to detect surface changes. According to Li and Roy (2017), considering Landsat 8 and Sentinel-2 together can lead to a median average interval of ~ 2.9 days.

In addition to geometric registration, bandpass adjustment, and spectral response harmonization, the combined use of Landsat 8 and Sentinel-2 requires the coordination of spatial resolution (Shao et al., 2019). A number of studies have upscaled Sentinel-2 imagery to 30 m to match it with that of Landsat 8 (Dong et al., 2020; Liu et al., 2020). For instance, the Harmonized Landsat and Sentinel-2 (HLS) project (Claverie et al., 2018), a NASA initiative, produces a harmonized 30-m surface reflectance dataset by resampling the Sentinel-2 10/20-m bands to 30-m. The "degraded resolution" scheme is generally easy to implement, but it sacrifices the fine spatial textures and limits the possibility of precisely mapping complex landscapes.

Different from the former option, the "enhanced resolution" scheme, i.e., enhancing Landsat 8 imagery to 10 m, is highly attractive. Li et al. (2017) downscaled the Landsat 30-m bands to 15 m by utilizing the panchromatic band and then resampled the result into registration with Sentinel-2 20-m bands. Pouliot et al. (2018) and Chen et al. (2021a) super-resolved Landsat 8 by using a convolutional neural network (CNN) and generative adversarial network (GAN). These works reconstructed the resolution-improved imagery by utilizing signals only from the Landsat source, and the results, although enhanced, have limited capacities to depict details relative to observed Sentinel-2 scenes (Pouliot et al., 2018). Given the deficiency, Sentinel-2 imagery temporally adjacent to the desired Landsat imagery is introduced as auxiliary data to obtain a robust output. The idea of merging multisource multitemporal imagery is similar to the spatiotemporal fusion concept originating from MODIS–Landsat fusion (Ma et al., 2021; Zhu et al., 2018). Various fusion models, such as the spatial and temporal adaptive reflectance fusion model (Gao et al., 2006), have been developed, but they require coincident fine- and coarse-resolution image pairs as input, which is challenging to satisfy for Landsat and Sentinel-2 because of their rarely matching acquisition dates. In our previous work (Wu et al., 2020), we combined an image simulation procedure with a spatiotemporal fusion model to present an initial solution. Agapiou (2020) exploited the pan-sharpening methods for this mission, and the derived 10-m results enhanced the image segmentation performance. Recently, approaches specifically for Landsat 8 and Sentinel-2 have also emerged. Wang et al. (2017) extended a geostatistical method called area-to-point regression Kriging (ATPRK) to downscale the Landsat 8 data to a 10-m resolution. Inspired by the advances in deep learning, Shao et al. (2019) and Luo et al. (2021) respectively presented an extended super-resolution convolutional neural network (ESRCNN) and a fusion generative adversarial network (FusGAN) to achieve data fusion. In addition to surface-reflectance-based studies, some cases aimed to construct 10-m normalized difference vegetation index (NDVI) time series by utilizing deep learning architecture (Ao et al., 2021; Bhogendra and Tej Bahadur, 2021).

Although attempts have been made to develop fusion models, several issues remain. First, as the core of fusion models, the mapping between inputs and outputs is statistically complex, and a simple linear model or a shallow network from previous studies (Shao et al., 2019; Wu et al., 2020) may be insufficient to obtain this mapping. Second, the existing deep-learning-based approaches rarely notice the inherent nature of the fusion task, such as the robust prediction for diverse temporal changes. Third, the scientific community needs benchmark datasets for Landsat 8 and Sentinel-2 fusion, and so, model validation cannot be performed fairly.

Given the limitations, we proposed a degradation-term constrained spatiotemporal fusion network (DSTFN) to fuse Landsat 8 and Sentinel-2 surface reflectance products and derive a 10-m dense time series. The model incorporates residual dense blocks with attention mechanism modules to enhance feature-level image fusion. A degradation constraint term is integrated into the loss function to enhance the prediction capacity for scenarios with abrupt changes. Two benchmark datasets featuring various levels of spatial heterogeneity and temporal dynamics are provided. On the basis of these datasets, our model is comprehensively evaluated against four baseline methods, including linear-weighting-based, geostatistical, and deep-learning-based methods.

2. Method

2.1. Two-stage data fusion framework

DSTFN is a deep-learning-based model that is proposed herein to merge Landsat 8 and Sentinel-2 observations and produce 10-m time-series imagery. Specifically, a 30-m Landsat image is downscaled to a 10-m resolution by fusing it with a temporally neighboring Sentinel-2 imagery. DSTFN considers the blue, green, red, near-infrared (NIR), and shortwave infrared (SWIR-1/2) bands that are common to both sensors and are widely accepted in applications. Fig. 1 illustrates the structure of DSTFN. The fusion framework is composed of two stages. The first stage accounts for the resolution disparity of Sentinel-2 bands and downscales the 20-m bands (B8A and B11–B12) to 10 m by taking the 10-m bands (B02–B04 and B08) as auxiliary data. The second stage coordinates the resolutions of Landsat and Sentinel-2 imagery and downscales the 30-m Landsat imagery (b2–b7) to 10 m. This stage uses the 10-m Sentinel-2 imagery (B02–B04, B8A, and B11–B12) temporally neighboring to the desired Landsat imagery as auxiliary data. The 15-m Landsat panchromatic band is included as auxiliary data to bridge the resolution gap between the two instruments, as suggested by previous studies (Shao et al., 2019; Wang et al., 2017).

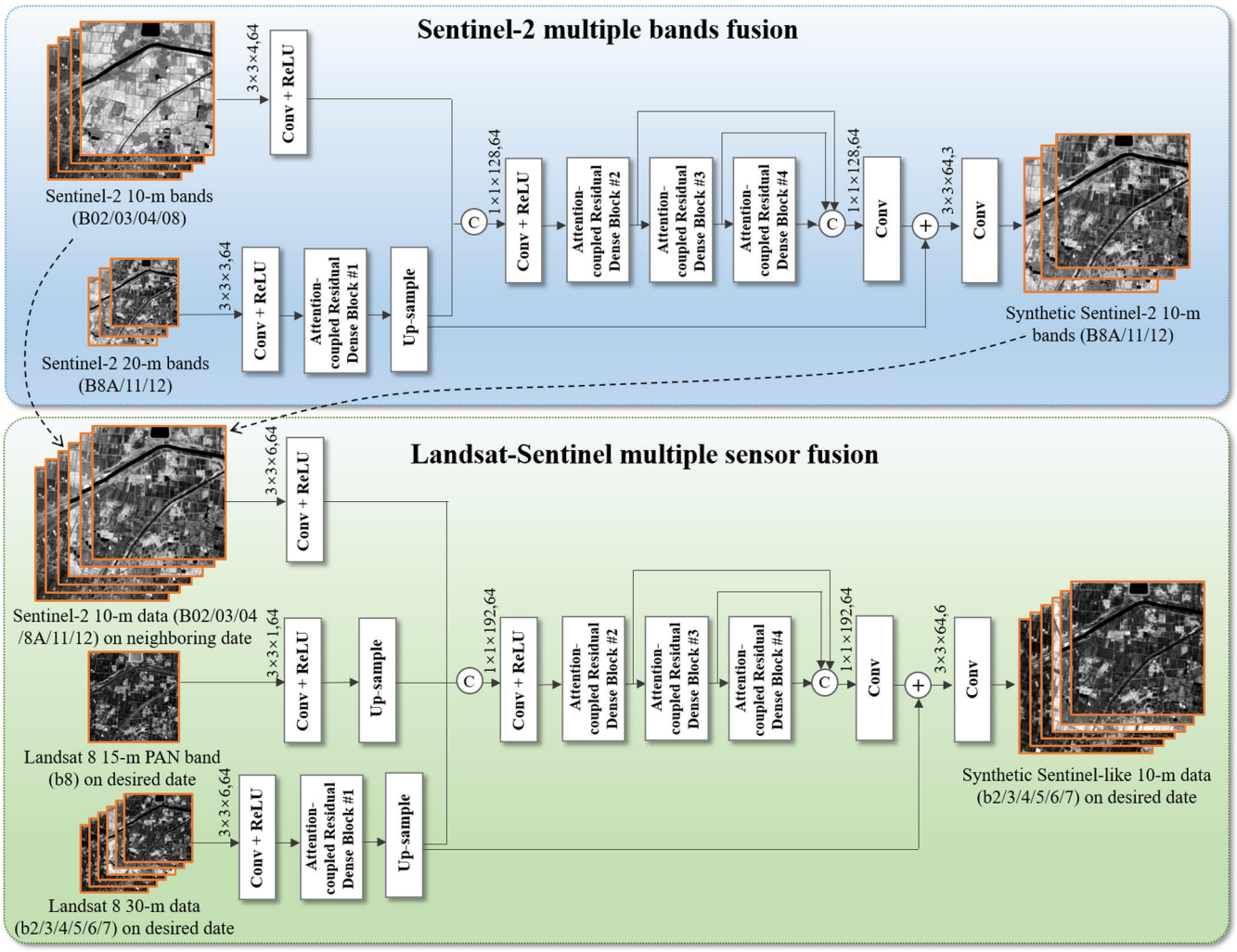


Fig. 1. Flowchart of the proposed DSTFN model.

2.2. Network architecture

2.2.1. Overview of fusion network

The proposed method has two subnetworks with similar structures, as shown in Fig. 1. For simplicity of description, the input data of the two networks are uniformly divided into two parts, namely, the coarse-resolution target data $\boldsymbol{Y}$ and the fine-resolution auxiliary data $\boldsymbol{Z}$. The output data $\boldsymbol{X}$ represent a resolution-enhanced version of $\boldsymbol{Y}$. As mentioned above, in the first stage focusing on the fusion of the Sentinel-2 bands, $\boldsymbol{Y}$ is the observed 20-m band group, $\boldsymbol{Z}$ is the observed 10-m band group, and $\boldsymbol{X}$ is the downscaled result of $\boldsymbol{Y}$; they are depicted as

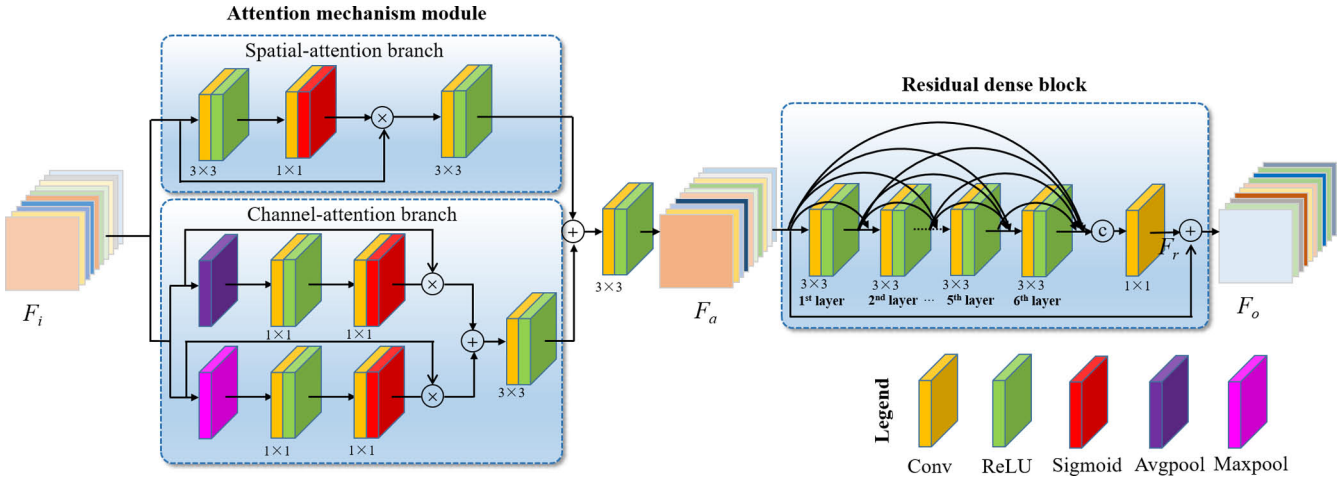


Fig. 2. Illustration of the attention-coupled residual dense block.

$$\begin{cases} \boldsymbol{Y} = \{B8A_{20},\ B11_{20},\ B12_{20}\} \\ \boldsymbol{Z} = \{B02_{10},\ B03_{10},\ B04_{10},\ B08_{10}\} \\ \boldsymbol{X} = \{B8A_{10},\ B11_{10},\ B12_{10}\} \end{cases} \tag{1}$$

where the band notations are consistent with those in Table 1, and the subscripts denote band resolutions. In the second stage, $\boldsymbol{Y}$ represents the 30-m Landsat 8 imagery, $\boldsymbol{Z}$ has two components depicting the 10-m neighboring Sentinel-2 imagery and the 15-m Landsat panchromatic band, and $\boldsymbol{X}$ is the downscaled version of $\boldsymbol{Y}$; they are described as

$$\begin{cases} \boldsymbol{Y} = \{b2_{30},\ b3_{30},\ b4_{30},\ b5_{30},\ b6_{30},\ b7_{30}\} \\ \boldsymbol{Z}_1 = \{B02_{10},\ B03_{10},\ B04_{10},\ B8A_{10},\ B11_{10},\ B12_{10}\} \\ \boldsymbol{Z}_2 = \{b8_{15}\} \\ \boldsymbol{X} = \{b2_{10},\ b3_{10},\ b4_{10},\ b5_{10},\ b6_{10},\ b7_{10}\} \end{cases} \tag{2}$$

The two networks have similar structures. First, the shallow feature maps are separately extracted from $\boldsymbol{Y}$ and $\boldsymbol{Z}$. The fine-resolution source $\boldsymbol{Z}$ adopts a convolutional layer with 64 filters of $3 \times 3 \times b_f$, where b_f denotes the fine-resolution band number. The coarse-resolution source $\boldsymbol{Y}$ applies a convolutional layer with 64 filters of $3 \times 3 \times b_c$ and an attention-coupled residual dense block (ARDB, the details are presented in Section 2.2.2), where b_c denotes the coarse-resolution band number. Subsequently, a bicubic resampling procedure is performed to coordinate the feature map size. Second, the features mapped from the two sources are dimensionally concatenated and processed via three ARDBs. The output feature maps from the three blocks are concatenated by skip connections so that the feature maps at different levels can be fully considered and utilized in the network. Then, the derived feature maps are combined with the upsampled features from the coarse-resolution source via elementwise addition. Finally, the feature maps are processed by a convolutional layer with f_c filters of $3 \times 3 \times 64$ to produce the fine-resolution result $\boldsymbol{X}$.

2.2.2. Attention-coupled residual dense block (ARDB)

Given that the mapping between input and output data is complex and nonlinear, it is a better choice to use a deep network and complex structures to fit this mapping. Therefore, we combine a residual dense block and an attention mechanism module to form the basic unit of the network called attention-coupled residual dense block (ARDB, Fig. 2). The unit applies an attention mechanism module to recalibrate the feature maps and then uses a residual dense block to extract features via dense connections and residual learning. As illustrated in Fig. 1, each network has four ARDBs: one for extracting features from the coarse-resolution source and the other three for extracting features from the combined sources.

As an interpretation of human intuition to pay more attention to the area of interest than the background (Woo et al., 2018), the attention mechanism has been applied in many tasks, such as image super-resolution (Lin et al., 2022a; Lin et al., 2022b) and classification (Tong et al., 2020). The attention mechanism module adaptively adjusts the feature response in the spatial and channel dimension, and thus, it recalibrates the extracted features and enhances the feature expressions. The module in our network has two branches accounting for spatial attention and channel attention. The spatial attention branch employs the first 3×3 convolutional layer to extract local features and the second 1×1 layer to estimate the spatial weights. The channel attention branch consists of two sub-branches, in which the global statistical features are extracted by pooling operators, and then the channel weights are derived by two convolutional layers with 1×1 kernels, followed by elementwise multiplication for signal recalibration. The recalibrated maps are combined from the two sub-branches and processed via a 3×3 convolutional layer. Finally, the output feature maps from the two branches are added elementwise to generate the adjusted feature maps.

After the recalibration, a residual dense block is used to make full use of hierarchical features and focuses on deriving the residual features. The block incorporates the ideas of dense connection (Huang et al., 2017) and residual learning (He et al., 2016; Shen et al., 2020). As illustrated in Fig. 2, each block adopts six "Conv + ReLU" layers, and each "Conv + ReLU" layer applies a 3×3 convolutional layer (Conv), followed by a rectified linear unit (ReLU). The feature maps pass within the block via dense connections; that is, each layer inputs the features from all preceding layers and passes the current features to all succeeding layers (Huang et al., 2017), and this strengthens feature propagation and encourages feature reuse. Mathematically, the output of the l-th layer ($1 \leqslant l \leqslant 6$) in a block is depicted as

$$F_l = f_l(concat(F_a,\ F_1,\ \ldots,\ F_{l-1})) \tag{3}$$

where F_a denotes the recalibrated feature map from the attention module, andF_1, …,F_{l-1}, and F_l denote the feature maps from the *1*, …, *l-1*, and *l*-th layer, respectively. f_l is the "Conv + ReLU" in the l-th layer. $concat(\bullet)$ represents the concatenation operation. To improve the feature representation ability, this study also employs the residual learning strategy. The final output F_o of a block can be expressed as:

$$F_o = F_a + F_r \tag{4}$$

where F_r is the residual feature map derived by 1×1 convolutional layer imposed on the feature maps concatenated from the six previous layers.

2.3. Loss function term based on the image degradation process

The loss function constrains the model and guides the optimization of the network. In the two networks, the loss function $L_t(\Theta)$ can be uniformly depicted as

$$L_t(\Theta) = \alpha L_1(\Theta) + \beta L_f(\Theta) + \gamma L_d(\Theta) \tag{5}$$

where $L_1(\Theta)$ and $L_f(\Theta)$ respectively denote the l_1-norm term and Frobenius norm term that are widely employed to constrain the errors between predictions and labels. $L_d(\Theta)$ is the degradation constraint term specifically designed to ease the temporal variation issue. Θ represents the network parameters.α,β, and γ are the regularization parameters adaptively determined by

$$\begin{aligned} \alpha &= \frac{L_1(\Theta)}{L_1(\Theta) + L_f(\Theta) + L_d(\Theta)}, \\ \beta &= \frac{L_f(\Theta)}{L_1(\Theta) + L_f(\Theta) + L_d(\Theta)}, \\ \gamma &= \frac{L_d(\Theta)}{L_1(\Theta) + L_f(\Theta) + L_d(\Theta)} \end{aligned} \tag{6}$$

The degradation constraint term $L_d(\Theta)$ is employed to solve the temporal variation problem. Specifically, if a dramatic change occurs between two multitemporal input images, the information from the coarse-resolution target image **Y** should be preferentially considered and maximally used to reconstruct the output **X**, because both **X** and **Y** represent the same-day land surface condition. Therefore, we use **Y** to build a constraint for the model. According to the observation degradation model (Shen et al., 2016), **Y** can be considered a degraded version of **X** by introducing warping, blurring, downsampling, and noise operators. Given that the Landsat and Sentinel-2 images are highly consistent after pre-processing steps, including atmospheric correction, bandpass adjustment, and geometric registration (see Section 3.1), the motion, blurring, and noise operators show indistinctive influence. Accordingly, **Y** is approximately equal to a downsampled version of **X**:

$$Y = DX \tag{7}$$

where **D** denotes the downsampling operator. By embracing this idea, we design the degradation terms in the two networks as follows.

In the first stage, given a training dataset $\left\{ \left(z^i_{S_{10m}}, y^i_{S_{20m}} \right); x^i_{S_{10m}} \right\}_{i=1}^{N}$,

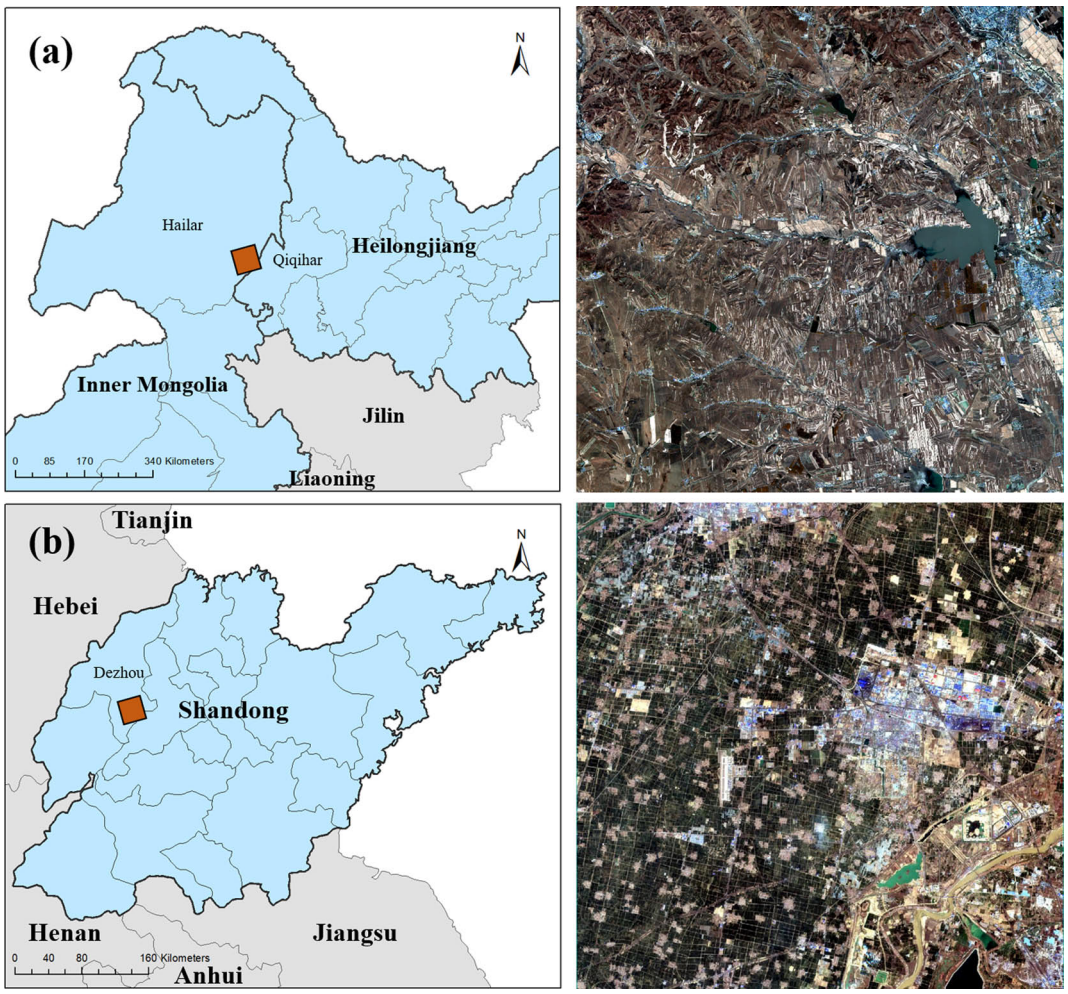


Fig. 3. Geographic locations and collected Sentinel-2 images at Hailar (a) and Dezhou (b).

where $z^i_{S_{10m}}$ and $y^i_{S_{20m}}$ are the 10-m and 20-m band groups, $x^i_{S_{10m}}$ is the label data (i.e., the enhanced version of $y^i_{S_{20m}}$), and N is the training data number. The degradation term $L_d^{net1}(\Theta)$ is expressed as

$$L_d^{net1}(\Theta) = \frac{1}{2N}\sum_{i=1}^{N}\left\| y^i_{S_{20m}} - f_d\left(\xi_{net1}\left(z^i_{S_{10m}}, y^i_{S_{20m}}\right) + f_u\left(y^i_{S_{20m}}\right)\right)\right\|_F^2 \quad (8)$$

where $\xi_{net1}(\bullet)$ denotes the residual output from the first-stage network. The residual output is combined with the upsampled coarse-resolution data to derive the fusion result. $f_d(\bullet)$ and $f_u(\bullet)$ denote the downsampling and upsampling operators, respectively.

In the second stage, the training dataset is depicted as $\left\{\left(z^i_{S_{10m}}, z^i_{L_{15m}}, y^i_{L_{30m}}\right); x^i_{L_{10m}}\right\}_{i=1}^{M}$, where $z^i_{S_{10m}}$, $z^i_{L_{15m}}$, and $y^i_{L_{30m}}$ are the 10-m Sentinel-2 imagery, 15-m Landsat panchromatic band, and 30-m Landsat imagery, respectively; $x^i_{L_{10m}}$ is the label data (i.e., the resolution-enhanced result of $y^i_{L_{30m}}$); and M is the training data number. In this case, the degradation term $L_d^{net2}(\Theta)$ is depicted as

$$L_d^{net2}(\Theta) = \frac{1}{2M}\sum_{i=1}^{M}\left\| y^i_{L_{30m}} - f_d\left(\xi_{net2}\left(z^i_{S_{10m}}, z^i_{L_{15m}}, y^i_{L_{30m}}\right) + f_u\left(y^i_{L_{30m}}\right)\right)\right\|_F^2 \quad (9)$$

where $\xi_{net2}(\bullet)$ denotes the residual output from the second-stage network.

3. Test datasets, network training, and baseline methods

3.1. Study sites and test datasets

Two time-series datasets are constructed for model evaluation. The first site ("Hailar" herein, Fig. 3(a)) is located at the border across Hailar and Qiqihar in Northeast China, with an area of about 1,568 km^2 (3,960 × 3,960 10-m pixels). It has various land covers, including farmlands, woodlands, lakes, and residential settlements. The woodlands are generally homogeneous and show slow phenological variations through time, while the farmlands are more heterogeneous and change more significantly due to human-induced activities. The farmlands at this site are dominantly covered by two crop species, namely, soybean (sowed in late April and matures in early-to-mid September) and maize (sowed in late May and matures in mid-to-late September). The second site ("Dezhou" herein, Fig. 3(b)) is located in Dezhou, Shandong Province, China, and it covers a spatial extent of 882 km^2 (2,970 × 2,970 10-m pixels). Although primarily covered by farmlands, this site has more built-up regions (e.g., urban regions and scattered villages) than the previous site. The farmlands at this site vary through time with growing cycles of two crops, namely, winter wheat (sowed in the previous year and matures in early June) and maize (sowed in late June and matures in early October).

Landsat 8 Level-2 and Sentinel-2 Level-1C products were collected from the United States Geological Survey web portal (https://earthexplorer.usgs.gov/). The Hailar dataset comprises 23 cloud-free scenes from 2019, with 10 obtained from Landsat 8 (path/

Table 2
Imagery used for model training and validation in the two sites.

	Training		Validation	
	Landsat 8 (MM/DD/YY)	Sentinel-2 (MM/DD/YY)	Landsat 8 (MM/DD/YY)	Sentinel-2 (MM/DD/YY)
Hailar site	01/23/2019	01/02/2019	01/07/2019	01/07/2019
	02/24/2019	01/22/2019	02/08/2019	01/27/2019
	05/15/2019	02/16/2019	04/13/2019	02/26/2019
	09/04/2019	03/03/2019	06/16/2019	03/13/2019
	10/22/2019	03/23/2019	10/06/2019	04/02/2019
		05/02/2019		09/29/2019
		10/19/2019		
Dezhou site	01/11/2018	01/05/2018	02/12/2018	02/04/2018
	04/17/2018	03/26/2018	03/16/2018	03/16/2018
	05/03/2018	04/20/2018	06/20/2018	03/31/2018
	09/08/2018	09/07/2018	09/24/2018	07/19/2018
	10/26/2018	10/02/2018	10/10/2018	09/22/2018
		11/01/2018	12/13/2018	10/17/2018
				12/16/2018

row: 121/027) and 13 (tile: T51UWP) from Sentinel-2. The Dezhou dataset contains 24 scenes from 2018, with 11 obtained from Landsat 8 (122/035) and 13 from Sentinel-2 (T50SMF). The Landsat Level-2 product was atmospherically corrected by Landsat 8 Surface Reflectance Code (LaSRC) before data distribution, and the Sentinel-2 L1C product was corrected by Sen2Cor. We extracted the visible, near infrared, and shortwave infrared bands from the two sources. The panchromatic band was also collected from Landsat. The collected data were geometrically aligned and spatially clipped to ensure the same extent. The slight differences caused by bandpass configuration were adjusted via the linear correction model developed by Zhang et al. (2018) with the linear coefficients regressed from coincident Landsat and Sentinel-2 surface reflectance data.

3.2. Model training

Each dataset was divided into two parts for network training and testing (Table 2). In the first stage, given that the label data (i.e., 10-m B8A and B11-B12) are unavailable, we followed Wald's protocol and trained the model based on resolution-degraded data with a scale factor of 2 (Shao et al., 2019). Specifically, the Sentinel-2 10-m and 20-m bands were degraded to 20-m and 40-m, respectively. The degraded bands (i.e., 20-m B02–B04 and B08; 40-m B8A and B11–B12) were used as input, and the 20-m B8A, B11, B12 bands as output labels. In the second stage, we adopted the same strategy but with a degradation factor of 3. Each Landsat image was combined with a temporally adjacent Sentinel-2 imagery before/after the Landsat. The Sentinel-2 imagery was degraded to 30-m, and the Landsat multispectral imagery and the panchromatic band were degraded to 45-m and 90-m, respectively. The degraded data were used as input, and the observed Landsat 30-m multispectral imagery was used as output labels. By constructing the mapping at degraded resolutions, the two deep networks can be trained.

3.3. Baseline methods

Four baseline methods, namely, bilinear interpolator, STARFM under simplified input modality (STARFM-SI) (Wu et al., 2020), ATPRK (Wang et al., 2017), and ESRCNN (Shao et al., 2019), were used for comparison. The bilinear interpolator directly interpolates the coarse-resolution imagery to a fine resolution while the other three approaches restore spatial structures by adding fine-resolution auxiliary imagery. STARFM-SI adopts a weighted linear relationship to achieve the fusion. ATPRK, a geostatistical algorithm, involves semi-variogram modeling from the cokriging matrix. ESRCNN applies deep learning to formulate the mapping between different resolutions. Fusion results were visually and quantitatively assessed against ground truth. Eight quantitative measures, namely, mean absolute error (MAE), mean

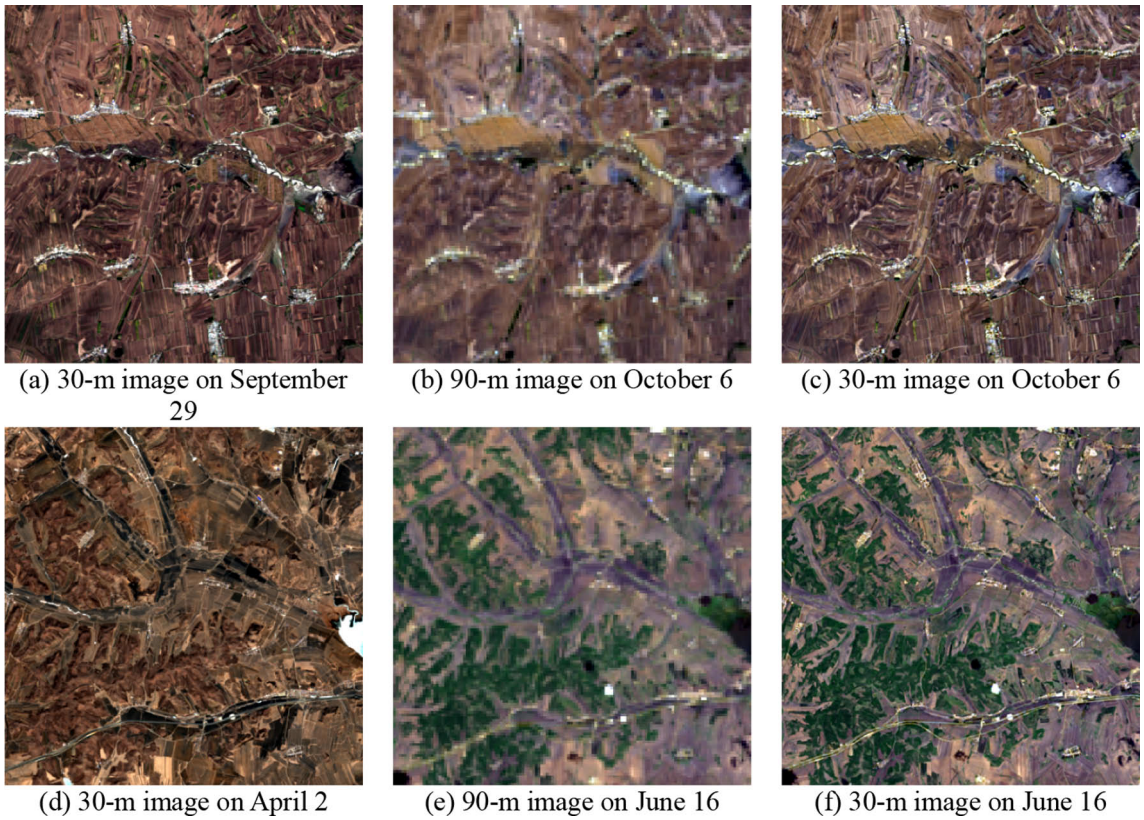


Fig. 4. Resolution-degraded input imagery and ground truth imagery in the two tests with slight (a–c) and dramatic (d–f) changes at the Hailar site. (a) and (d) are degraded from Sentinel-2, (b) and (e) are degraded from Landsat 8, and (c) and (f) are the observed Landsat 8 images.

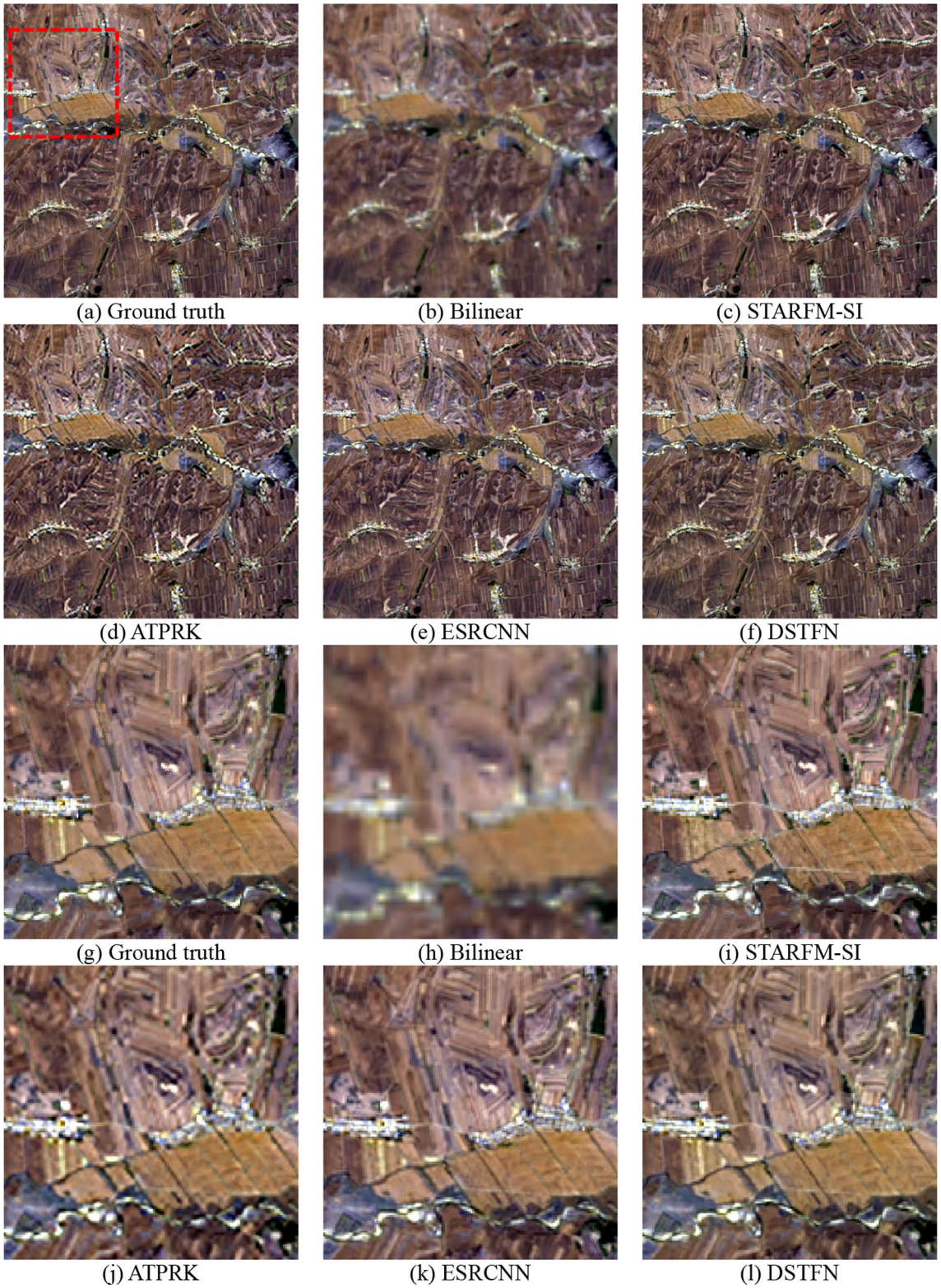


Fig. 5. Experimental results in the test featuring slight change at the Hailar site. (g)–(l) are detailed views of (a)–(f) in the subset region marked with a red square.

relative error (MRE), root-mean-square error (RMSE), erreur relative global adimensionnelle de synthèse (ERGAS), spectral angle mapper (SAM), correlation coefficient (CC), peak signal-to-noise ratio (PSNR), and structural similarity (SSIM), were used to reveal the fusion performance. For details on the quantitative measures, please refer to Text S1 in the supplementary material. The lower MAE, MRE, RMSE, SAM, and ERGAS scores and higher CC, PSNR, and SSIM scores indicated superior fusion outputs.

4. Experimental results

4.1. Experiments based on resolution-degraded data at Hailar site

The collected Hailar dataset was used to validate the model. Each

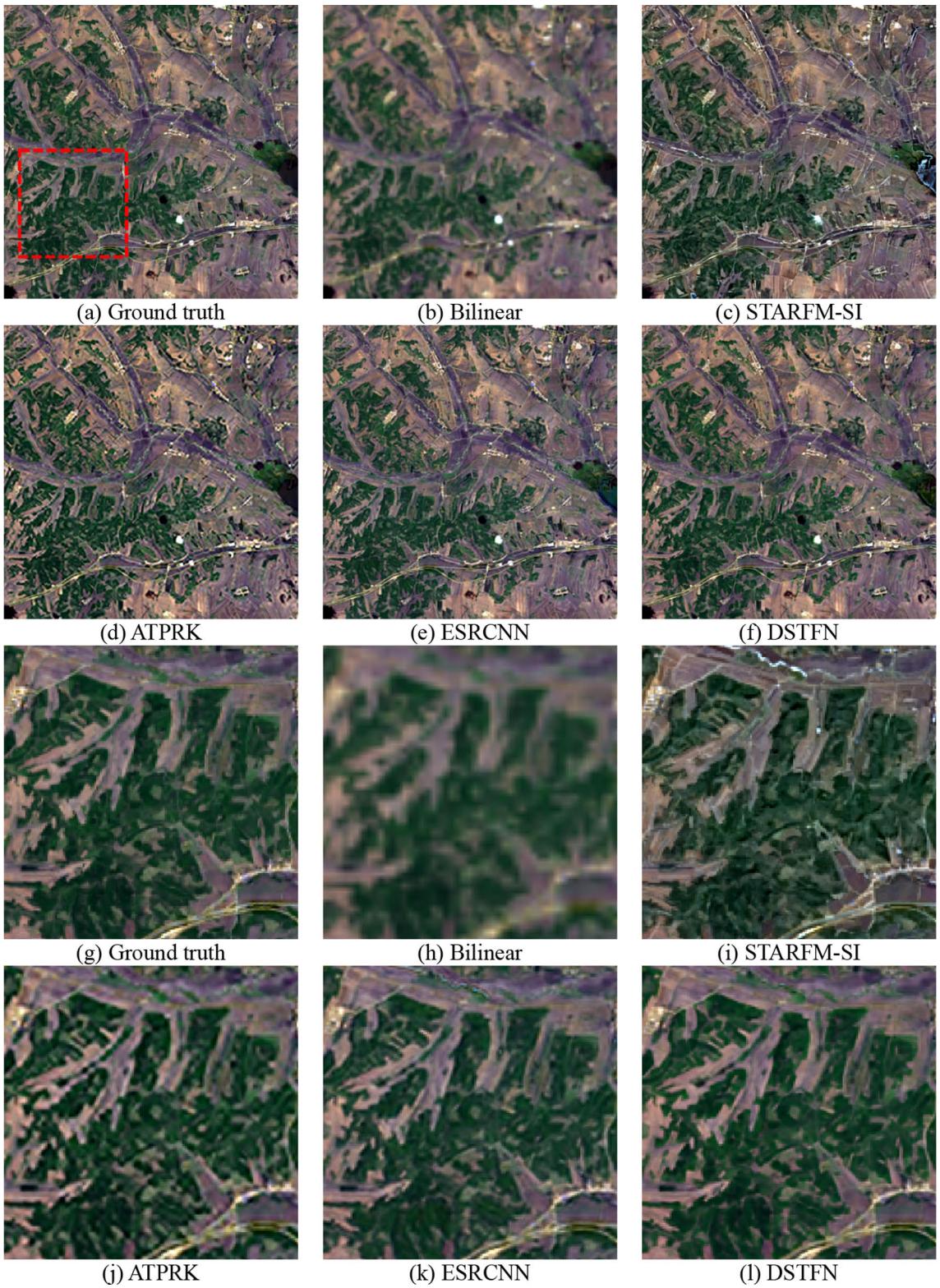


Fig. 6. Experimental results in the test featuring dramatic change at the Hailar site. (g)–(l) are detailed views of (a)–(f) in the subset region marked with a red square.

Landsat imagery was combined with the temporally nearest Sentinel-2 imagery before/after Landsat, resulting in eight data groups for testing. Following the training strategy, the eight tests were performed on the basis of resolution-degraded imagery to obtain the baseline data as ground truth. Specifically, the 90/45-m Landsat-degraded data and the 30-m Sentinel-degraded data were merged to derive the 30-m imagery (i.e., the resolution-enhanced output of the Landsat-degraded data), and the result was assessed against the observed 30-m Landsat imagery.

The two tests showing slight and dramatic changes are presented for visual comparison. The resolution-degraded inputs and ground truth in the tests are illustrated in Fig. 4. The two cases have time gaps of 7 days (Fig. 4(a)–(c)) and 75 days (Fig. 4(d) –(f)), respectively, in which the first case shows inapparent change while the second has dramatic

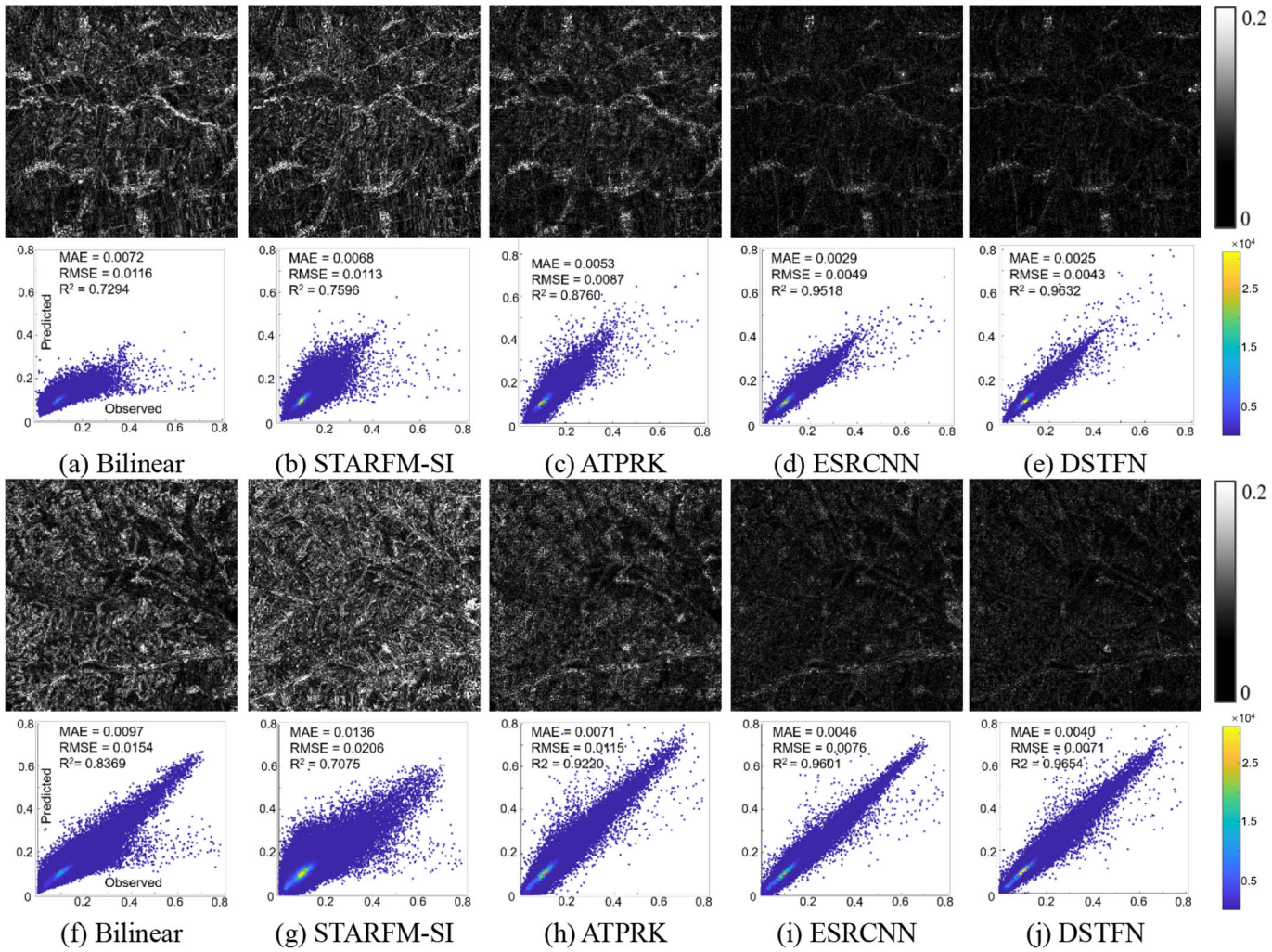


Fig. 7. Red-band error maps and scatter plots of fusion results against the ground truth in the two temporally diverse tests at the Hailar site. (a)–(e) and (f)–(j) belong to the slight and dramatic change cases, respectively.

Table 3
Quantitative results of the two temporally diverse tests at the Hailar site.

		Ideal	Bilinear	STARFM-SI	ATPRK	ESRCNN	DSTFN
Slight change test	MAE	0	0.0087	0.0077	0.0067	0.0038	**0.0034**
	MRE	0	0.0671	0.0641	0.0559	0.0334	**0.0282**
	RMSE	0	0.0142	0.0131	0.0111	0.0064	**0.0058**
	SAM	0	2.0089	1.8457	1.7612	1.1179	**1.0329**
	ERGAS	0	0.7505	0.7476	0.5831	0.3459	**0.3148**
	CC	1	0.8755	0.8897	0.9354	0.9753	**0.9800**
	SSIM	1	0.9905	0.9916	0.997	0.9994	**0.9995**
	PSNR	$+\infty$	36.2956	37.1957	38.3869	43.0591	**43.7915**
dramatic change test	MAE	0	0.0120	0.0153	0.0101	0.0070	**0.0059**
	MRE	0	0.0892	0.1311	0.0815	0.0631	**0.0492**
	RMSE	0	0.0189	0.0238	0.0157	0.0112	**0.0100**
	SAM	0	3.3871	4.2133	3.3312	2.3773	**1.9917**
	ERGAS	0	0.9029	1.2498	0.7194	0.5097	**0.4643**
	CC	1	0.9242	0.858	0.9559	0.9753	**0.9798**
	SSIM	1	0.9875	0.9576	0.9949	0.9982	**0.9988**
	PSNR	$+\infty$	33.9028	32.1334	35.2991	38.1388	**39.2244**

change, with forest phenology stages transforming from dormancy to greening. The results of the two tests and detailed views are shown in Figs. 5 and 6. In the slight change case (Fig. 5), the bilinear interpolator completely fails to recover the spatial structures. ATPRK presents a slightly blurry output while STARFM-SI, ESRCNN, and DSTFN produce results visually similar to the ground truth. In the dramatic change case (Fig. 6), according to the zoomed-in views, STARFM-SI predicts the result with spectral distortion in the green forest area, ATPRK still has room for recovering the sharp edges of ground features, and the deep-learning-based DSTFN and ESRCNN outperform the others from the spatial and spectral aspects.

The quantitative descriptions reflect the visually imperceptible

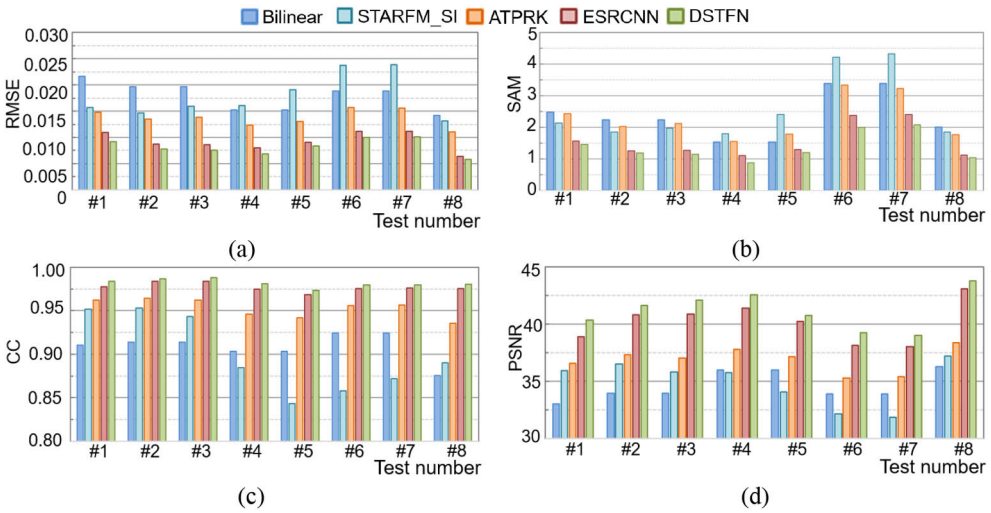


Fig. 8. Quantitative results of the eight experiments at the Hailar site. The observed dates of input images in these tests are shown in Table 2.

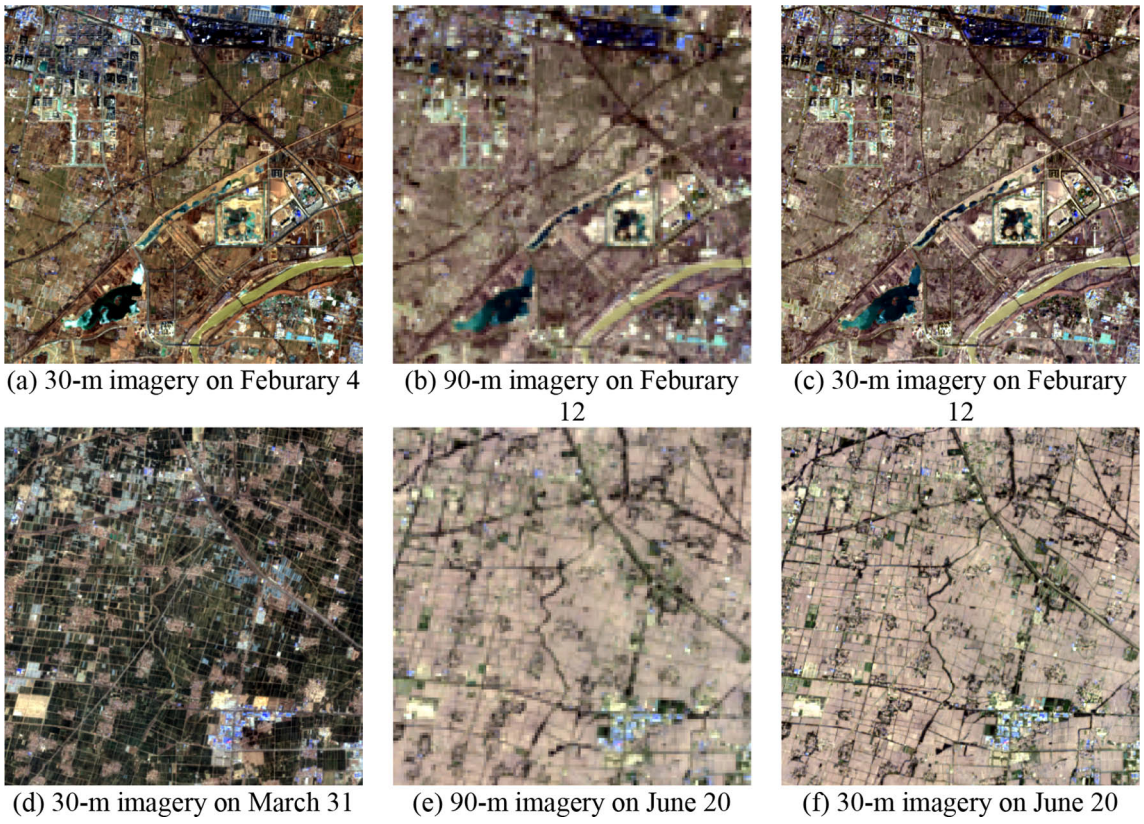


Fig. 9. Resolution-degraded input imagery and ground truth imagery in two tests with slight (a–c) and dramatic (d–f) changes at the Dezhou site. (a) and (d) are degraded from Sentinel-2, (b) and (e) are degraded from Landsat 8, and (c) and (f) are the observed Landsat 8 images.

differences. According to Fig. 7 which shows the red-band error maps and scatter plots of the fusion results against the ground truth, DSTFN and ESRCNN have considerably fewer errors than the others because the error maps have less bright area and the scatters are more fixed around the ideal 1:1 line. The quantitative descriptions in the scatter plots reveal that DSTFN outperforms ESRCNN to a certain degree. For example, DSTFN gives red-band predictions with large R^2 scores (DSTFN vs. ESRCNN:0.9632 vs. 0.9518) in the first case. Table 3 presents the quantitative results averaged from the six bands. DSTFN achieves the quantitative scores closest to the ideal values across the eight measures; this result suggests its superiority. Eight tests were performed at the Hailar site, and their quantitative results in terms of RMSE, SAM, CC, and PSNR are mapped in Fig. 8. DSTFN stably achieves the highest CC and PSNR scores and the lowest RMSE and SAM scores and is thus robust under various scenes. Specifically, relative to the benchmark methods, DSTFN shows a 0.79%–4.3% decrease in MRE and a 1.00–6.57 increase in PSNR on average in these tests.

4.2. *Experiments based on resolution-degraded data at Dezhou site*

Based on the Dezhou dataset, we applied the testing strategy in Section 4.1 to perform twelve tests. The Dezhou site has more complex

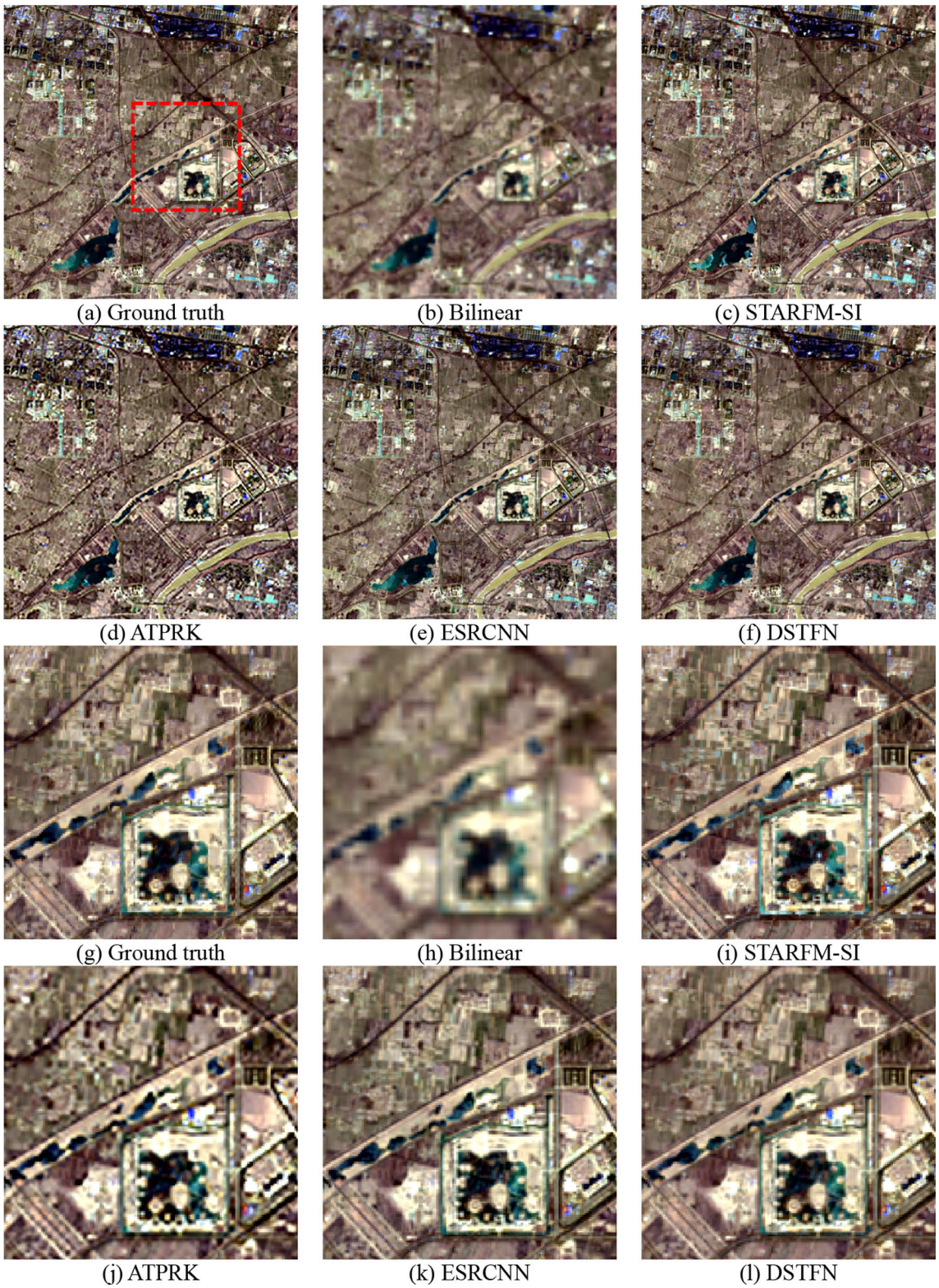


Fig. 10. Experimental results in the test featuring slight change at the Dezhou site. (g)–(l) are the detailed views of (a)–(f) in the subset region marked with a red square.

landscapes than the previous site, such as spatially heterogeneous artificial buildings and temporally dynamic farming systems. Similarly, the two temporally diverse cases are displayed, and the input data, ground truth, and results in two tests are shown in Figs. 9–11. In the slight change case (Fig. 9(a)–(c) and Fig. 10), the ground surface shows insignificant change during the 8-day time gap. The bilinear interpolator fails to reconstruct spatial details. ATPRK produces the result with slight blurry effects. By contrary, DSTFN, ESRCNN, and STARFM-SI derive plausible outputs. In the dramatic change case (Fig. 9(d)–(f) and Fig. 11), STARFM-SI fails to capture subtle features such as line-shaped roads. ATPRK continues to show limited capacities to recover edges. ESRCNN suffers from slight spectral inconsistency. DSTFN is generally

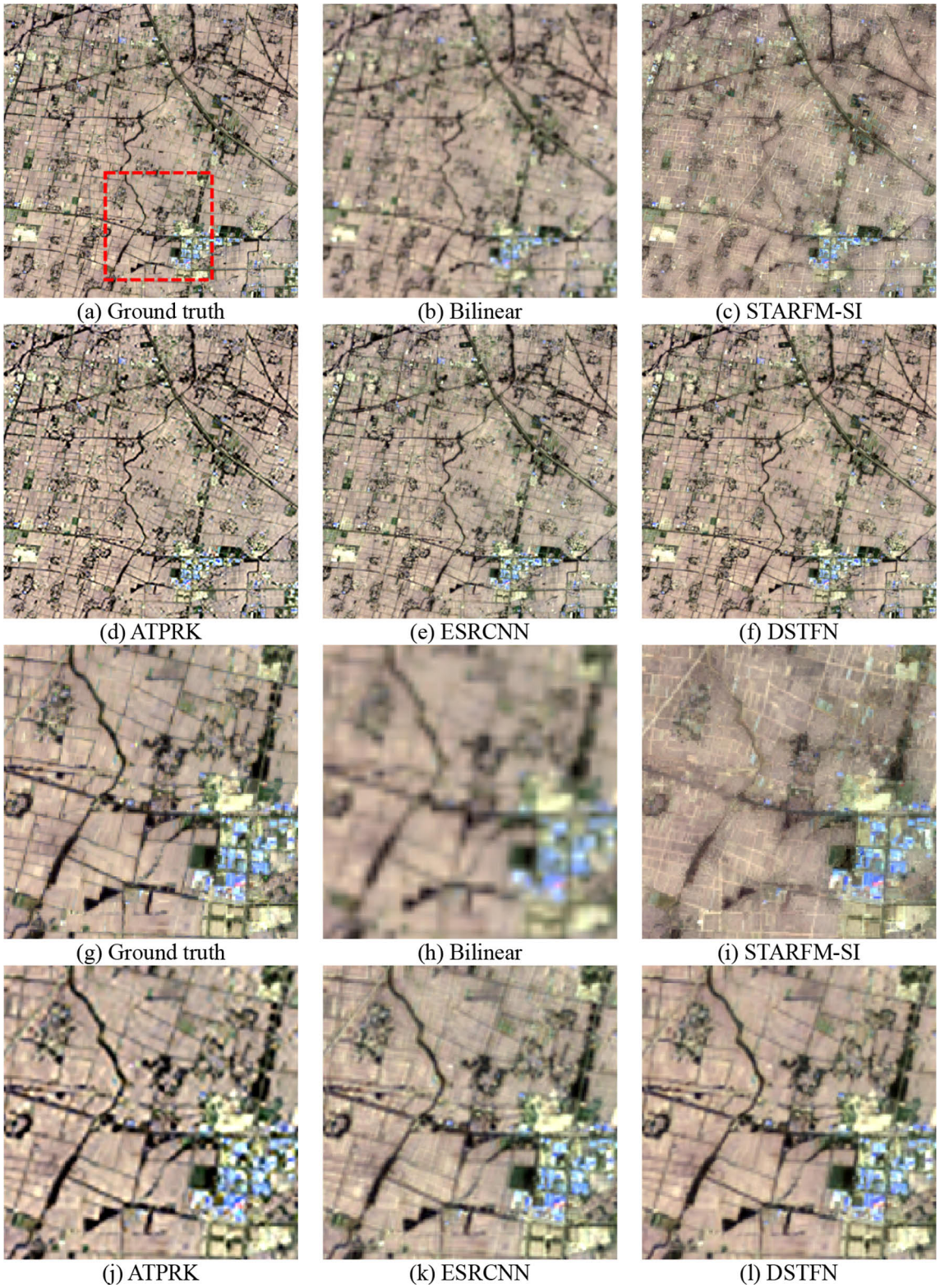


Fig. 11. Experimental results in the test featuring dramatic change at the Dezhou site. (g)–(l) are the detailed views of (a)–(f) in the subset region marked with a red square.

closet to the ground truth from the spatial and spectral perspectives.

Fig. 12 shows the red-band error maps and scatter plots of the above two tests, and it can be observed that DSTFN and ESRCNN have remarkable advantages over the others. Between the two deep learning approaches, DSTFN is more effective because it has fewer errors (e.g., DSTFN vs. ESRCNN in terms of RSME: 0.0084 vs. 0.0103). The quantitative results involving full bands are presented in Table 4. The finding generally conforms to the visual comparison that DSTFN has the best performance since its quantitative results are closest to the ideal quantitative values. Twelve tests were performed at the Dezhou site, and the quantitative descriptions of RMSE, SAM, CC, and PSNR in these tests are shown in Fig. 13. DSTFN shows robustness across various landscapes

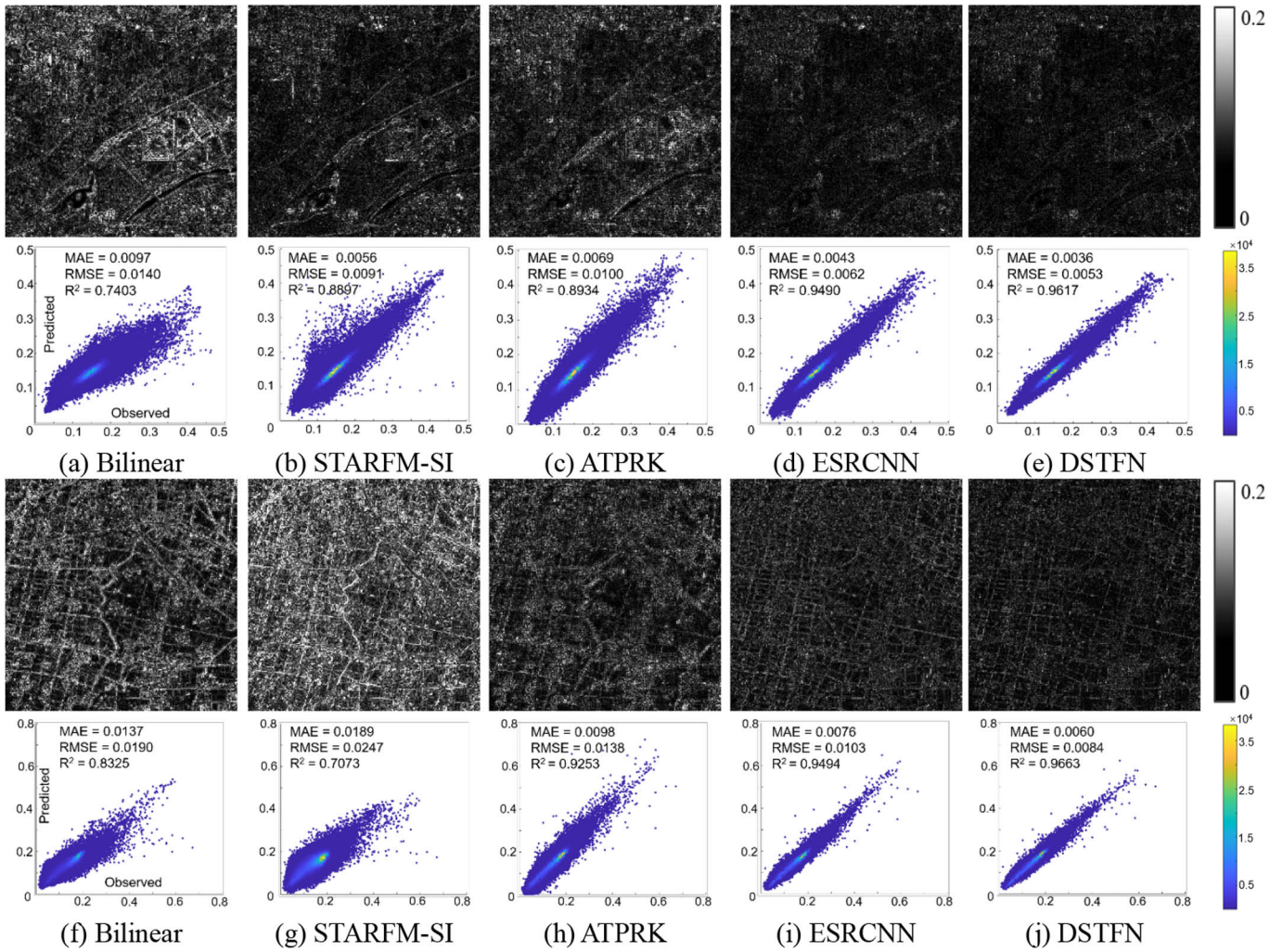


Fig. 12. Red-band error maps and scatter plots of fusion results against the ground truth in the two temporally diverse tests at the Dezhou site. (a)–(e) and (f)–(j) belong to the slight and dramatic change cases, respectively.

Table 4
Quantitative results of the two temporally diverse tests at the Dezhou site.

		Ideal	Bilinear	STARFM-SI	ATPRK	ESRCNN	DSTFN
slight-change test	MAE	0	0.0108	0.0063	0.0068	0.0048	**0.0044**
	MRE	0	0.0645	0.0401	0.0490	0.0320	**0.0277**
	RMSE	0	0.0161	0.0101	0.0103	0.0071	**0.0066**
	SAM	0	2.0824	1.4031	1.6708	1.2138	**1.0726**
	ERGAS	0	0.5846	0.3963	0.3933	0.2637	**0.2410**
	CC	1	0.8725	0.9441	0.9532	0.9749	**0.9789**
	SSIM	1	0.9797	0.9919	0.9941	0.9983	**0.9987**
	PSNR	$+\infty$	35.4524	39.6289	39.4981	42.4609	**43.0676**
dramatic-change test	MAE	0	0.0153	0.0207	0.0123	0.0105	**0.0087**
	MRE	0	0.0873	0.1447	0.0797	0.0568	**0.0472**
	RMSE	0	0.0218	0.0275	0.0175	0.0144	**0.0123**
	SAM	0	3.4373	4.8723	2.9265	2.7399	**2.2969**
	ERGAS	0	0.7255	0.9137	0.5682	0.4478	**0.3836**
	CC	1	0.9070	0.8419	0.9481	0.9579	**0.9691**
	SSIM	1	0.9694	0.8992	0.9875	0.9930	**0.9953**
	PSNR	$+\infty$	32.5511	30.5379	34.3971	35.8314	**37.2402**

and temporal dynamics since it obtains the highest CC and PSNR values, and the lowest RMSE and SAM values in every experiment. Compared with the other approaches, DSTFN shows a 0.91%–6.40% decrease in MRE and a 0.95–6.60 increase in PSNR. In addition, DSTFN outperforms ESRCNN, and its quantitative superiority is more obvious in the scenes with dramatic change, demonstrating that DSTFN has stronger capacities to deal with temporal variations.

4.3. Experiments based on original data at Dezhou site

The Dezhou site has a more complex landscape than the Hailar site. Thus, we used this site to perform the tests on the basis of the original

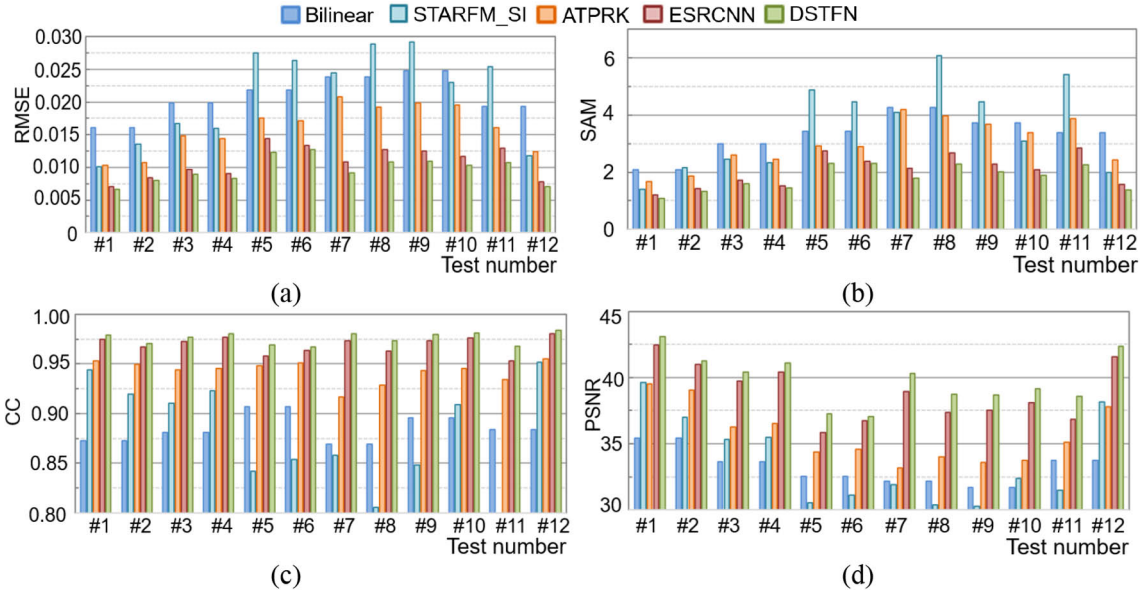


Fig. 13. Quantitative results of the twelve experiments at the Dezhou site. The observed dates of input images in these tests are shown in Table 2.

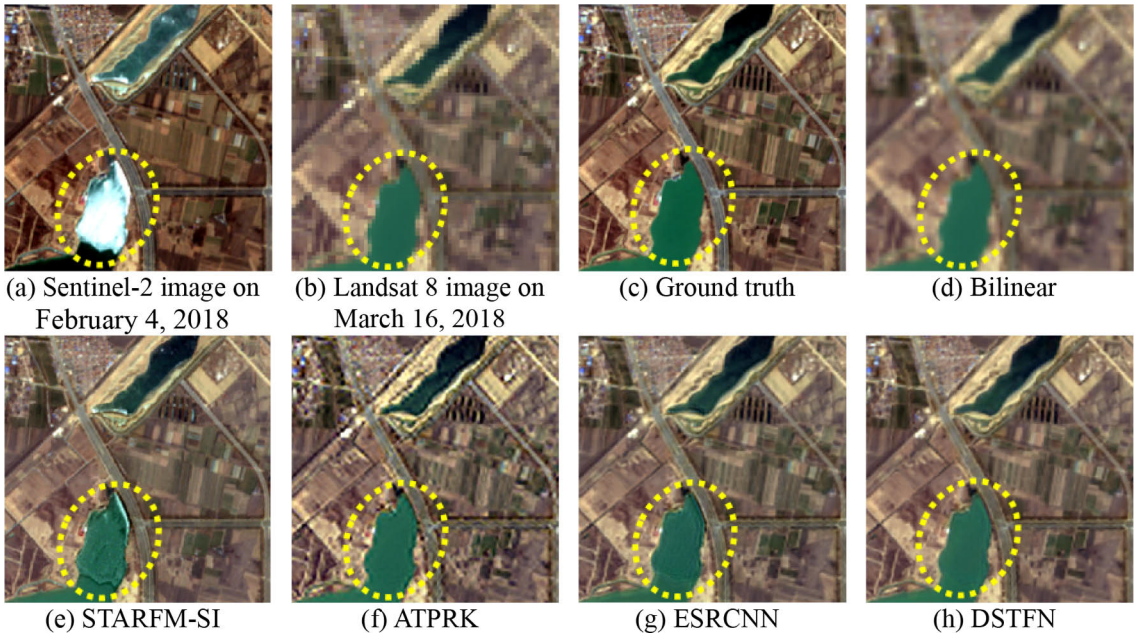


Fig. 14. Experimental results of a subset region in the test featuring a slight change at the Dezhou site.

observations. The Landsat–Sentinel image pair was collected on March 16, 2018. By fusing the Landsat image with a temporally adjacent Sentinel image, we synthesized a Sentinel-like image on March 16, 2018. Then, the fused image was compared with the collected Sentinel image to assess the model performance. Similar to those described in the previous sections, the results of two tests featuring diverse temporal dynamics are displayed in Fig. 14 and Fig. 15, and their quantitative results are given in Table 5. To facilitate the visual comparison, we present two subset regions involving land cover changes.

The first experiment focuses on a scene with a slight temporal change (Fig. 14). In this case, the farmlands experienced mild phenological variations while the region marked by a yellow ellipse showed significant changes, transforming from bright ice to dark green water. The visual comparison indicates that the bilinear interpolator causes serious image blur and that ATPRK shows slight artifacts. Although the results reveal slight radiometric differences from the ground truth, STARFM-SI, ESRCNN, and DSTFN generally reproduce spatial structures and provide reliable outputs. As for the capability of capturing land cover changes, the ice outlines are more evident in the results of STARFM-SI and ESRCNN than in the output of DSTNF. The second experiment (Fig. 15) features a dramatic change. The nearly four-month time gap caused strong temporal variations because the crops were at contrasting phenological stages in multitemporal imagery. The subset region marked by a yellow ellipse shows significant land cover transformations caused by artificial constructions. Generally, STARFM-SI shows weakness in recovering transformed land surface, ATPRK continues to show noticeable artifacts, while the deep learning methods visually outperform the others. The quantitative results in Table 5 reveal that DSTFN is superior to ESRCNN in the two tests. By comparing the two tests, we find that the four fusion methods show degraded performance along with an

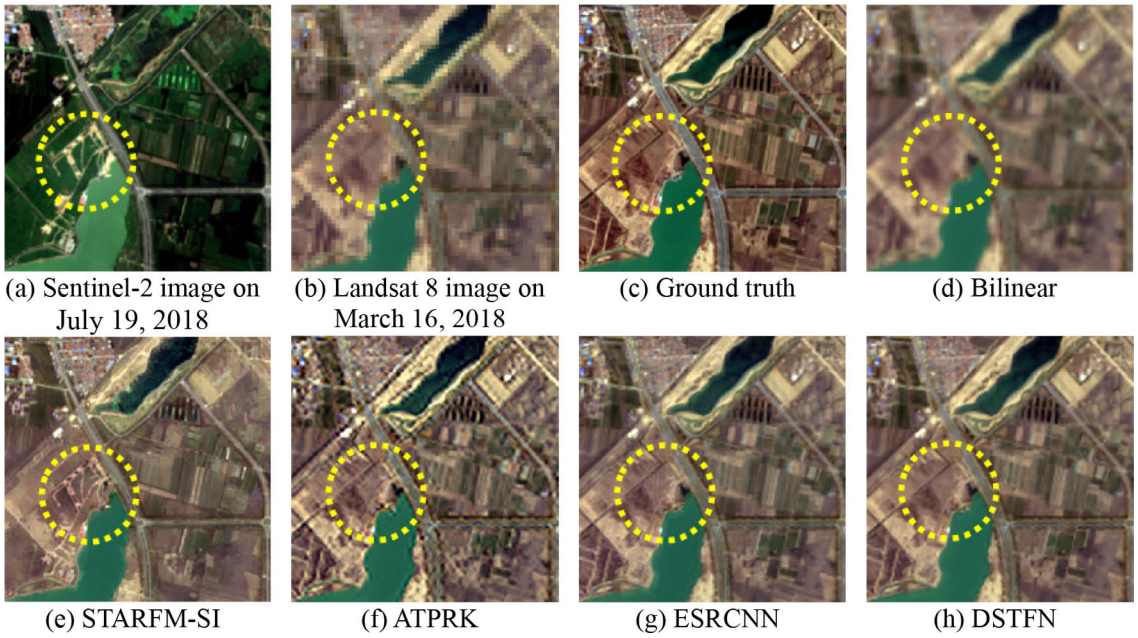


Fig. 15. Experimental results of a subset region in the test featuring a dramatic change at the Dezhou site.

Table 5
Quantitative results of the two tests based on original data at the Dezhou site.

		Ideal	Bilinear	STARFM-SI	ATPRK	ESRCNN	DSTFN
slight-change test	MAE	0	0.0146	0.0120	0.0147	0.0114	**0.0111**
	MRE	0	0.0987	0.0834	0.0896	0.0806	**0.0781**
	RMSE	0	0.0206	0.0168	0.0209	0.0158	**0.0155**
	SAM	0	3.3264	3.0509	3.8422	2.9786	**2.8486**
	ERGAS	0	0.8726	0.7031	0.8450	0.6583	**0.6499**
	CC	1	0.8970	0.9339	0.9030	0.9433	**0.9456**
	SSIM	1	0.9870	**0.9889**	0.9885	0.9886	0.9886
	PSNR	$+\infty$	33.6062	35.3616	33.4213	35.8348	**36.0158**
dramatic-change test	MAE	0	0.0146	0.0184	0.0158	0.0122	**0.0118**
	MRE	0	0.0987	0.1195	0.1388	0.0841	**0.0809**
	RMSE	0	0.0206	0.0259	0.0226	0.0170	**0.0167**
	SAM	0	3.3264	4.5354	4.1161	3.1478	**3.0249**
	ERGAS	0	0.8726	1.0222	0.8835	0.7017	**0.6869**
	CC	1	0.8970	0.8433	0.8931	0.9344	**0.9376**
	SSIM	1	0.9870	0.9838	0.9883	0.9884	**0.9886**
	PSNR	$+\infty$	33.6062	31.5148	32.6897	35.1782	**35.3382**

extended magnitude of ground variations. Nevertheless, DSTFN has the minimum decreasing tendency quantitatively. For example, the MRE scores decrease by 3.61%, 4.92%, 0.35%, and 0.28% for STARFM-SI, ATPRK, ESRCNN, and DSTFN, respectively. These results demonstrate the stronger ability of DSTFN to capture dramatic changes.

Lastly, we present a synthetic 10-m image in Fig. 16, which was downscaled from the Landsat observation by using the proposed deep network. Four representative regions featuring complex landscapes, such as agricultural farmlands and urban artificial constructions were zoomed in for comparison between the 30-m and 10-m scenes. Generally, the original 30-m Landsat imagery cannot provide sufficient spatial structures of heterogeneous ground features. By contrast, the synthetic 10-m imagery significantly enhances the spatial resolution and demonstrates a strong capacity to distinguish ground features.

5. Discussion

5.1. Analysis of the five resolution enhancement methods

The five methods herein can be categorized into three groups, i.e., interpolation-based, linear-regression-based, and deep-learning-based methods. The bilinear interpolation method upsamples the Landsat images to a 10-m resolution without using auxiliary data and shows incapability to reproduce spatial structures. The linear-regression-based group assumes a linear mapping between inputs and outputs. STARFM involves a linear weighting function with respect to spatial, temporal, and spectral information from similar pixels (Gao et al., 2006). STARFM searches similar pixels based on the fine-resolution auxiliary image, the selected pixels are not so reliable to give accurate estimations when the land surface remarkably changes in the multitemporal input images; Thus STARFM works less effectively in cases featuring land cover changes (e.g., Fig. 11). ATPRK produces estimations by combing linear prediction with residual compensation (Wang et al., 2017). The residual compensation helps to preserve the spectral coherence between fine-resolution outputs and coarse-resolution inputs. Nevertheless, the residuals are estimated at the coarse resolution, and even they are downscaled by geostatistical techniques, the outputs still show slight blur and artifacts (e.g., Fig. 6).

The deep-learning-based group applies the deep learning techniques to fit the mapping, and the data-driven models spontaneously learn a complex nonlinear relationship by exploiting training samples. As clearly observed from our tests, the deep learning methods perform

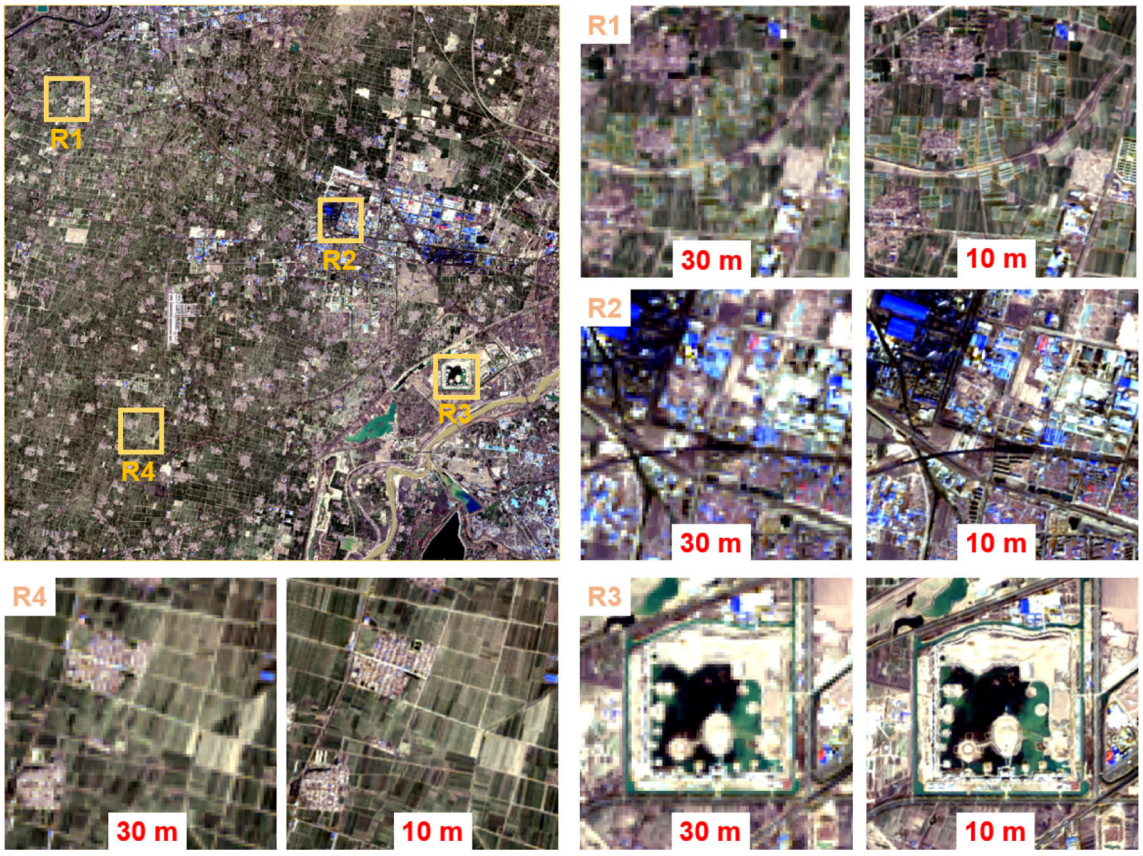


Fig. 16. Synthetic 10-m imagery produced by DSTFN and the comparison between 30-m and 10-m scenes.

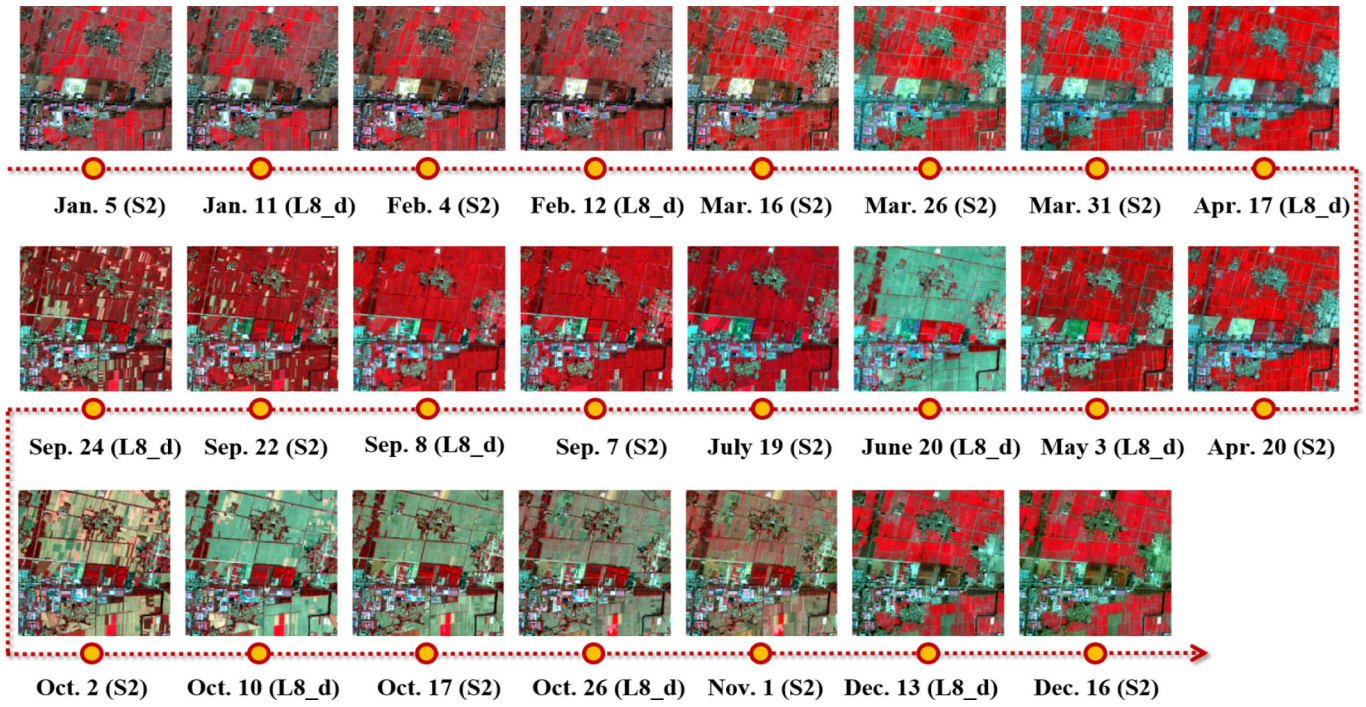


Fig. 17. 10-m dense time series in 2018 in a Dezhou subset region by using DSTFN. "S2" and "L8_d" represent Sentinel-2 and downscaled Landsat 8 imagery, respectively.

more accurately and robustly than the other groups. The two methods herein are developed based on CNN, but with different components. Compared with ESRCNN, DSTFN has two advantages: it has much deeper layers and more complex network structures to fit the mapping and uses a degradation term to constrain the relationship between fine-resolution outputs and coarse-resolution inputs. The constraint term can

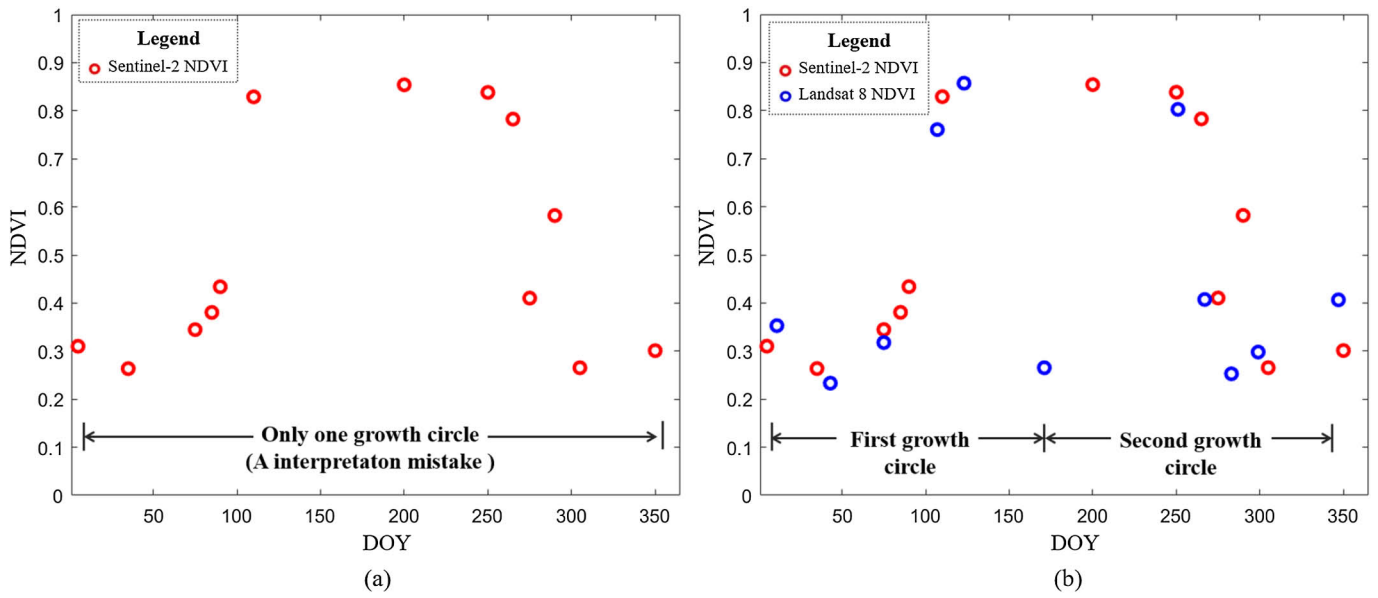


Fig. 18. NDVI dynamics recorded by the Sentinel-2 source (a) and the combined source (b).

maximize the use of information from coarse-resolution inputs and reduce the adverse effect of the fine-resolution auxiliary inputs, which is very useful for dealing with significant surface changes. Due to these improvements, DSTFN performs over ESRCNN. Even in the cases with land cover changes (e.g., Fig. 11), DSTFN can avoid the spectral distortion issue and get robust estimations.

5.2. Implications for time-series analysis

Through the synergistic use of Landsat 8 and Sentinel-2 imagery, time-series analysis can be performed at a frequency denser than that when using a single data source. For instance, at the Dezhou site, although the twin Sentinel-2 satellites have a 5-day revisit cycle, we collected only 13 cloud-free Sentinel-2 images in 2018. Meanwhile, Landsat 8 provided another 10 usable images. Fig. 17 exhibits the DSTFN-derived time-series imagery at a 10-m resolution at the Dezhou site. In this case, the combined use of the two data sources led to a relatively dense time series and more efficient temporal variation characterization. As shown in Fig. 18, we mapped the NDVI dynamics of the crop pixel at the center of the region in Fig. 17 by using only Sentinel-2 imagery and combined imagery, respectively. Since the collected Sentinel-2 images did not cover May and June during which the land cover was transformed from winter wheat to maize, one growth circle was mistakenly interpreted when only Sentinel-2 imagery was used (Fig. 18(a)). By contrast, the combined sources offered 23 images in total and the Landsat 8 image acquired on June 20 recorded the maturity stage of winter wheat. Thus, the dense time-series images were more likely to distinguish the two crop growth circles (Fig. 18(b)). In sum, the combined dense images provide new implications and opportunities for time-series applications. Launched recently, Landsat 9 (Masek et al., 2020) can further improve the temporal frequency when it is involved in the data processing chain in future applications.

5.3. Benchmark dataset for Landsat 8 and Sentinel-2 fusion

Some methods have been proposed to merge Landsat 8 and Sentinel-2 observations, and more are expected in the future. The cross-comparison of models is essential in establishing guidelines for choosing ideal approaches. In this study, we offered two datasets that can be potentially employed as standard datasets for model assessment. The two datasets, Hailar and Dezhou, contain 23 and 24 scenes, respectively, and involve bands across visible, near infrared, and shortwave infrared ranges. Generally, the two datasets cover diverse landscapes (homogeneous land covers such as woodlands and heterogeneous land covers such as urban buildings), and reveal various surface dynamics (mild changes such as forest phenological changes and abrupt changes such as crop harvesting activities); thus, they are ideal datasets for assessing the models under scenarios across different levels of spatial and temporal variations.

5.4. Generalized definition of spatiotemporal data fusion

Traditionally, spatiotemporal data fusion is defined as the fusion of observations from two sensors with the complementary spatial and temporal resolution, i.e., a sensor with a fine spatial resolution and a sparse temporal coverage, and another with a coarse resolution but a frequent coverage (Zhu et al., 2018). A typical example is the widely-used MODIS–Landsat fusion that combines daily 500-m MODIS and 16-day 30-m Landsat observations to yield daily Landsat-like imagery (Gao et al., 2006). However, as new remote sensors emerge, the traditional definition shows limitations. For example, the fusion of Landsat 8 and Sentinel-2 can produce 10-m dense time series, but compared with Sentinel-2, Landsat has a coarser resolution and a sparser frequency. Thus, we suggested a more generalized spatiotemporal fusion definition as the fusion of two or more sensors with different spatial resolutions and temporal coverages. In this context, we do not strictly require the sensors to have complementary spatial and temporal resolutions, and the combination of Landsat 8 and Sentinel-2 observations can be included in spatiotemporal data fusion.

6. Conclusion

In this work, a deep network is proposed to produce 10-m dense time-series imagery by merging Landsat 8 and Sentinel-2 observations. The presented DSTFN model takes advantage of residual dense blocks and attention mechanism modules to enhance the feature representation and extraction. A constraint term designed on the basis of the image degradation process is embedded into the loss function, thus enabling the model to maximize the use of coarse-resolution imagery and ease the temporal variation issue. A series of experiments based on resolution-degraded data and original data indicate that DSTFN robustly outperforms the benchmark methods and stably achieves the state-of-the-

art performance. Generally, the proposed method can effectively downscale Landsat imagery to 10 m and produce dense time series by combing the downscaled Landsat imagery with Sentinel-2 imagery. A case study of mapping NDVI annual variations is also provided to illustrate the potential of the 10-m dense time series to reflect crop temporal dynamics at the field scale. The two experimental datasets in this work can be employed as standard datasets for future model validation. The experimental datasets and the DSTFN code can be accessed at https://github.com/andywu456 or by sending requests to the authors.

CRediT authorship contribution statement

Jingan Wu: Conceptualization, Methodology, Data curation, Software, Investigation, Formal analysis, Writing – original draft. **Liupeng Lin:** Methodology, Software, Formal analysis, Writing – original draft. **Tongwen Li:** Visualization, Investigation, Formal analysis. **Qing Cheng:** Resources, Supervision. **Chi Zhang:** Visualization, Investigation. **Huanfeng Shen:** Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing.

Declaration of Competing Interest

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.

Acknowledgments

The research was supported by the National Key Research and Development Program of China (No. 2019YFB2102900), the Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515110567) and a grant from the State Key Laboratory of Resources and Environmental Information System. We acknowledge the NASA and ESA teams for free access to Landsat 8 and Sentinel-2 data and the authors of ATPRK and ESRCNN for providing the source codes. We would also like to thank Koi Lin from Wuhan University for his suggestions on the writing of this manuscript.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jag.2022.102738.

References

Agapiou, A., 2020. Evaluation of Landsat 8 OLI/TIRS Level-2 and Sentinel 2 Level-1C Fusion Techniques Intended for Image Segmentation of Archaeological Landscapes and Proxies. Remote Sens. 12 (3), 579.

Ao, Z., Sun, Y., Xin, Q., 2021. Constructing 10-m NDVI Time Series From Landsat 8 and Sentinel 2 Images Using Convolutional Neural Networks. IEEE Geosci. Remote Sens. Lett. 18 (8), 1461–1465.

Bhogendra, M., Tej Bahadur, S., 2021. Deep learning-based framework for spatiotemporal data fusion: an instance of Landsat 8 and Sentinel 2 NDVI. J. Appl. Remote Sens. 15 (3), 1–13.

Chen, B., Li, J., Jin, Y., 2021a. Deep Learning for Feature-Level Data Fusion: Higher Resolution Reconstruction of Historical Landsat Archive. Remote Sensing 13 (2), 167.

Chen, N., Tsendbazar, N.-E., Hamunyela, E., Verbesselt, J. and Herold, M., 2021b. Sub-annual tropical forest disturbance monitoring using harmonized Landsat and Sentinel-2 data. International Journal of Applied Earth Observation and Geoinformation, 102: 102386.

Claverie, M., Ju, J., Masek, J.G., Dungan, J.L., Vermote, E.F., Roger, J.-C., Skakun, S.V., Justice, C., 2018. The Harmonized Landsat and Sentinel-2 surface reflectance data set. Remote Sens. Environ. 219, 145–161.

Dong, T., Liu, J., Qian, B., He, L., Liu, J., Wang, R., Jing, Q., Champagne, C., McNairn, H., Powers, J., Shi, Y., Chen, J.M., Shang, J., 2020. Estimating crop biomass using leaf area index derived from Landsat 8 and Sentinel-2 data. ISPRS J. Photogramm. Remote Sens. 168, 236–250.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., Bargellini, P., 2012. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. Remote Sens. Environ. 120, 25–36.

Gao, F., Masek, J., Schwaller, M., Hall, F., 2006. On the blending of the Landsat and MODIS surface reflectance: predicting daily Landsat surface reflectance. IEEE Trans. Geosci. Remote Sens. 44 (8), 2207–2218.

Gutman, G., Byrnes, R.A., Masek, J., Covington, S., Justice, C., Franks, S., Headley, R., 2008. Towards monitoring land-cover and land-use changes at a global scale: the global land survey 2005. Photogramm. Eng. Remote Sens. 74 (1), 6–10.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. In: In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.

Huang, G., Liu, Z., Maaten, L.V.D., Weinberger, K.Q., 2017. Densely Connected Convolutional Networks. In: In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269.

Irons, J.R., Dwyer, J.L., Barsi, J.A., 2012. The next Landsat satellite: The Landsat Data Continuity Mission. Remote Sens. Environ. 122, 11–21.

Ju, J., Roy, D.P., 2008. The availability of cloud-free Landsat ETM+ data over the conterminous United States and globally. Remote Sens. Environ. 112 (3), 1196–1211.

Korhonen, L., Hadi, Packalen, P., Rautiainen, M., 2017. Comparison of Sentinel-2 and Landsat 8 in the estimation of boreal forest canopy cover and leaf area index. Remote Sens. Environ. 195, 259–274.

Li, J., Roy, D.P., 2017. A Global Analysis of Sentinel-2A, Sentinel-2B and Landsat-8 Data Revisit Intervals and Implications for Terrestrial Monitoring. Remote Sensing 9 (9), 902.

Li, Z., Zhang, H.K., Roy, D.P., Yan, L., Huang, H., Li, J., 2017. Landsat 15-m Panchromatic-Assisted Downscaling (LPAD) of the 30-m Reflective Wavelength Bands to Sentinel-2 20-m Resolution. Remote Sensing 9 (7), 755.

Lin, L., Li, J., Shen, H., Zhao, L., Yuan, Q., Li, X., 2022a. Low-Resolution Fully Polarimetric SAR and High-Resolution Single-Polarization SAR Image Fusion Network. IEEE Trans. Geosci. Remote Sens. 60, 1–17.

Lin, L., Shen, H., Li, J., Yuan, Q., 2022b. FDFNet: A Fusion Network for Generating High-Resolution Fully PolSAR Images. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Liu, L., Xiao, X., Qin, Y., Wang, J., Xu, X., Hu, Y. and Qiao, Z., 2020. Mapping cropping intensity in China using time series Landsat and Sentinel-2 images and Google Earth Engine. Remote Sensing of Environment, 239: 111624.

Luo, X., Tong, X., Hu, Z., 2021. Improving Satellite Image Fusion via Generative Adversarial Training. IEEE Trans. Geosci. Remote Sens. 59 (8), 6969–6982.

Ma, Y., Wei, J., Tang, W. and Tang, R., 2021. Explicit and stepwise models for spatiotemporal fusion of remote sensing images with deep neural networks. International Journal of Applied Earth Observation and Geoinformation, 105: 102611.

Masek, J.G., Wulder, M.A., Markham, B., McCorkel, J., Crawford, C.J., Storey, J. and Jenstrom, D.T., 2020. Landsat 9: Empowering open science and applications through continuity. Remote Sensing of Environment, 248: 111968.

Pan, L., Xia, H., Yang, J., Niu, W., Wang, R., Song, H., Guo, Y. and Qin, Y., 2021. Mapping cropping intensity in Huaihe basin using phenology algorithm, all Sentinel-2 and Landsat images in Google Earth Engine. International Journal of Applied Earth Observation and Geoinformation, 102: 102376.

Pouliot, D., Latifovic, R., Pasher, J., Duffe, J., 2018. Landsat Super-Resolution Enhancement Using Convolution Neural Networks and Sentinel-2 for Training. Remote Sensing 10 (3), 394.

Roy, D.P., Huang, H., Boschetti, L., Giglio, L., Yan, L., Zhang, H.H. and Li, Z., 2019. Landsat-8 and Sentinel-2 burned area mapping - A combined sensor multi-temporal change detection approach. Remote Sensing of Environment, 231: 111254.

Roy, D.P., Wulder, M.A., Loveland, T.R., C.e., W., Allen, R.G., Anderson, M.C., Helder, D., Irons, J.R., Johnson, D.M., Kennedy, R., Scambos, T.A., Schaaf, C.B., Schott, J.R., Sheng, Y., Vermote, E.F., Belward, A.S., Bindschadler, R., Cohen, W.B., Gao, F., Hipple, J.D., Hostert, P., Huntington, J., Justice, C.O., Kilic, A., Kovalskyy, V., Lee, Z. P., Lymburner, L., Masek, J.G., McCorkel, J., Shuai, Y., Trezza, R., Vogelmann, J., Wynne, R.H., Zhu, Z., 2014. Landsat-8: Science and product vision for terrestrial global change research. Remote Sens. Environ. 145, 154–172.

Sánchez-Espinosa, A., Schröder, C., 2019. Land use and land cover mapping in wetlands one step closer to the ground: Sentinel-2 versus landsat 8. J. Environ. Manage. 247, 484–498.

Shao, Z., Cai, J., Fu, P., Hu, L. and Liu, T., 2019. Deep learning-based fusion of Landsat-8 and Sentinel-2 images for a harmonized surface reflectance product. Remote Sensing of Environment, 235: 111425.

Shen, H., Lin, L., Li, J., Yuan, Q., Zhao, L., 2020. A residual convolutional neural network for polarimetric SAR image super-resolution. ISPRS J. Photogramm. Remote Sens. 161, 90–108.

Shen, H., Meng, X., Zhang, L., 2016. An Integrated Framework for the Spatio–Temporal–Spectral Fusion of Remote Sensing Images. IEEE Trans. Geosci. Remote Sens. 54 (12), 7135–7148.

Shen, H., Wu, J., Cheng, Q., Aihemaiti, M., Zhang, C., Li, Z., 2019. A Spatiotemporal Fusion Based Cloud Removal Method for Remote Sensing Images With Land Cover Changes. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12 (3), 862–874.

Tong, W., Chen, W., Han, W., Li, X., Wang, L., 2020. Channel-Attention-Based DenseNet Network for Remote Sensing Image Scene Classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 4121–4132.

Wang, Q., Blackburn, G.A., Onojeghuo, A.O., Dash, J., Zhou, L., Zhang, Y., Atkinson, P. M., 2017. Fusion of Landsat 8 OLI and Sentinel-2 MSI Data. IEEE Trans. Geosci. Remote Sens. 55 (7), 3885–3899.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. CBAM: Convolutional Block Attention Module. In: In: 2018 European Conference on Computer Vision (ECCV), pp. 3–19.

Woodcock, C.E., Allen, R., Anderson, M., Belward, A., Bindschadler, R., Cohen, W., Gao, F., Goward, S.N., Helder, D., Helmer, E., Nemani, R., Oreopoulos, L., Schott, J., Thenkabail, P.S., Vermote, E.F., Vogelmann, J., Wulder, M.A., Wynne, R., 2008. Free Access to Landsat Imagery. Science 320 (5879), 1011.

Wu, J., Cheng, Q., Li, H., Li, S., Guan, X., Shen, H., 2020. Spatiotemporal Fusion With Only Two Remote Sensing Images as Input. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 6206–6219.

Zhang, H.K., Roy, D.P., Yan, L., Li, Z., Huang, H., Vermote, E., Skakun, S., Roger, J.-C., 2018. Characterization of Sentinel-2A and Landsat-8 top of atmosphere, surface, and nadir BRDF adjusted reflectance and NDVI differences. Remote Sens. Environ. 215, 482–494.

Zhu, X., Cai, F., Tian, J., Williams, T.-K.-A., 2018. Spatiotemporal Fusion of Multisource Remote Sensing Data: Literature Survey, Taxonomy, Principles, Applications, and Future Directions. Remote Sens. 10 (4), 527.