Remote Sensing Applications: Society and Environment 32 (2023) 101005

Contents lists available at ScienceDirect

Remote Sensing Applications: Society and Environment

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

A review of remote sensing image spatiotemporal fusion: Challenges, applications and recent trends

Juan Xiao [a], Ashwani Kumar Aggarwal [b], Nguyen Hong Duc [c], Abhinandan Arya [d], Uday Kiran Rage [e], Ram Avtar [a, f, *]

[a] Graduate School of Environmental Science, Hokkaido University, Sapporo, Japan
[b] Electrical and Instrumentation Engineering Department, Sant Longowal Institute of Engineering and Technology, Punjab, India
[c] Department of Water Resources, College of Environment and Natural Resources, Can Tho University, Can Tho, Viet Nam
[d] Synspective Inc., Koto-ku, Miyoshi, Tokyo, 135-0022, Japan
[e] Division of Information Systems, The University of Aizu, Aizu Wakamatsu, Fukushima, Japan
[f] Faculty of Environmental Earth Science, Hokkaido University, Sapporo, Japan

ARTICLE INFO

Keywords:
Spatiotemporal fusion
Spatial resolution
Temporal resolution
Deep learning

ABSTRACT

In remote sensing (RS), use of single optical sensors is frequently inadequate for practical Earth observation applications (e.g., agricultural, forest, ecology monitoring) due to trade-offs between spatial and temporal resolution. The advent of spatiotemporal fusion (STF) of RS images has allowed the production of images with high resolution at both spatial and temporal scales. Despite the development of more than 100 STF models in the past two decades, many of these models have not been practically applied due to the possibility of limited understanding of the models. Therefore, this study aims to provide a comprehensive review of STF methods, including their conception, development, challenges, and applications. This study focuses primarily on deep learning-based STF models, which achieved superior performance and significantly increased the number of STF models. This review can guide the selection and design of STF models, as well as proposes future directions for STF modeling. The findings of this review facilitate further STF research to improve the accuracy and application of fused RS images in the field of agriculture, forestry, and ecological monitoring.

1. Introduction

Remote sensing (RS) images have been widely used in various fields such as agriculture, forestry, and ecology field for their ability to capture rich information from the Earth's surface (Mustafa et al., 2011; Pettorelli et al., 2014; Xiao et al., 2023). RS data can be acquired using spaceborne (e.g., satellites), airborne (e.g., Unmanned Aerial Vehicles [UAVs]), and ground-based (e.g., hand-held devices) platforms, on which various sensors can be mounted, covering different levels of Earth observation. As the number of RS platforms and images has increased exponentially in recent decades, big data and data science have emerged as associated fields (Ghamisi et al., 2019). However, due to physical constraints and other factors, individual RS optical imaging sensors typically have trade-offs between spatial and temporal resolution (Ghassemian, 2016; Tan et al., 2019). For example, the Landsat Thematic Mapper (TM) has a spatial resolution of 30 m and a temporal resolution of 16 days. The temporal resolution of Landsat TM thus limits the frequency of Earth observations. As a complementary data source, the Moderate-resolution Imaging Spectroradiometer (MODIS) mounted on the Terra and Aqua satellites provides daily revisits, with spatial resolution ranging from 250 to 1000 m. The large amount of RS data

* Corresponding author. Faculty of Environmental Earth Science, Hokkaido University, Sapporo, Japan.
E-mail address: ram@ees.hokudai.ac.jp (R. Avtar).

https://doi.org/10.1016/j.rsase.2023.101005
Received 10 March 2023; Received in revised form 31 May 2023; Accepted 1 June 2023
Available online 3 June 2023
2352-9385/© 2023 Published by Elsevier B.V.

with limited capability on both spatial and temporal resolution scales has raised concerns about the effective processing and application of these big data to meet burgeoning demands. In turn, the increasing demand for high spatial and temporal resolution images for Earth observation has driven the development of spatiotemporal fusion (STF) technique in processing RS images. Therefore, a comprehensive understanding of STF is essential to leverage their potential for applications in Earth observation.

STF is the fusion of images with high spatial and low temporal resolution (hereafter, fine images) and those with a high temporal and low spatial resolution (hereafter, coarse image) to simulate images with high spatiotemporal resolution. Fine and coarse input images are captured by different sensors and cover the same area, which is a prerequisite for STF. In addition, input image preprocessing, including radiometric calibration, atmospheric correction, reprojection, co-registration, is also required before images are fed into the STF model. A schematic diagram of different inputs for STF process is provided in Fig. 1. Four categories of input images are used in STF models: three input images, including one fine-coarse image pair for reference date t_1 and a coarse image for prediction date t_2 (Gao et al., 2006; Wang and Atkinson, 2018; Zhu et al., 2016); five input images including two fine-coarse image pairs for reference dates t_1 and t_3, and a coarse image for prediction date t_2 (Liu et al., 2019; Peng et al., 2012; Zhu et al., 2010); two input images including one fine image for reference date t_1 and one coarse image for prediction date t_2 that is used to predict a fine image at t_2 (Fung et al., 2019; Y. Sun et al., 2019; Tan et al., 2021); and a time series of input images where time series coarse images are used to generate time series fine images (Amorós-López et al., 2013; M. Wu et al., 2015; Zhou et al., 2020). The first two categories are referred to as universal input in STF, and five images as input provide better performance in predicting the reflectance in land cover change and heterogeneous landscapes (Wei et al., 2017b). However, the approach uses five images as input has limited applicability in areas with severe cloud contamination and cannot predict images for near-real-time applications because the fusion process includes hindcasting (Hazaymeh and Hassan, 2015). This issue also limits the application of time-series input images, which has additional requirements regarding the number of input images. Higher demands are placed on the STF model when two images are used as input, leading to unsatisfactory fusion accuracy (W. Li et al., 2020). However, few input images facilitate STF in areas with high cloud contamination.

Image fusion can be performed at the pixel, feature, or decision levels. Pixel-level fusion refers to the direct fusion of raw or preprocessed images, which involves resampling coarse image resolution to match fine image resolution. However, this approach is highly sensitive to the accuracy of input image co-registration and is time-consuming (Li et al., 2017). Nonetheless, pixel-level fusion retains the most image information, and significant progress has been made in this field to date (Dogra et al., 2017). Feature-level or object-oriented fusion focuses on extracting features of interest from source images before the fusion process. Feature-level fusion is more sensitive to input noise and less sensitive to misregistration issues, as it eliminates unnecessary features and retains important information (S. Li et al., 2017). Both the feature extraction method and the features' robustness are considered important factors influencing the result of feature-level fusion. Decision-level fusion is performed after feature interpretation of the input images to obtain more reliable decision knowledge. However, the high information loss rate is a major disadvantage of decision-level fusion (Xiao et al., 2020).

STF is an ill-posed problem because the results are not unique and are not real but a simulation produced by an uncertain inference process (Ghamisi et al., 2019). STF models have been proposed based on various theories during the past two decades; however, most have not been incorporated into practical Earth observation applications. Therefore, the goal of this review is to explore the past development of STF models, their advantages and disadvantages, and the key challenges of the different STF mehods. Additionally, the applications of proposed STF models were also surveyed and clarified. Future directions for STF modeling were then proposed, such as the requirement of benchmark datasets, the promotion of practical Earth observation applications, and the inclusion of cen-

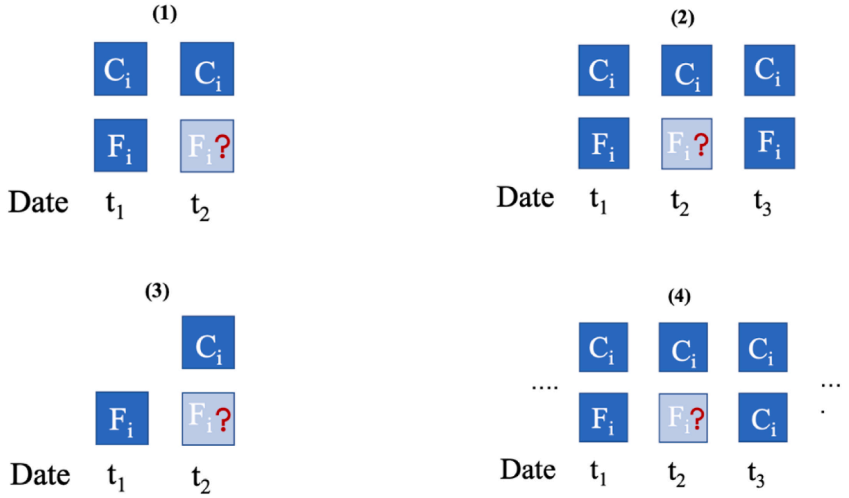

Fig. 1. Schematic diagram of different inputs for spatiotemporal fusion (STF), where C_i and F_i are coarse fine images, respectively, and $t_1 < t_2 < t_3$. Blue squares denote the inputs, and the output is represented by a light blue square with a question mark. (1) three input images; (2) five input images; (3) two input images, and (4) time series input images. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

timeter-scale images acquired by unmanned aerial vehicles. As a result, this review can serve as a valuable resource for guiding the selection and design of STF models. Furthermore, it has the potential to promote the application of these models, thereby contributing to the advancement of STF in Earth observation applications.

2. Methodology

In response to the research concerns and to document the literature review process, this study adheres to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher et al., 2009). The implementation process of this study includes three steps: (1) identification, (2) screening, and (3) eligibility assessment. Fig. 2 depicts the detailed procedure for conducting the literature search. Firstly, to collect the publications for review, this study performed a keyword search in the Web of Science (https://www.webofscience.com) database using the advanced search function. The following Boolean string was used (((ALL= (spatiotemporal fusion)) OR ALL= (high spatial and temporal image fusion)) AND ALL= (remote sensing)). A manual search with keywords in Google Scholar was conducted to identify relevant publications, including non-indexed journals. In this review, relevant information and materials which are non-conventional publications will be considered grey literature. By 9 October 2022, the total identification result was 919 publications. Then, these identified publications were screened by reading their abstract and keywords to answer the following questions: (1) is the research proposed an STF model? and (2) is the research related to the application of the STF model? After screening and removing duplicate publications, a total of 351 results were discovered to meet the defined inclusion criteria. In the final assessment of eligibility, two more criteria were considered: (1) literature was written in English for the international community, and (2) publications with full text for detailed information. Consequently, 203 publications were kept for review, of which 194 peer-reviewed articles, and 6 books/book chapters were directly falling under the scope of this review. Therefore, the full-text review was performed from these 203 publications. The review process involved collecting key information, such as the published year, the definition of the proposed STF model, its advantages and disadvantages, and its application.

3. Results

3.1. Basic terminology

Optical imaging system parameters such as the signal-to-noise ratio (SNR) and modulation transfer function (MTF) make it challenging to obtain high resolution in both spatial and temporal domains simultaneously, which limits the application of RS in many fields such as agricultural and forest monitoring (Zhang and Ma, 2021). Thus, some STF models have incorporated these parameters during the design process. This subsection clarifies terms related to spatial and temporal resolution, pixels, endmembers, and platforms and sensors commonly used in STF modeling.

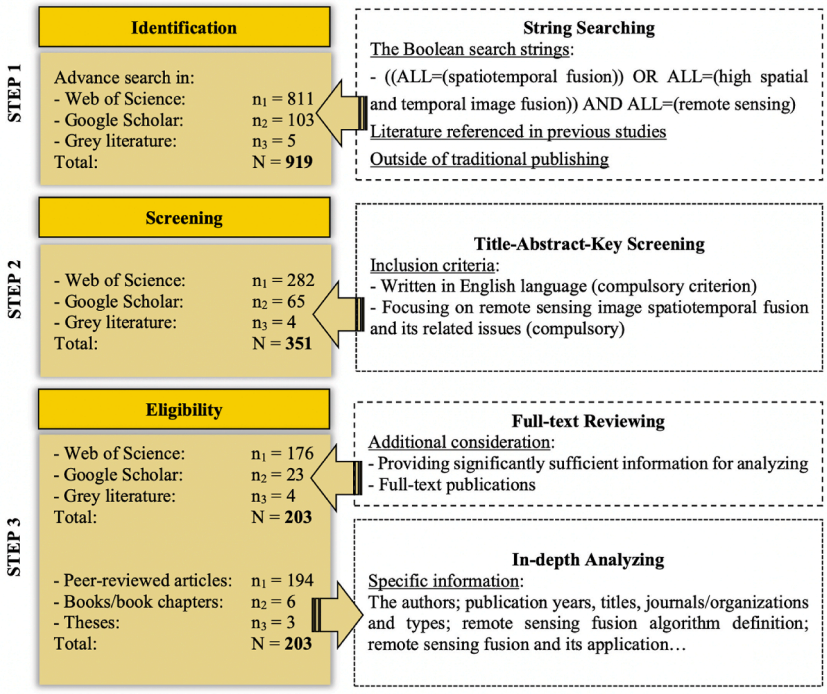

Fig. 2. Flowchart of the systematic review methodology.

3.1.1. Spatial and temporal resolution

In RS modeling, the spatial resolution represents the ability to detect the Earth's surface in detail. In this context, spatial resolution does not refer to pixel resolution, which can be altered by resampling. The spatial resolution of RS images is related to the point spread function (PSF) and MTF of the optical imaging system. PSF describes the ability of an optical system to resolve a point source. A point source forms an expanded image point due to diffraction after passing through any optical system. Under the assumption that the optical system is linear and shift-invariant, the spatial response of an optical system is characterized by a PSF (Markham, 1984). Therefore, image information may be extracted more accurately by measuring the PSF of the system. MTF is defined as the Fourier transform (FT) of the PSF under the assumption of linear and shift invariance. Thus, the PSF is one of the most comprehensive criterion used to evaluate the performance judgment of optical imaging systems (Zhao and Rowlands, 2014). The temporal resolution of an RS image refers to the minimum revisit time interval of a sensor, with a large revisit time corresponding to low temporal resolution and vice versa. High-temporal-resolution images are desirable for areas with frequent changes or where time-series monitoring is required.

3.1.2. Pixel and endmember

Pixel and endmember values are critical for understanding object features captured by RS. The pixel is the smallest unit of RS images, and a mixed pixel is a combination of multiple features captured in a single pixel. For example, Fig. 3 represents a mixed pixel consisting of ground surface features A, B, C, and D. Mixed pixels occur when the spatial resolution of a sensor is relatively lower than the scale of spatial heterogeneity at the ground surface (Shi and Wang, 2016). Endmembers, on the other hand, are pure features such that a mixed pixel can have multiple endmembers. Therefore, a mixed pixel can be decomposed to describe its endmembers quantitatively. Endmember change and land cover change represent temporal changes in reflectance (X. Li et al., 2020).

3.1.3. Remote sensing platforms and sensors commonly used in spatiotemporal fusion

Landsat and MODIS images are commonly used for STF. Two open-source datasets used Landsat and MODIS images from the Coleambally Irrigation Area (CIA) and the Lower Gwydir Catchment (LGC) study sites, respectively, have been extensively used for STF. The CIA study site is a rice-based irrigation system in southern New South Wales (NSW, Australia, 34.0034°E, 145.0675°S). It covers large irrigated croplands with irregular shapes and other land types, such as dryland agriculture and woodlands (Ao et al., 2022). The CIA dataset has 17 cloud-free Landsat 7 ETM -MODIS (MOD09GA) pairs with 6 bands and an image size of 1720 × 2040, during the austral summer growing season from October 2001 to May 2002. It can be used to evaluate the performance of STF in heterogeneous cropland landscapes. The LGC study site is located in northern New South Wales (NSW, 149.2815°E, 29.0855°S), and contains 14 cloud–free Landsat 5 TM-MODIS (MOD09GA) pairs with 6 bands and an image size of 3200 × 2720 from April 2004 to April 2005 (Emelyanova et al., 2013). This area is mainly covered by croplands, bare soil, and natural vegetation. Additionally, flood events and abrupt changes in land cover types on the predicted date can be observed at the LGC study site (Ao et al., 2022). Therefore, the LGC dataset was commonly used to evaluate fusion performance in abruptly changing areas.

A study by Li et al. (2020) provides a new STF datasets benchmark that consists of three Landsat-MODIS datasets with three essential characteristics: (1) diversity of regions, (2) long timespan, and (3) challenging scenarios. As a result, they made a significant contribution to evaluating the STF models. Apart from Landsat and MODIS datasets, some studies have used a higher spatial resolution satellite images in recent years, such as sentinel 2 (Liu et al., 2019; Sadeh et al., 2020; Shao et al., 2019) and GaoFen satellite images (Cui et al., 2018; Ge et al., 2020; Jiang et al., 2020). In addition, considering the uncertainties introduced by radiometric correction and geometric registration, some studies have used simulated images as inputs (Peng et al., 2012; Y. Sun et al., 2019; Wang and Atkinson, 2018). Table 1 lists the most commonly used datasets for STF.

3.2. Overview of STF

Based on the characteristics of the model framework and the procedures used to implement the models, Chen et al. (2015) classified the STF models into those based on transformation, reconstruction, and learning processes. It discussed the advantages and disadvantages of each category and evaluated the fusion performance of four STF models. However, their comparison experiment did not

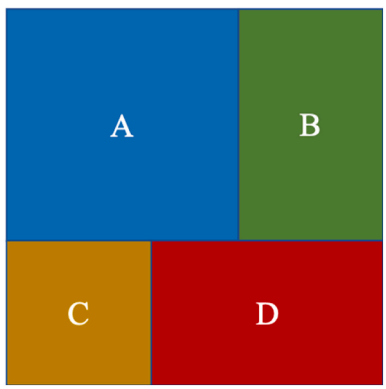

Fig. 3. Representation of a single-pixel containing four endmembers.

Table 1
Summary of commonly used remote sensing (RS) datasets for spatiotemporal fusion (STF).

Platform	Sensor	Launch Date	Spatial resolution (m)	Revisit frequency (days)
Landsat-8	OLI	2013	VNIR/TIR:30/100	16
Landsat-7	ETM +	1999	VNIR/TIR:30/60	16
Landsat-4/5	TM	1982/1984	VNIR/TIR:30/120	16
Terra/Aqua	MODIS	1999/2002	B1-2: 250 B3-7: 500 B8-36: 1000	0.5
GF-1	PAN/MSC	2013	VNIR: 8	4
GF-2	PAN/MSC	2014	VNIR: 4	5
Sentinel-2A/2B	MSI	2015	B2, 3, 4, 8: 10 B5, 6, 7, 8A, 11, 12: 20	10
NOAA	AVHRR	1978	1100	0.5
ENVISAT	MERIS	2002	300	35
Terra	ASTER	1999	VNIR/SWIR/TIR: 15/30/90	16

(VNIR, visible and near-infrared; SWIR, shortwave infrared; TIR, thermal infrared).

cover all three categories, and Bayesian and deep learning (DL) methods were not mentioned in this study Chen et al. (2015). A similar study by Belgiu and Stein (2019) divided the STF models into three categories and examined their ability to respond to gradual land surface changes, but Bayesian and DL methods were only briefly mentioned. Based on the specific methodology used to link coarse and fine images, Zhu et al. (2018) classified STF models into five categories. Namely, weighted function-based, unmixing-based, Bayesian-based, learning-based, and hybrid-based methods. This study adequately surveyed and discussed the principles underlying each category, but the development of DL methods and various neural networks for STF received little attention. Subsequently, Li et al. (2020) divided existing STF methods into the same five categories and provided three Landsat 8 OLI-MODIS datasets as the benchmark to test the accuracy of STF models. Previous studies have adopted several classification methods to categorize STF models (Shen et al., 2016b; Y. Sun et al., 2019). In this study, the STF classification method developed by Zhu et al. (2018) was used, with the Bayesian approach being classified under the learning-based category.

About 146 STF models have been proposed within the collected literature, among which learning-based models account for 43% and weight function-based methods account for 27%. Hybrid and unmixing-based models are the rarest models, representing only 15% of the total, respectively (Fig. 4). The rapid increase in the number of STF models in recent years suggests that an up-to-date and comprehensive review is required to summarize the progress in STF model development.

Over the past decade, there has been a significant surge in the number of proposed STF models. This rise can be attributed to the increasing demands for Earth observation applications necessitating high spatial and temporal resolution images, as well as the development of advanced theoretical and massive hardware computing power. Zhukov et al. (1999) proposed the first unmixing-based STF model, subsequent unmixing-based models began to appear in 2008, and their number has increased steadily since 2011. Gao et al. (2006) proposed the first weight function-based STF model in 2006, which has been cited up to 1500 times. The number of weight function-based models predominates during the period 2005–2016. Since its introduction in 2012, the number of learning-based models has become increasingly dominant, particularly since 2017. DL was first implemented in STF in 2015 and has contributed to excellent fusion performance, such that learning-based approaches have gradually become mainstream in the STF field. Hybrid methods combine methods belonging to the other three categories. The number of hybrid STF models has expanded relatively quickly since their introduction during 2014–2016 due to their better performance over single STF models. The numbers of unmixing-based and weight function-based models have not grown as rapidly as those of learning and hybrid-based models, likely due to their dependence on assumptions and the low accuracy of the fusion results.

3.2.1. Unmixing-based method

Unmixing-based STF methods rely on the linear spectral mixing theory, which extracts endmembers reflectance from prior fine images and predicts the information of fine image pixels by unmixing the coarse image pixels on the prediction date (Settle and Drake, 1993; M. Wu et al., 2015). In unmixing-based models, the fine reference image is typically supplemented by an unsupervised land cover classification map or an open-source high spatial resolution land cover map. Fig. 5 illustrates the conceptual model of the

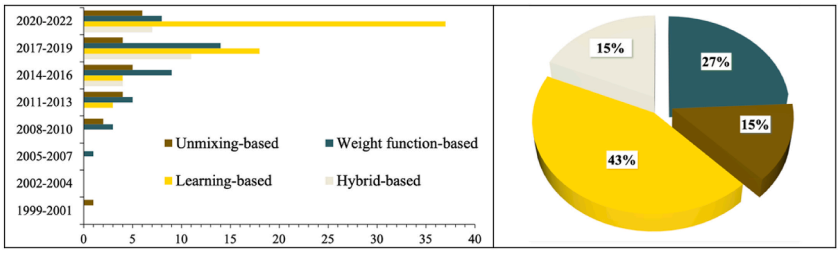

Fig. 4. The number and ratio of STF model proposals across four categories.

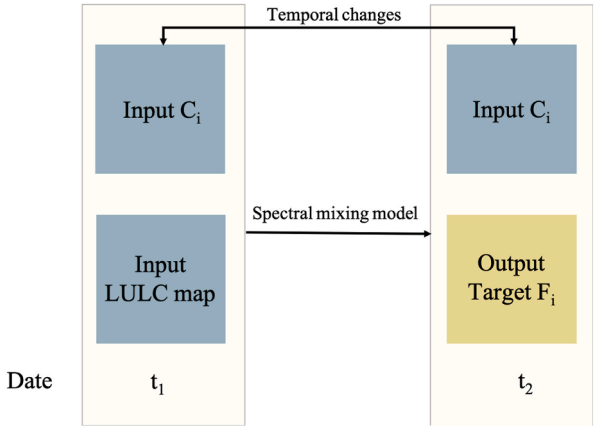

Fig. 5. Conceptual model of unmixing-based method (t1 < t2).

unmixing-based method, including the temporal changes in coarse input images and the spectral mixing model between the coarse image and the land use land cover (LULC) map of a fine image at t_1. The spectral mixing model, along with information about temporal changes, is used to predict the fine image at t_2.

The first unmixing-based STF model, multisensor multiresolution technique (MMT), was proposed in 1999 by Zhukov et al. (1999). MMT includes four operations: (1) land cover classification of the fine image on the reference date; (2) calculation of the proportions of each land cover class in each coarse image pixel; (3) unmixing of each coarse image pixel at the prediction date in the moving window mode; (4) reconstruction of the fine image on the prediction date. MMT has low spectral unmixing accuracy and low intra-class spectral variability. Therefore, other unmixing-based models were later developed to address these challenges and handle heterogeneous landscapes.

Within the proposed unmixing-based STF models, some assume constant temporal variation within land cover types and negligible within-class variability (Wu et al., 2012; Zurita-Milla et al., 2008). However, these assumptions are unreasonable because land surfaces are dynamic and heterogeneous. To address this limitation and improve prediction accuracy, mixed pixels are considered in both coarse and fine images (Amorós-López et al., 2011; Liu et al., 2020; Wang et al., 2021a). By performing fusion on a sliding window, the variability between windows is increased, whereas the intra-class variation within the window is eliminated (Amorós-López et al., 2013). Therefore, an adaptive window size has been adopted to reduce unmixing error (Liu et al., 2020; M. Wu et al., 2015). Furthermore, additional strategies have been used to reduce unmixing errors further. These strategies include considering image pairs before and after the prediction date (Huang and Zhang, 2014; Lu et al., 2016; M. Wu et al., 2015; Wu et al., 2012), classifying all input images (Xu et al., 2015; J. Yang et al., 2020; Zhong and Zhou, 2019), incorporating sensor bias into the models (Shi et al., 2022; M. Wu et al., 2015; J. Yang et al., 2020). By using these strategies within the unmixing-based STF model, their capability to capture abrupt land cover changes is enhanced (Chen et al., 2018; Huang and Zhang, 2014; Shi et al., 2022; J. Yang et al., 2020; Zhong and Zhou, 2019, 2018).

In the unmixing-based STF method, fine and coarse images are not required to have corresponding spectral bands (Wang et al., 2021b). Therefore, the number of unmixing-based STF models has steadily increased over time. However, the assumption of endmember consistency between reference and prediction dates (Jia et al., 2021; Liu et al., 2019) limits STF model application in landscapes with changing land cover. In addition, the rate of change in spectral reflectance of fine and coarse images is assumed to be the same, making applications in heterogeneous landscapes challenging (Ao et al., 2022; Liu et al., 2020). Because spectral variability is common in heterogeneous areas. Unreasonable assumptions are likely to result in large estimation errors and significant loss of spatial and spectral detail (Liu et al., 2019; Wang et al., 2021a), resulting in low fusion accuracy. Notably, the fusion results from unmixing-based STF models are commonly affected by the block effect (Wang et al., 2021b). As a result, the number of unmixing-based models is one of the rarest among the four STF categories.

The classification of fine images has a significant influence on the fusion results of the unmixing-based STF models. In particular, the ISODATA algorithm is a widely used LULC classification algorithm in unmixing-based models (Chen et al., 2018; Rao et al., 2015; Wu et al., 2012; Xu et al., 2015; Zhang et al., 2013; Zhong and Zhou, 2019; Zurita-Milla et al., 2008). Alternatively, maximum likelihood (Maselli et al., 2019; Wu et al., 2015), object-based (Chen et al., 2021), and other classification methods (Yang et al., 2020; Zhong and Zhou, 2019) have also been adopted. Several unmixing-based models used high spatial resolution land cover maps from Finer Resolution Observation and Monitoring of Global Land Cover (FROM–GLC) data (Wu et al., 2015), GlobeLand 30 product (Chen et al., 2018), or other auxiliary maps (Zurita-Milla et al., 2009) instead of classifying fine images. Some unmixing-based models also used fine image segmentation (Huang and Zhang, 2014).

To improve the fusion results of models in this category, alternative image classification approaches can be explored. Machine learning-based supervised classification algorithms such as support vector machine, random forest, and decision tree offer the advantage of not requiring pure endmembers (Wang et al., 2021a). However, in the absence of training data, unsupervised classification is a better option because it is more user-friendly and can be implemented automatically. The number of clusters is a key parameter for unsupervised classification. Typically, estimating the optimal number of clusters has relied on empirical or prior knowledge. To ad-

dress this limitation, Xie and Beni (2011) proposed the cluster validity index, known as XB index, to determine the optimal number of clusters in unsupervised classification by quantifying the compactness and separation of partitions. Unmixing-based model applications include normalized difference vegetation index (NDVI) products (Chen et al., 2018; Lu et al., 2016; Qiu et al., 2021; Rao et al., 2015), agricultural monitoring (Amorós-López et al., 2013), imperious surface mapping in urban areas (R. Chen et al., 2021), change detection (Lu et al., 2016), inland water monitoring (Mizuochi et al., 2017), and land surface temperature (LST) monitoring (J. Wang et al., 2020).

3.2.2. *Weight function-based method*

Weight function-based STF methods estimate pixel values in a desired fine image according to the weighted sum of adjacent similar pixels in the input images. The method of weight sum of adjacent similar pixels follows Tobler's First Law of Geography (Tobler, 1970), which states that "everything is related to everything else, but near things are more related than distant things". Therefore, the weights assigned to the pixels should decrease as the distance increases (K. Peng et al., 2022; Zhang et al., 2013). Typically, the weight function-based methods predict images by performing mathematical operations on similar endmembers within a moving window. Fig. 6 depicts the conceptual model of the weight function-based STF method, which includes the resampling of input coarse images, the temporal difference between two resampled coarse images, the spatial distance between the resampled coarse image and the fine image at reference date t_1, and a moving window with the central pixel and similar pixels. The number of STF models in this category has grown dramatically due to its simple and clear theoretical framework.

The weight function-based method divides the relationship assumptions between the input images into scale and temporal models (Fig. 7). The scale model (also referred to as structure similarity) assumes that the relationships between fine and coarse image pairs are invariant in the same period (Liu et al., 2019; Xue et al., 2017). It involves using the PSF of sensors to map fine image pixels to coarse image pixels and takes STF as the super-resolution problem (Zhu et al., 2018). This mapping, however, may not yield accurate predictions because the relationship between fine and coarse images may change during the observation period. Studies conducted by (Bhattarai et al., 2015; Liao et al., 2017) have adopted the scale model assumption. The temporal model (also called temporal dependence) assumes that the changes between coarse images are equivalent to the changes between fine images. However, as fine images capture detail changes more precisely than coarse images, changes between coarse images may differ from those changes between fine images within the same period (Zhang et al., 2018). Studies conducted by (Cheng et al., 2017; Hazaymeh and Hassan, 2015; Kwan et al., 2018; Liao et al., 2017; Malleswara Rao et al., 2015; Roy et al., 2008; Wang and Huang, 2018; Wang et al., 2017a) have adopted the temporal model to propose the STF model.

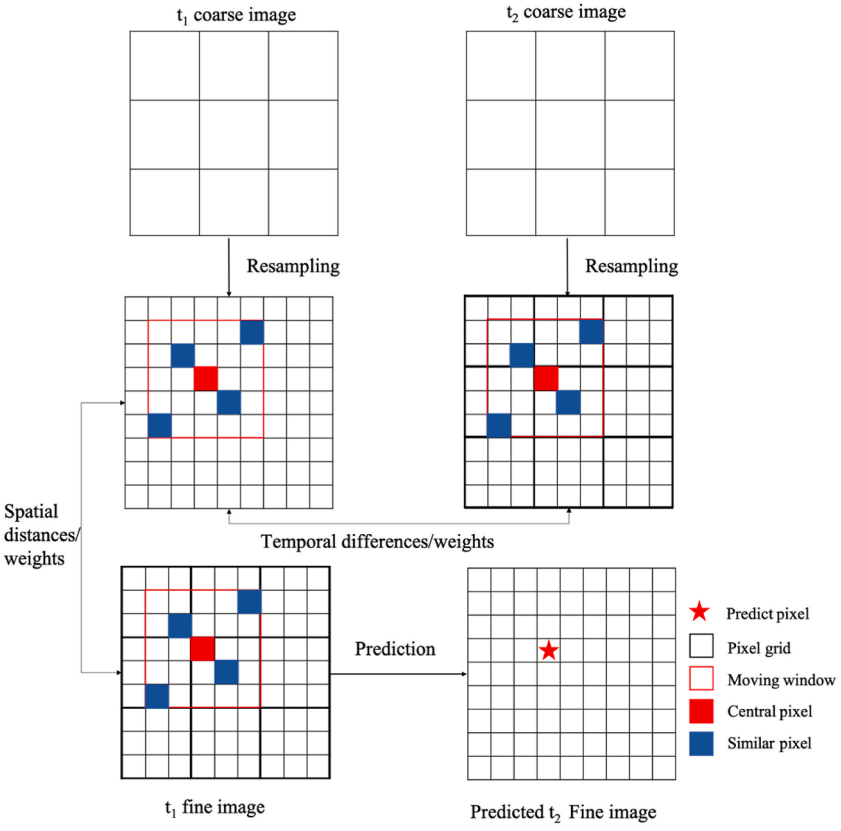

Fig. 6. Conceptual model of weight function-based STF methods ($t_1 < t_2$).

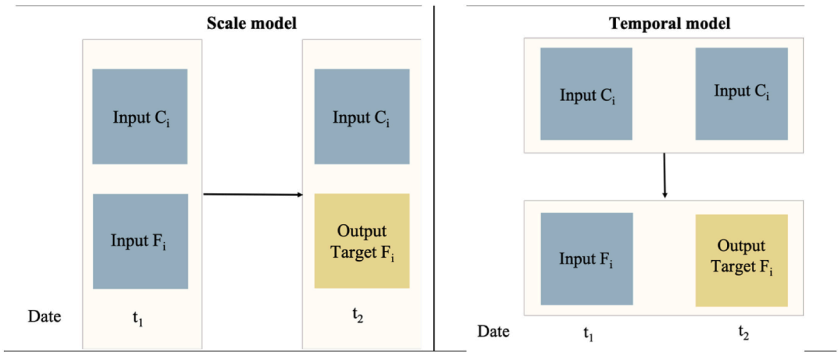

Fig. 7. Scale model and temporal model in STF.

The spatial and temporal adaptive reflectance fusion model (STARFM) was the first open-source weight function-based STF model (Zhu et al., 2018) and has been extensively compared with newly proposed models. The STARFM consists of three main steps: (1) selecting spectrally similar pixels within a window of interest using fine images; (2) determining a weighting function factor as a function of both fine and coarse images; (3) predicting fine images at prediction date. STARFM assumes that pixels in coarse images have only one endmember, which can capture phenological changes but has limitations in heterogeneous landscapes. Therefore, several modifications to the STARFM have been proposed, referred to as STARFM-based models. These models have been improved through different data processing procedures (Wang et al., 2014), selection of similar neighbor pixels (Fu et al., 2013; Zhao et al., 2018), expansion into land cover change applications (Hilker et al., 2009a; Wu et al., 2017) and heterogeneous landscape (Wu et al., 2017; Zhu et al., 2010). Additional techniques, such as ATPRK (Wang et al., 2017b) and super-resolution (Teo and Fu, 2021), have also been incorporated to modify STARFM. Furthermore, some studies have separately considered each spectral band (Ghosh et al., 2020) and accounted for sensor differences (Shen et al., 2013).

Weight function-based models have the flexibility to use different weight functions to improve STF accuracy. For example, the rigorously-weighted spatiotemporal fusion model (RWSTFM) employs ordinary kriging based on geostatistical assumptions to calculate the weights of neighboring pixels (Wang and Huang, 2017). Similarly, the spatio-temporal vegetation index image fusion model (STVIFM) uses a new disaggregation weighting system to predict the NDVI change for each fine pixel by calculating the total NDVI change within a moving window (Liao et al., 2017). To deal with strong seasonal changes, the Fit-Fc (Wang and Atkinson, 2018) conducted STF between Sentinel-2 and Sentinel-3 images, while Enhanced Fit-Fc (EFF) (Y. Li et al., 2022) was proposed to estimate leaf area index (LAI). Studies by Sun et al. (2019) and Zhang et al. (2021) have approached STF as the pansharpening problem and generated an initial fusion image by introducing a linear injection model. The high-resolution spatiotemporal image fusion method (HISTIF) was proposed by Jiang et al. (2020) for phenological change monitoring within crop fields, which includes filtering for cross-scale spatial matching and multiplicative modulation of temporal change.

The weight function-based models have the advantage of simple mathematical inference, which is simple to implement by performing arithmetic operations and searching similar image pixels within the moving windows. However, the current weight function-based models have the following disadvantages: (1) The acquirement dates of reference and prediction images should not be far apart; (2) Unreasonable assumptions of the relationship between input images; and (3) The weight function is highly dependent on assumptions. In detail, the time interval between reference and prediction should not be too long because the relationship between fine and coarse images learned from known pairs may change with time (Rao et al., 2015). Furthermore, assumptions made when deducing the relationship between input images may not be valid in reality. For example, some models, such as STARFM, ESTARFM, and STAVFM assume that no land cover changes occur during the study period and that changes in different sensors are nearly identical over time. However, these assumptions are unreasonable for the fact of land surface dynamics (Liu et al., 2018). In addition, some weight function-based models assume that the changes in spectral reflectance are linear during the study period (Cheng et al., 2017). This assumption may be only reasonable for a limited time and is difficult to apply in practice. In summary, weighting function-based STF methods are suitable for phenological change prediction, and their accuracy decreases in heterogeneous landscape and land cover change areas (Jia et al., 2021). The superior performance of other STF methods and the limitations of weight function-based theory have resulted in a stagnant development after 2019.

As summarized in Table 2, the STARFM has been used for various Earth surface monitoring applications, including monitoring vegetation, crops, and forests. It has also been applied in land use classification, evapotranspiration estimation, public health, and LST studies. Furthermore, several studies modified STARFM to generate LST (Huang et al., 2013; Weng et al., 2014; P. Wu et al., 2015; Wu et al., 2013; Xia et al., 2019); vegetation indices (Houborg et al., 2016; Liu et al., 2018; Meng et al., 2013), and surface soil moisture data (Xu et al., 2018).

3.2.3. Learning based method

Learning-based STF models predict images by learning the feature patterns of input images. A range of advanced theories, including Bayesian estimation, sparse representation, compressed sensing, machine learning, and DL, have contributed significantly to developing the learning-based STF models. In this study, the learning-based method includes Bayesian learning, sparse representation learning, machine learning, and DL. The earliest learning-based STF model was proposed in 2012 based on sparse representation

Table 2
Summarization of STARFM model's application.

Applications	Studies
Vegetation monitoring	(Hilker et al., 2009b); (Udelhoven, 2012); (Tian et al., 2013); (Zhang et al., 2016); (Yan et al., 2018).
Land use classification	(Watts et al., 2011); (M. Zhang et al., 2019); (R. Sun et al., 2019).
Crop monitoring	(Devendra Singh, 2011); (D. Singh, 2011).
Evapotranspiration	(Anderson et al., 2011); (Ke et al., 2017).
Public health study	Liu and Weng (2012).
Forest monitoring	(Walker et al., 2012); (Walker et al., 2015).
Land surface temperature	Shen et al. (2016a).

learning. Over the following 10 years, most proposed STF models were learning-based due to the widespread adoption of DL networks. Their numbers increased even more rapidly after 2017, with the number of such models proposed during 2020–2022 exceeding the total number proposed before 2020.

3.2.3.1. Bayesian learning. The STF models use Bayesian estimation theory to learn data distribution probability on the prediction date belonging to the Bayesian learning category. STF is an example of a maximum posterior (MAP) issue in this category. Several Bayesian-based STF models have been developed for STF, each with its unique approach to image prediction. Bayesian Maximum Entropy (BME) (Li et al., 2013) developed an error model to link fine and coarse images. Xue et al. (2017) combined the scale and temporal model of input images. Zhu et al. (2019) used a bias correction model in the robust Fixed Rank Filter (R–FRF) before developing a hierarchical Bayesian model based on R–FRF. The Best Linear Unbiased Estimation spatiotemporal fusion (BLUE) (Yin et al., 2018) provides the fine image as the background field and coarse images as the observation field to model the relationship and predict the image using the best linear unbiased estimator. Additionally, the HIghly Scalable Temporal Adaptive Reflectance Fusion Model (HISTARFM) (Moreno-Martínez et al., 2020) uses an optimal interpolator and a Kalman filter operating synergistically to reduce the amount of noise and decrease possible biases. Given that there are few fine images in a long time series and that coarse images contain the majority of the change information, He et al. (2017) and Yang et al. (2020) reconstruct fine difference images rather than generating fine images directly. These methods have demonstrated their performance in terms of land cover change and phenological change. The main advantage of Bayesian-based models is their ability to cope well with the uncertainty of input images and predict the most likely fine image. However, existing methods have either been developed for specific applications (e.g., sea surface temperature (Li et al., 2013; Zhu et al., 2019), NDVI (Yin et al., 2018), land vegetation monitoring (Moreno-Martínez et al., 2020)) or have more stringent requirements for input images (Liu et al., 2019), which are the main disadvantage of these models.

3.2.3.2. Sparse representation learning. Compressive sensing and sparse representation have gained wide interest in the last decade, particularly in image processing. These techniques stemmed from the observation that natural images tend to be sparse. In STF modeling, the sparse representation learning method uses a dictionary between image patches to learn the differences between fine and coarse images. The design of a dictionary is a critical factor, and the K-singular value decomposition (K-SVD) algorithm is widely used due to its efficiency and simplicity (Chen et al., 2017; Song and Huang, 2013).

The sparse-representation-based spatiotemporal reflectance fusion model (SPSTFM) (Peng et al., 2012) is the first learning-based STF model, which uses a non-linear approach to extract fine-coarse dictionary pairs to reconstruct the fine image at the predicted date. However, it is not reasonable to assume that different image pairs share the same dictionary and that the fine and coarse images share the same sparse coding coefficients for image reconstruction (B. Wu et al., 2015). Several approaches have been proposed in the literature to improve the accuracy of sparse-representation-based STF models. These include reducing the number of input images and utilized high-pass modulation (Song and Huang, 2013), increasing the training data with spatially or temporally extended samples (Ge et al., 2020; Li et al., 2018), introducing error bound regularization technique to account for dictionary element perturbations (B. Wu et al., 2015), proposing a prior detection of temporal change and a two-level selection strategy of similar pixels (Chen et al., 2017), considering the structural knowledge within sparse coefficients (Wei et al., 2017b), using coupled dictionary to constrain the similarity of sparse coefficients (Wei et al., 2017a), considering spectral band structure similarity in the fusion task and using edge information by employing adaptive multi-band constraints (Ying et al., 2018), integrating image super-resolution technique into the fusion process by considering significant spatial resolution difference between fine and coarse images (Chen et al., 2017; Peng et al., 2012; Song and Huang, 2013; L. Wang et al., 2020). Existing sparse representation learning-based STF models have generally neglected the spectral correlations of RS images. However, a recent study by Peng et al. (2022) alleviated the loss of spectral information by integrating the degradation relationship, spatial-spectral-nonlocal correlation, and semi-coupled mapping priors of inputs.

Although several sparse representation-learning STF models have accurately predicted images with land cover changes, they have limitations in accurately maintaining the object shapes (Liu et al., 2016). This is because sparse learning is always computationally complex and cannot extract sufficient local structural information from large input patches (Liu et al., 2016). Furthermore, the sparse representation learning-based STF models are also computationally expensive because they use a nonanalytic optimization in the sparse domain to achieve mapping between fine and coarse images (Jia et al., 2021). These factors limit the application of sparse representation learning models at large-area scales (Jia et al., 2021). Moreover, the fusion accuracy of these models can be low due to insufficient prior knowledge (Wei et al., 2017b), and their reliance on strong assumptions within the dictionary and the sparse coefficient can degrade their performance (Wu et al., 2015). Additionally, although the dictionary training process is critical in sparse representation learning-based STF models, its impact on fusion quality remains unknown (Li et al., 2018).

3.2.3.3. Machine learning and DL. Instead of relying on assumptions, machine learning and DL based-STF model focus on establishing complex relationships between input and output images. Various machine learning-based STF models have been investigated, such as random forests (Hutengs and Vohland, 2016; Ke et al., 2016), regression trees (Boyte et al., 2018), and decision trees (L. Zhang et al., 2021). However, machine learning algorithms may not be effective in high-dimensional RS image fusion. DL-based STF models on the other hand, use customized neural network architecture to effectively establish nonlinear mapping for RS image fusion. DL-based STF models automatically learn and extract abstract features from prior images and use these features to reconstruct fine images. DL-based STF models have also been proposed based on the temporal model of inputs (Ao et al., 2022; Y. Li et al., 2020; Tan et al., 2018), the scale model of inputs (Liu et al., 2019), and the spatial structure consistency between the reference and the prediction dates (Fung et al., 2019). These assumptions provide more information to the DL network to automatically learn the relationship between the reference and ground truth images. Based on the survey of existing STF models, the number of DL-based STF models outnumber machine learning-based STF models by 10-fold. The first DL-based STF model based on an artificial neural network (ANN) was proposed in 2015 (Moosavi et al., 2015), and the subsequent adoption of Convolutional Neural Networks (CNNs) in STF has resulted in a rapid increase in the number of DL-based STF models.

The DL-based STF process mainly includes the following five steps: (1) Preprocessing images; (2) Designing a network for feature fusion; (3) Designing loss function; (4) Training model; (5) Model validation. DL-based STF models may differ in terms of their feature extraction network, architecture design, and loss function design. This review found that except for the first DL-based STF model, which is based on the ANN, existing DL-based STF models commonly use CNNs as the feature extractor due to their ability to automatically learn and extract complex features from images (Yamashita et al., 2018). CNNs achieve invariance in shift, scaling, and distortion by combining three fundamental architectural ideas: local receptive fields, shared weights, and subsampling (Shao and Cai, 2018). In RS image-related applications, CNNs have demonstrated significant advantages, including vegetation monitoring (Kattenborn et al., 2021), classification (W. W. Zhang et al., 2019), and object detection (Ren et al., 2018). This review divides CNNs in STF models into five groups for further analysis: deep convolutional networks, Generative Adversarial Network (GAN), AutoEncoder, Long Short-Term Memory Network (LSTM), and Transformer. Deep convolutional networks are networks that only use stacked convolutional layers without no other architectures. Table 3 summarizes the five groups of CNNs that are used in the existing STF models. Deep convolutional networks are the most commonly used CNNs in the STF, followed by GAN and AutoEncoder. Two studies have investigated LSTM and Transformer, respectively.

3.2.3.3.1. Deep convolutional networks. Convolution in RS images involves image cross-correlation, which enables learning relative spatial location information. By stacking convolutional layers, deep convolutional networks are commonly used in the STF models to extract representational image features. Various STF models based on deep convolutional networks have been proposed using different data inputs and CNN architecture. For example, STFDCNN (Song et al., 2018) and VDCNSTF (Zheng et al., 2019) trained MODIS with downsampled Landsat images and downsampled Landsat with original Landsat images, respectively, using non-linear mapping CNN and super-resolution CNN. ESRCNN (Shao et al., 2019) and DSTFN (Wu et al., 2022) were proposed to predict Sentinel 2 images using Landsat 8 and Sentinel 2 images. The self-adoption fusion of Sentinel 2 was first performed to predict bands 10 and 11 with 10 m resolution using a CNN, followed by a multi-temporal fusion of Landsat 8 and Sentinel 2 using another CNN. Other models, such as StfNet (Liu et al., 2019), DL-SDFM (Jia et al., 2020), DMNet (W. Li et al., 2020), AMNet (Li et al., 2021c), MOST (Wei et al., 2021), MTDL-STF (Jia et al., 2022), STFDSC (Zhang et al., 2022), PDCNN (Li et al., 2022) and STFMCNN (Chen et al., 2022) trained coarse difference images instead of extracting original coarse image features. Two-stageSTF (Sun and Zhang, 2019), Residual-CNN (Wang and Wang, 2020), VDSR (Htitiou et al., 2021), ASRCNN (Ao et al., 2021) and MSFE-SCAM (Lei et al., 2022) fed fine-coarse images directly into the network, while DenseSTF (Ao et al., 2022) predicted image pixels using image patches. STF3DCNN (Peng et al., 2020) and LTSC3D (M. Peng et al., 2022) used three-dimensional CNN. BiaSTF (Li et al., 2020) used different CNN to learn temporal change and sensor bias. HDLSFM (Jia et al., 2021) predicted land cover and phenological changes using Laplacian super-resolution and linear regression, respectively. The final prediction was obtained through a weighted combination of land cover change and phenological change predictions using a sliding window. In addition, MCDNet (Li et al., 2021b) and DPSTFN (Cai et al., 2022) trained fine and coarse images separately using different CNN.

Table 3
A summary of different CNNs used in deep learning-based spatiotemporal fusion models.

CNNs	Name of model
Deep Convolutional Networks	STFDCNN (Song et al., 2018), ESRCNN (Shao et al., 2019), StfNet (Liu et al., 2019), VDCNSTF (Zheng et al., 2019), two-stageSTF (Sun and Zhang, 2019) STF3DCNN (Peng et al., 2020), DL-SDFM (Jia et al., 2020), BiaSTF (Y. Li et al., 2020), Residual-CNN (Wang and Wang, 2020). DMNet (W. Li et al., 2020), VDSR (Htitiou et al., 2021), HDLSFM (Jia et al., 2021), AMNet (Li et al., 2021c), MOST (Wei et al., 2021), MCDNet (Li et al., 2021b), ASRCNN (Ao et al., 2021), MSFE-SCAM (Lei et al., 2022), DenseSTF (Ao et al., 2022), LTSC3D (M. Peng et al., 2022), MTDL-STF (Jia et al., 2022), STFDSC (Zhang et al., 2022), DSTFN (Wu et al., 2022), DPSTFN (Cai et al., 2022), PDCNN (W. Li et al., 2022), STFMCNN (Chen et al., 2022).
GAN	STFGAN (Hongyan Zhang et al., 2021), GAN-STFM (Tan et al., 2021), SSTSTF (Ma et al., 2021), RSFN (Tan et al., 2022), GASTFN (Shang et al., 2022), MLFF-GAN (Song et al., 2022).
AutoEncoder	DCSTFN (Tan et al., 2018), EDCSTFN (Tan et al., 2019), GAN-STFM (Tan et al., 2021), SSTSTF (Ma et al., 2021), MLFF-GAN (Song et al., 2022), STF-EGFA (Cheng et al., 2022).
LSTM	HDLM (Yang et al., 2021),
Transformer	MSNet (Li et al., 2021a).

3.2.3.3.2. GAN and AutoEncoder. GAN (Goodfellow et al., 2014) and AutoEncoder (Hinton and Salakhutdinov, 2006) are generative models using unsupervised learning to learn data distribution effectively. The generated data from generative models is as similar as possible to the actual data distribution, which has high potential in STF. GAN has been applied in RS image processing, including image super-resolution (Jiang et al., 2019), classification (Tao et al., 2017), and object detection (Rabbi et al., 2020). The generator in GAN can generate data with similar statistical distribution as the training samples and tries to fool the discriminator with the generated data, making it difficult to distinguish between generated fake data and real data. However, vanilla GAN suffers from poor stability and mode collapse. Thus, different GANs variants have been proposed to improve the performance of the vanilla GAN. For example, the conditional GAN (CGAN) (Mirza and Osindero, 2014) introduces condition labels to the generator and discriminator to guide the direction of generated data. The least-squares GAN (LSGAN) (Mao et al., 2017) proposes a new loss function to replace the standard min–max loss to enhance the GAN's stability. GAN-based STF models have been proposed using different architectures and input images. STFGAN (Hongyan Zhang et al., 2021) adopted a two-stage framework with an end-to-end image fusion GAN to handle a 16-fold resolution difference between Landsat and MODIS images. GAN-STFM (Tan et al., 2021) only used a coarse image on the prediction date and a randomly selected fine image to feed into the generator network, while SSTSTF (Ma et al., 2021) designed different networks for modeling spatial, temporal, and sensor differences. The inputs of RSFN (Tan et al., 2022) were one coarse image on the prediction date and two fine images before and after the prediction date, with each input image fed to a different network to extract features. GASTFN (Shang et al., 2022) downscaled the coarse difference image through super-resolution and included it in the fine input image. In the feature extraction stage, MLFF-GAN (Song et al., 2022) downsampled fine and coarse images separately and used LSGAN as the loss function.

AutoEncoder consists of three main components: the encoder, bottleneck, and decoder. The encoder compresses the input images into a latent-space representation, while the bottleneck module contains compressed knowledge representations and restricts the flow of information to the decoder, allowing only the most vital information to pass through. The decoder helps the network decompress the knowledge representations and reconstruct the data to its original dimension. AutoEncoder has been applied in RS image processing, such as classification (Li et al., 2016) and change detection (Luppino et al., 2022). When it comes to the STF models with the AutoEncoder architecture, DCSTFN (Tan et al., 2018), EDCSTFN (Tan et al., 2019), and STF-EGFA (Cheng et al., 2022) used encoder and decoder directly to predict fine image. GAN-STFM (Tan et al., 2021), SSTSTF (Ma et al., 2021), and MLFF-GAN (Song et al., 2022), on the other hand, used encoder and decoder in the generator network of GAN to enhance the performance of STF.

3.2.3.3.3. LSTM. LSTM (Hochreiter and Schmidhuber, 1996) performed admirably in extracting time series features. The feature extraction of time series RS images can be viewed as a sequence learning problem, and it can be addressed using a recurrent connection operator (Liu et al., 2017). The applications of LSTM in RS image processing include classification (Liu et al., 2017), super-resolution (Chang and Luo, 2019), crop yield estimation (Tian et al., 2021), and change detection (Sun et al., 2022). An LSTM unit consists of a cell state, input gate, forget gate, and output gate, which can effectively learn temporal patterns over a long time. However, LSTM architectures are designed for processing one-dimensional data, which may pose challenges for high-dimensional RS images STF. For example, HDLM (Yang et al., 2021) used LSTM to learn phenological changing patterns from super-resolution CNN-enhanced time-series images. Note that the input and output of the HDLM are a pixel with six bands, which may limit its ability to extract comprehensive features from the RS image.

3.2.3.3.4. Transformer. The Transformer is a novel architecture proposed by Vaswani et al. (2017), which has shown promising results in time-series change feature extraction and has been widely used in the natural language processing (NLP) and computer vision (CV) fields. Its potential in RS image processing has also been tested. Transformer has been applied in RS for tasks such as image classification (Deng et al., 2022), object detection (Xu et al., 2021), and semantic segmentation (L. L. Wang et al., 2022). In MSNet (Li et al., 2021a), the encoder component of the Transformer was used to extract feature change information to learn the degree of the global time correlation information, making it the first modification of the Transformer for STF. However, since Transformer is completely based on self-attention, it often ignores the pixel-level internal structural features of small blocks, leading to a loss of shallow features (Lin et al., 2022). Additionally, the high resolution of the RS images results in a quadratic increase in the input image size, making computational complexity (Su et al., 2022). Fortunately, several recent studies have improved Transformer architecture in CV. For example, CNN-enhanced Transformer, Local Attention Enhanced Transformer, and Hierarchical Transformer (Liu et al., 2021). These improvements in Transformer are promising to promote the development of STF in the future.

Fig. 8 shows the distribution of STF models based on different CNNs by year for further analysis. Deep convolutional networks and AutoEncoder were first used in STF models in 2018, with the number of deep convolutional networks-based STF models increasing yearly since then. The number of AutoEncoder and GAN-based STF models is equal, with GAN-based STF models having been proposed since 2021. Both LSTM and Transformer networks were applied in the STF models only in 2021. The effectiveness of convolution in DL and RS image processing has contributed to an overwhelming number of deep convolutional networks-based STF models. However, the potential of GAN, AutoEncoder, LSTM, and Transformer networks in the STF field needs further exploration.

Apart from the DL networks discussed above, incorporating various strategies to construct DL network architecture also significantly impacts the STF result. Therefore, Table 4 summarizes the most frequently used DL strategies in the STF models. These strategies include residual learning, attention mechanism, super-resolution, multi-stream, compound loss function, multiscale mechanism, and transfer learning.

3.2.3.3.5. Residual learning. A study by He et al. (2016) introduced a deep residual learning framework, namely ResNet, to harness the degradation problem in a deep neural network, making it possible to build extremely deep networks. It is believed that the DL network extracted features will be more representative as the network become deeper. However, as the depth of the DL network increases, the gradient tends to vanish or explode, leading to convergence difficulties (Li et al., 2020). The concept of skip connection, which reuses the information from previous layers, is ResNet's most significant contribution, as it avoids gradient vanishing or

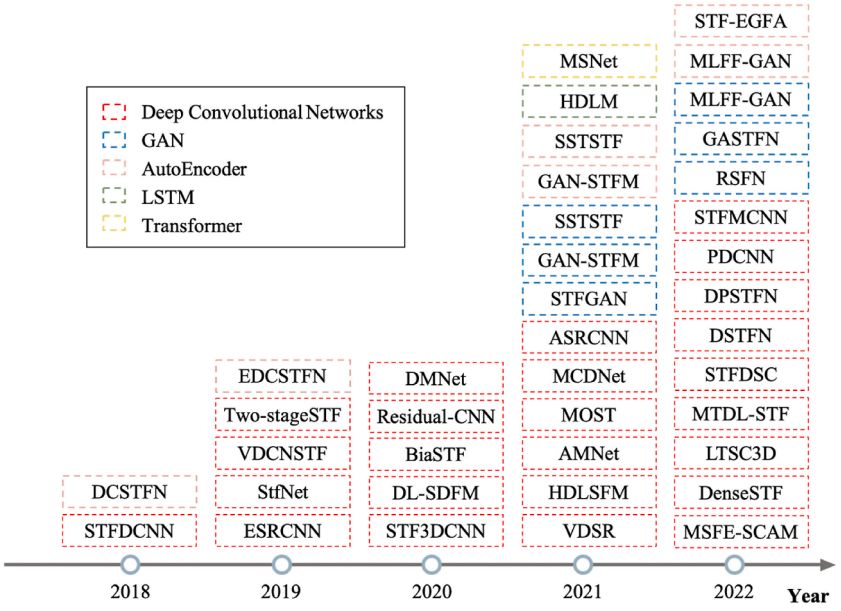

Fig. 8. Development of different convolutional neural networks in STF models.

Table 4
Summary of the deep learning strategies used in the existing DL-based STF.

Deep learning strategy	Name of models
Residual learning	STFDCNN (Song et al., 2018), two-stageSTF (Sun and Zhang, 2019), Residual-CNN (Wang and Wang, 2020), HDLSFM (Jia et al., 2021), STFGAN (Zhang et al., 2021), VDSR (Htitiou et al., 2021), GAN-STFM (Tan et al., 2021), SSTSTF (Ma et al., 2021), MOST (Wei et al., 2021), MSFE-SCAM (Lei et al., 2022), RSFN (Tan et al., 2022), MTDL-STF (Jia et al., 2022), DSTFN (Wu et al., 2022), PDCNN (W. Li et al., 2022), STFMCNN (Chen et al., 2022), MLFF-GAN (Song et al., 2022).
Attention mechanism	SSTSTF (Ma et al., 2021), AMNet (Li et al., 2021c), MSNet (Li et al., 2021a), ASRCNN (Ao et al., 2021), MSFE-SCAM (Lei et al., 2022), RSFN (Tan et al., 2022), DenseSTF (Ao et al., 2022), MTDL-STF (Jia et al., 2022), DSTFN (Wu et al., 2022), DPSTFN (Cai et al., 2022), PDCNN (W. Li et al., 2022), MLFF-GAN (Song et al., 2022), STF-EGFA (Cheng et al., 2022).
Super-resolution	STFDCNN (Song et al., 2018), StfNet (Liu et al., 2019), ESRCNN (Shao et al., 2019), VDCNSTF (Zheng et al., 2019), VDSR (Htitiou et al., 2021), HDLSFM (Jia et al., 2021), MOST (Wei et al., 2021), MCDNet (Li et al., 2021b), HDLM (Yang et al., 2021), DPSTFN (Cai et al., 2022), GASTFN (Shang et al., 2022).
Multi-stream	DCSTFN (Tan et al., 2018), StfNet (Liu et al., 2019), DL-SDFM (Jia et al., 2020), MSNet (Li et al., 2021a), STFDSC (Zhang et al., 2022), PDCNN (W. Li et al., 2022), STFMCNN (Chen et al., 2022), STF-EGFA (Cheng et al., 2022).
Compound loss function	EDCSFTN (Tan et al., 2019), MCDNet (Li et al., 2021b), GAN-STFM (Tan et al., 2021), SSTSTF (Ma et al., 2021), RSFN (Tan et al., 2022), MSFE-SCAM (Lei et al., 2022), STF-EGFA (Cheng et al., 2022).
Multiscale mechanism	DMNet (W. Li et al., 2020), AMNet (Li et al., 2021c) MCDNet (Li et al., 2021b), MSFE-SCAM (Lei et al., 2022), PDCNN (W. Li et al., 2022).
Transfer learning	MOST (Wei et al., 2021).

exploding problems. To improve STF results, numerous DL-based STF models use residual learning to avoid image information loss and reduce computational complexity (Jia et al., 2022; Li et al., 2022). Residual learning architecture is also used as a feature extractor to extract RS image features for STF (Hongyan Zhang et al., 2021).

3.2.3.3.6. Attention mechanism. In recent years, attention mechanisms have gained significant attention in the DL. These mechanisms are inspired by the human visual system that when humans observe an object, they generally do not look at it in its entirety but instead selectively access certain important parts of the observed object based on their needs. Attention mechanisms in DL focus on the most relevant part of the input for a given task. Different variants of attention mechanisms, such as global attention, local attention (Luong et al., 2015), self-attention (Vaswani et al., 2017), spatial attention, and channel attention, have been proposed and widely used in NLP and CV with great success. In the field of STF, spatial and channel attention mechanisms have been applied to better capture and preserve spatial and spectral information to improve fusion accuracy (Cai et al., 2022; Cheng et al., 2022; Lei et al., 2022; Ma et al., 2021). Additionally, an attention mechanism module has been used to focus on the temporal changes in the RS image (Song et al., 2022). Furthermore, attention mechanisms have been used to focus on feature maps in AMNet to improve fusion accuracy (Li et al., 2021c). However, further exploration is needed to better leverage the attention mechanisms based on different DL architectures for STF.

3.2.3.3.7. Super-resolution. Super-resolution is commonly used in DL-based STF models to bridge the large resolution scale gap between fine and coarse images. The super-resolution convolutional neural network (SRCNN) proposed by Dong et al. (2016) was

widely used for image enhancement. Some DL-based STF models divided the fusion process into two stages: image super-resolution and image fusion. However, this approach faces challenges because super-resolution is an ill-posed inverse problem, and changes in reflectance over time increase uncertainty (Shang et al., 2022). STFDCNN (Song et al., 2018) and VDCNSTF (Zheng et al., 2019) downsampled Landsat image by ten times to generate a low spatial resolution Landsat image to establish super-resolution mapping model between downsampled Landsat and original Landsat. Two-stageSTF (Sun and Zhang, 2019) preprocessed all input images to the same intermediate resolution images to predict preliminary fusion results. However, a limitation of these approaches is that the super-resolution and fusion networks are trained in different stages. The learning results of the super-resolution in the first stage tend to cause bottlenecks in the fusion network in the second stage, and these bottlenecks cannot be repaired in the training stage of the fusion network. Training STF models with end-to-end networks, on the other hand, may perform better. In addition, to fully exploit the pixel details of the fine image, the coarse images should be upsampled to the same resolution as the fine image and fed into the STF model simultaneously. Therefore, some super-resolution strategies may be redundant and limit the overall performance of STF models.

3.2.3.3.8. Multi-stream. A multi-stream strategy used in the STF model is to design two or more stream neural networks to extract features of different input images. This approach aims to improve the fusion accuracy using different networks in separate streams to focus on different features from the input images. For example, Tan et al. (2018) designed a two-stream neural network in DCSTFN to extract features of fine and coarse images separately. Similarly, StfNet (Liu et al., 2019), MSNet (Li et al., 2021a), STFDSC (Zhang et al., 2022), PDCNN (W. Li et al., 2022), and STFMCNN (Chen et al., 2022) trained two-stream networks to predict two fine difference images before reconstructing the target fine image. DL-SDFM (Jia et al., 2020) learned temporal change mapping with one neural network and spatial difference mapping with another. In contrast, STF-EGFA (Cheng et al., 2022) fed two pairs of fine-coarse images into one neural network and two fine images into the other two separate networks. However, multi-stream architecture increases the number of parameters and computations, as well as limits the network's ability to directly learn the complex mapping between all the inputs and output.

3.2.3.3.9. Compound loss function. DL networks use a loss function to calculate the errors between the predicted value and the actual value, which is crucial in evaluating their performance. The lower the value of the loss function, the closer the predicted result of the DL network is to the ground truth. It has been reported that a simple L1 or L2 loss function is difficult to optimize the parameters well and obtain high-quality fusion images (Lei et al., 2022). When only the L2 loss function is used, the fusion image is relatively smooth, but edge information is lost because it penalizes larger errors. L1 loss, on the other hand, is detrimental to the convergence and the learning of the STF model because the gradients of L1 are equal most time. However, the L1 loss function has demonstrated greater robustness and better performance compared to the L2 loss functions for the image restoration models (Zhao et al., 2015). Therefore, A compound loss was used to direct the DL-based STF model to enhance the clarity and sharpness of predictions (Lei et al., 2022; Li et al., 2021b; Tan et al., 2019, 2022). Additionally, some DL-based STF models have also used perceptual loss, which measures feature loss using a pre-trained network in the compound loss function (Cheng et al., 2022; Ma et al., 2021; Tan et al., 2019, 2021).

3.2.3.3.10. Multiscale mechanism. In DL neural network, the receptive field refers to the region of the input that produces the features. Typically, a large receptive field results in a low resolution of the feature map, and vice versa, and a shallow neural network has a small receptive field but a high-resolution feature map. To achieve a large receptive field, one approach is to modify the stride of the layer or use pooling operations for downsampling, and multiple downsampling processes can further increase the receptive field. Nevertheless, Downsampling inevitably reduces the resolution of the feature map. Therefore, the fusion of receptive fields or feature maps at multiple scales strives to exploit the benefits of each layer in the network to increase STF accuracy. Given RS images contain contextual information at different scales, using receptive fields or feature maps of various scales is crucial to avoid losing essential image features during convolutional layer operations. DL-based STF models such as DMNet (Li et al., 2020), MCDNet (Li et al., 2021b), and PDCNN (Li et al., 2022) used a 3×3, a 5×5 (or two 3×3 convolutional layers), and a 7×7 convolutional layer (or three 3×3 convolutional layers) to perceive multiscale information from feature maps more flexibly. The multiscale fusion strategy used in AMNet (Li et al., 2021c) and MSFE-SCAM (Lei et al., 2022) is similar to the idea of the Feature Pyramid Network, which fuses images of different resolutions.

3.2.3.3.11. Transfer learning. Transfer learning is the process of transferring the weights of the pre-trained model to a new DL model to aid in its training. This approach is valuable because training a DL model with random initial weights can be a time-consuming process. By applying transfer learning, the training process can be accelerated, and it can help overcome the gradient vanishing problem. In addition, to address the case of a small training dataset, a pre-trained model with features learned from other large datasets can be transferred to new models through transfer learning. MOST (Wei et al., 2021) used transfer learning to address the issue of insufficient training data for STF.

Fig. 9 shows the use of different DL strategies in the STF model over the year, as summarized in Table 4. In 2018, STF models began to use residual learning, super-resolution, and multi-stream strategies. Residual learning was the most frequently used strategy in STF models, followed by the attention mechanism and super-resolution strategies. Both residual learning and multi-stream strategies have continued to show upward trends, while the number of models employing super-resolution strategy increased but subsequently experienced a rapid decline in 2022. Compound loss functions were first introduced in STF models in 2019. Multiscale mechanism strategies were introduced in 2020 and increased over the following two years. Attention mechanism and transfer learning were applied in STF models beginning in 2021. Attention mechanism has received significant attention in STF models. In 2022, the number of STF models that used attention mechanism exceeded those using residual learning. Note that the efficacy of these strategies within a DL-based STF model needs to be verified because many factors influence the accuracy of a DL model.

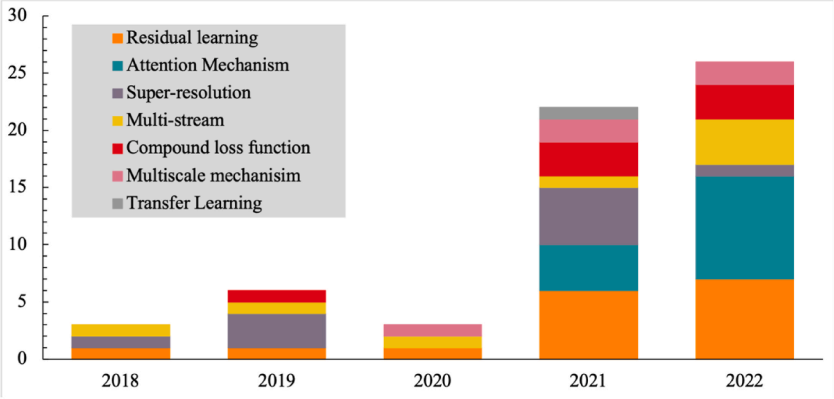

Fig. 9. Development of deep learning strategies.

Some of the proposed DL-based STF models (Shang et al., 2022; Tan et al., 2019, 2021) require a minimum number of input images, facilitating the STF task in areas with severe cloud contamination. Especially, GAN-STFM (Tan et al., 2021) only used a fine image at arbitrary dates before the prediction date and a coarse image at the prediction date, providing more flexibility for STF. DL-based STF models have employed simple architectures with few trainable parameters to shorten model training time (Li et al., 2020; Li et al., 2020; Peng et al., 2020; Tan et al., 2019). Furthermore, some DL-based STF models accept multiband input, allowing the model to exploit the spectral correlation between different bands to better determine land cover and phenological changes (Jia et al., 2020; Peng et al., 2020; Shang et al., 2022; Tan et al., 2019, 2021). In summary, DL-based STF models have progressed from complex, indirect modeling to simpler, lighter modeling (Zhang et al., 2022), with the aim of striking a balance between spatial detail preservation and spectral change reconstruction (Chen et al., 2022). Additionally, DL-based STF models have tried to enhance fusion performance by using various DL strategies. However, existing DL-based STF models only consider the statistical relationship between the fine and coarse images rather than the physical properties of RS signals, such that the predicted fine images cannot retain precise spatial details and object shapes (Chen et al., 2017; Shi et al., 2019). Further research is needed to determine how to incorporate physical temporal changes into DL-based fusion to preserve spectral fidelity and spatial details in predictions (Chen et al., 2017). Moreover, some STF models have been designed to take sensor differences into account (Li et al., 2020; Ma et al., 2021). Although the training speed of DL-based STF models can be accelerated by utilizing powerful graphics processing units and clipping the image to a smaller patch to feed into the model (Li et al., 2021a), further research is required to improve fusion accuracy while reducing computational complexity. DL-based STF models have been used for numerous applications, including NDVI (Ao et al., 2021; Htitiou et al., 2021) and LST modeling (Yin et al., 2021), color correction for mosaicked images (Wei et al., 2021), classification (Yang et al., 2022), and sea surface temperature monitoring (Zha et al., 2022).

3.2.4. Hybrid method

To capitalize on the advantages of the three STF categories introduced above, a hybrid STF method that combines multiple principles has been developed to produce more accurate fusion results (Xue et al., 2019). Hybrid STF models include those combining unmixing-based and weight function-based methods (Cui et al., 2018; Jiang and Huang, 2022; Liu et al., 2019; Ping et al., 2018; S. Wang et al., 2022; Xie et al., 2016; Zhu et al., 2016), unmixing-based and Bayesian learning methods (X. Li et al., 2017; Liao et al., 2016; Xue et al., 2019), or all three STF methods (Gevaert and García-Haro, 2015; Ma et al., 2018; Zhao et al., 2018).

One of the first proposed hybrid models, Flexible Spatiotemporal DAta Fusion (FSDAF), combines unmixing-based, weighted function-based, and spatial interpolation methods (Zhu et al., 2016). Several improved FSDAF-based models have been developed to better capture complex temporal changes and improve fusion accuracy. These modifications included introducing an enhanced linear regression model (Tang et al., 2019), introducing a constrained least-squares process to combine the increment from unmixing and coarse image interpolation (Liu et al., 2019), considering mixed pixels in fine images and introducing a new residual index (Shi et al., 2019), considering sub-pixel class fraction change information (Li et al., 2020), selecting spectral unmixing adaptively (Hou et al., 2020), adopting linear pixel unmixing to generate high-resolution LAI (Zhai et al., 2020), incorporating change detection technology for land cover changed areas (Guo et al., 2020), using an accelerated inverse-distance weight, parallelized computation, and an adaptive domain decomposition method (Gao et al., 2022). The applications of hybrid-based models include NDVI (Liao et al., 2016), LST (Quan et al., 2018), and LAI monitoring (Zhai et al., 2020). Although there is no clear pattern in developing hybrid STF models, existing models are mainly combinations of unmixing-based and weight function-based models. As a result, hybrid STF models offer improved fusion accuracy but are computationally costly.

4. Discussion

A systematic review of proposed STF models is essential for advancing the field of STF. In this review, up to 146 STF models have been proposed within the collected literature. These models are classified into four categories: unmixing-based, weight function-based, learning-based, and hybrid methods. The review specifically examines the concept, development, advantages, disadvantages,

challenges, and applications of each class. Specifically, it summarizes the superior performance and sharply increased number of the DL-based STF models, including their network architecture and used DL strategies. Modeling complex spatiotemporal interdependencies between images is a major challenge of STF, and it is especially difficult to achieve high performance when applying these methods in heterogeneous landscapes and for land cover retrieval under rapid changes (Htitiou et al., 2021).

4.1. Comparison of STF methods

Unmixing-based STF models assume the spectral mixing model to be linear, making it difficult to express complex land surface changes and may result in the loss of spatial and spectral details. Weight function-based STF models, on the other hand, make assumptions about constant temporal changes between fine and coarse images or spatial distances between fine and corresponding coarse images, limiting their ability to capture the temporal and spatial variability of images adequately (Zhu et al., 2018). However, weight function-based STF models retain spatial details better than unmixing models (Ma et al., 2018). Among the different types of learning-based STF methods, DL-based STF models particularly have obtained superior fusion results, making them the preferred method of STF models. This is attributed to the DL-based STF models' ability to automatically extract abstract features from training data and use these features to effectively fuse RS images (Lei et al., 2022; Tan et al., 2018). However, learning-based STF models focus on image predictions based on spatial similarity and often ignore spectral correlations among RS images, resulting in a significant loss of spectral information (Y. Peng et al., 2022). Hybrid STF models combine unmixing-, weight function-, and learning-based methods to retain their complementary strengths and outperform all three individual methods. However, the computational complexity of hybrid models limits their development.

4.2. STF datasets

Landsat and MODIS images are widely used in STF due to their similar bandwidth and radiation (Jia et al., 2021). However, fusion results with daily 30 m spatial resolution are inadequate for precision Earth observation applications. On the other hand, fusion results with 10 m spatial resolution for Sentinel and Landsat data are increasingly preferred for their higher spatial resolution. Due to the high demands on UAV images with centimeter-scale resolution in practical precision Earth observation applications, STF of UAV and satellite images exhibits great potential in future applications. In addition to the STF model, the fusion result is also influenced by the RS image dataset. Previous studies have discovered that selecting different base image pairs for a given prediction date can significantly impact STF performance (Qiu et al., 2021; Xie et al., 2018). Moreover, if reflectance changes in the test dataset are not learned in the training dataset, a DL-based STF model will not learn such cases, resulting in lower fusion accuracy. The size of the input images for STF models also differs between DL-based and other methods. A region of interest (ROI) is subdivided into smaller patches for DL-based STF models to account for computer memory consumption, whereas other models can directly use ROI images as input.

4.3. Challenges

The large spatial ratio between fine and coarse RS images, as well as the multidimensionality of RS images, make STF more difficult than natural image fusion (Liu et al., 2019). Specifically, the primary difficulties for RS image STF are spatial differences, sensor differences, and temporal changes (Song et al., 2022). In addition, one of the challenges in the STF field is the practical application of proposed STF models. More than 100 STF models have been proposed to date, but only a few of these have been implemented in real-world applications. This failure may be partly due to the lack of datasets, model codes, and a user-friendly graphical user interface. Many STF studies have claimed to have achieved state-of-the-art results using datasets from specific regions. However, to comprehensively assess the performance of different models, benchmark datasets that cover the heterogeneity of the Earth's surface and incorporate images from various sensors are needed, as current public benchmark datasets are insufficient. Additionally, the lack of open-source code and limited reproducibility also pose challenges in STF, impeding objective evaluation and hindering progress in this field. Moreover, there is a need to enhance the generalization, robustness, and transferability of STF models. Current research on DL-based STF often inputs small patches of RS images obtained by clipping, which greatly compresses the data utilization and presents a significant barrier for industrial applications. Furthermore, accurately capturing land cover changes and landscape heterogeneity remain unsolved challenges in STF, as discussed by several studies (Htitiou et al., 2021). Therefore, addressing these challenges is crucial to improve the performance and applicability of STF models.

4.4. Frameworks to improve STF results

Several studies have developed frameworks to improve the robustness of STF models. One approach is using virtual image pair, which is an image pair closer to the prediction date than the original reference image pairs. This approach was developed to reduce the differences between images on the reference and prediction dates (Q. Wang et al., 2020). This flexible framework can accommodate any number of image pairs and is suitable for unmixing and weight function-based STF models. The blocks-removed spatial unmixing and geographically weighted spatial unmixing frameworks were developed by Wang et al. (2021b) to improve the fusion accuracy of unmixing-based models by removing the blocks effect on the fusion image and addressing spatial variation in land cover, respectively. In situations where there are insufficient data, a data augmentation method using cycle GAN architecture has been proposed to generate images for STF (J. Chen et al., 2021). Another approach is the use of a similarity weight learning block to replace manually designed weight calculation rules that are common in weight function-based models or integrated into the DL model to better utilize neighboring fine images (Sun and Xiao, 2022).

4.5. STF outlooks

4.5.1. Benchmark datasets for STF

The widely used benchmark datasets in the STF field are the CIA and LGC datasets. Nevertheless, these datasets only include Landsat and MODIS image pairs and cover a limited landscape. Existing proposed STF models were used to predict Sentinel 2 or other higher-resolution images in addition to Landsat images. However, no corresponding benchmark datasets are available to assess their performance. As RS technology advances, a variety of different RS images may be predicted using STF techniques in the future, further expanding the need for additional benchmarks. Additionally, to facilitate the application of STF models across various landscapes (e.g., homogenous and heterogeneous) with different temporal changes (e.g., phenological and land cover changes), additional benchmark datasets acquired by various sensors and from diverse landscapes for the STF model are needed. By creating such datasets, future STF models can systematically and robustly address various data fusion tasks under complex environmental conditions.

4.5.2. Practical applications of STF images

Despite the numerous STF models proposed to date, only a few have been implemented in practice Earth observation applications. In addition to refining these STF models theoretically, expanding their practical application is also vital to advance STF technology. To facilitate the application and comparison of newer models, open-source STF model codes should be made available to the public to remove research barriers (Sdraka et al., 2022). In addition, STF models are written in different programming languages (e.g., C++, Python) and adopted on different platforms (e.g., MATLAB, GEE, IDE) for the fusion process, necessitating the development of a graphical user interface (GUI) to facilitate the application of high-performance STF models. Another important fusion topic in the RS community is pansharpening, which involves the fusion of panchromatic and multispectral bands. The high demand and impressive results of pansharpened image products have led to extensive use in Google Earth and Bing Maps (Alparone et al., 2016). Notably, the classical Gram-Schmidt pansharpening method has been involved in commercial image-processing software, such as ENVI, ERDAS Imagine, and PCI Geomatica. Therefore, implementing high-performance STF models into commercial software is another way to expand STF applications.

4.5.3. STF based on centimeter-scale images and advanced theory

The rapid evolution of RS technology, including UAVs, and advanced theories such as DL have highlighted the necessity for STF models to explore the fusion of centimeter-scale of UAV and satellite images with >100-fold image resolution differences. Although several state-of-the-art STF models have been proposed and applied to various satellite datasets, the current STF models can only handle a spatial resolution scale difference of up to 16-fold between fine and coarse images. Thus, extending the accuracy and efficiency of STF models to handle more precise image fusion is a promising future direction for STF. The computational load for massive very-high-resolution RS data is very heavy, necessitating a balance between efficiency and accuracy. Therefore, understanding how to efficiently model the relationship between input and output with high accuracy using different strategies is critical for advancing STF.

5. Conclusion

This study provides a comprehensive analysis of proposed STF models, revealing that most STF models have adopted DL techniques since 2018, which have shown superior performance compared to other methods. However, despite the considerable attention received by STF in recent decades, practical Earth observation applications of these models have remained limited. Therefore, this review discussed future directions for STF, including developing benchmark datasets, promoting practical STF model applications, and incorporating centimeter-scale UAV images with advanced DL techniques. Examining various STF methods and recent advances in DL-based STF models has the potential to inspirefurther development of high-performance STF models for Earth observation applications.

Ethical Statement

Hereby, I/Ram Avtar/consciously assure that for the manuscript/**A review of remote sensing image spatiotemporal fusion: Challenges, applications and recent trends**/the following is fulfilled.

1) This material is the authors' own original work, which has not been previously published elsewhere.
2) The paper is not currently being considered for publication elsewhere.
3) The paper reflects the authors' own research and analysis in a truthful and complete manner.
4) The paper properly credits the meaningful contributions of co-authors and co-researchers.
5) The results are appropriately placed in the context of prior and existing research.
6) All sources used are properly disclosed (correct citation). Literally copying of text must be indicated as such by using quotation marks and giving proper reference.
7) All authors have been personally and actively involved in substantial work leading to the paper, and will take public responsibility for its content.

The violation of the Ethical Statement rules may result in severe consequences.

To verify originality, your article may be checked by the originality detection software iThenticate. See also http://www.elsevier.com/editors/plagdetect.

I agree with the above statements and declare that this submission follows the policies of Solid State Ionics as outlined in the Guide for Authors and in the Ethical Statement.

Declaration of competing interest

The authors declare that they have no conflict of interests with the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgment:

We would like to thank the Tenure-track grant of Hokkaido University, Hirose grant, and APN-GCR grant no. CBA2021-02MY-Avtar, and JST SPRING, Grant Number JPMJSP2119 for financial support.

References

Alparone, L., Baronti, S., Aiazzi, B., Garzelli, A., 2016. Spatial methods for multispectral pansharpening: multiresolution analysis demystified. IEEE Trans. Geosci. Rem. Sens. 54, 2563–2576. https://doi.org/10.1109/TGRS.2015.2503045.

Amorós-López, J., Gómez-Chova, L., Alonso, L., Guanter, L., Moreno, J., Camps-Valls, G., 2011. Regularized multiresolution spatial unmixing for ENVISAT/MERIS and landsat/TM image fusion. Geosci. Rem. Sens. Lett. IEEE 8, 844–848. https://doi.org/10.1109/LGRS.2011.2120591.

Amorós-López, J., Gómez-Chova, L., Alonso, L., Guanter, L., Zurita-Milla, R., Moreno, J., Camps-Valls, G., 2013. Multitemporal fusion of Landsat/TM and ENVISAT/MERIS for crop monitoring. Int. J. Appl. Earth Obs. Geoinf. 23, 132–141. https://doi.org/10.1016/j.jag.2012.12.004.

Anderson, M.C., Kustas, W.P., Norman, J.M., Hain, C.R., Mecikalski, J.R., Schultz, L., González-Dugo, M.P., Cammalleri, C., D'Urso, G., Pimstein, A., Gao, F., 2011. Mapping daily evapotranspiration at field to continental scales using geostationary and polar orbiting satellite imagery. Hydrol. Earth Syst. Sci. 15, 223–239. https://doi.org/10.5194/hess-15-223-2011.

Ao, Z., Sun, Y., Pan, X., Xin, Q., 2022. Deep learning-based spatiotemporal data fusion using a patch-to-pixel mapping strategy and model comparisons. IEEE Trans. Geosci. Rem. Sens. 60, 1–18. https://doi.org/10.1109/TGRS.2022.3154406.

Ao, Z., Sun, Y., Xin, Q., 2021. Constructing 10-m NDVI time series from landsat 8 and sentinel 2 images using convolutional neural networks. Geosci. Rem. Sens. Lett. IEEE 18, 1461–1465. https://doi.org/10.1109/LGRS.2020.3003322.

Belgiu, M., Stein, A., 2019. Spatiotemporal image fusion in remote sensing. Rem. Sens. 11. https://doi.org/10.3390/rs11070818.

Bhattarai, N., Quackenbush, L.J., Dougherty, M., Marzen, L.J., 2015. A simple Landsat–MODIS fusion approach for monitoring seasonal evapotranspiration at 30 m spatial resolution. Int. J. Rem. Sens. 36, 115–143. https://doi.org/10.1080/01431161.2014.990645.

Boyte, S.P., Wylie, B.K., Rigge, M.B., Dahal, D., 2018. Fusing MODIS with Landsat 8 data to downscale weekly normalized difference vegetation index estimates for central Great Basin rangelands, USA. GIScience Remote Sens. 55, 376–399. https://doi.org/10.1080/15481603.2017.1382065.

Cai, J., Huang, B., Fung, T., 2022. Progressive spatiotemporal image fusion with deep neural networks. Int. J. Appl. Earth Obs. Geoinf. 108, 102745. https://doi.org/10.1016/j.jag.2022.102745.

Chang, Y., Luo, B., 2019. Bidirectional convolutional LSTM neural network for remote sensing image super-resolution. Rem. Sens. 11. https://doi.org/10.3390/rs11202333.

Chen, B., Chen, L., Huang, B., Michishita, R., Xu, B., 2018. Dynamic monitoring of the Poyang Lake wetland by integrating Landsat and MODIS observations. ISPRS J. Photogrammetry Remote Sens. 139, 75–87. https://doi.org/10.1016/j.isprsjprs.2018.02.021.

Chen, B., Huang, B., Xu, B., 2017. A hierarchical spatiotemporal adaptive fusion model using one image pair. Int. J. Digit. Earth 10, 639–655. https://doi.org/10.1080/17538947.2016.1235621.

Chen, B., Huang, B., Xu, B., 2015. Comparison of spatiotemporal fusion models: a review. Rem. Sens. 7, 1798–1835. https://doi.org/10.3390/rs70201798.

Chen, J., Wang, L., Feng, R., Liu, P., Han, W., Chen, X., 2021. CycleGAN-STF: spatiotemporal fusion via CycleGAN-based image generation. IEEE Trans. Geosci. Rem. Sens. 59, 5851–5865. https://doi.org/10.1109/TGRS.2020.3023432.

Chen, R., Li, X., Zhang, Y., Zhou, P., Wang, Y., Shi, L., Jiang, L., Ling, F., Du, Y., 2021. Spatiotemporal continuous impervious surface mapping by fusion of landsat time series data and google earth imagery. Rem. Sens. 13. https://doi.org/10.3390/rs13122409.

Chen, Y., Shi, K., Ge, Y., Zhou, Y., 2022. Spatiotemporal remote sensing image fusion using multiscale two-stream convolutional neural networks. IEEE Trans. Geosci. Rem. Sens. 60. https://doi.org/10.1109/TGRS.2021.3069116.

Cheng, F., Fu, Z., Tang, B., Huang, L., Huang, K., Ji, X., 2022. STF-EGFA: a remote sensing spatiotemporal fusion network with edge-guided feature attention. Rem. Sens. 14. https://doi.org/10.3390/rs14133057.

Cheng, Q., Liu, H., Shen, H., Wu, P., Zhang, L., 2017. A spatial and temporal nonlocal filter-based data fusion method. IEEE Trans. Geosci. Rem. Sens. 55, 4476–4488. https://doi.org/10.1109/TGRS.2017.2692802.

Cui, J., Zhang, X., Luo, M., 2018. Combining linear pixel unmixing and STARFM for spatiotemporal fusion of Gaofen-1wide field of view imagery and MODIS imagery. Rem. Sens. 10, 1–22. https://doi.org/10.3390/rs10071047.

Deng, P., Xu, K., Huang, H., 2022. When CNNs meet vision transformer: a joint framework for remote sensing scene classification. Geosci. Rem. Sens. Lett. IEEE 19, 1–5. https://doi.org/10.1109/LGRS.2021.3109061.

Dogra, A., Goyal, B., Agrawal, S., 2017. From multi-scale decomposition to non-multi-scale decomposition methods: a comprehensive survey of image fusion techniques and its applications. IEEE Access 5, 16040–16067. https://doi.org/10.1109/ACCESS.2017.2735865.

Dong, C., Loy, C.C., He, K., Tang, X., 2016. Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. 38, 295–307. https://doi.org/10.1109/TPAMI.2015.2439281.

Emelyanova, I.V., McVicar, T.R., Van Niel, T.G., Li, L.T., van Dijk, A.I.J.M., 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 Sens. Environ. 133, 193–209. https://doi.org/10.1016/j.rse.2013.02.007.

Fu, D., Chen, B., Wang, J., Zhu, X., Hilker, T., 2013. An improved image fusion approach based on enhanced spatial and temporal the adaptive reflectance fusion model. Rem. Sens. 5, 6346–6360. https://doi.org/10.3390/rs5126346.

Fung, C.H., Wong, M.S., Chan, P.W., 2019. Spatio-temporal data fusion for satellite images using hopfield neural network. Rem. Sens. 11. https://doi.org/10.3390/rs11182077.

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. Rem. Sens. 44, 2207–2218. https://doi.org/10.1109/TGRS.2006.872081.

Gao, H., Zhu, X., Guan, Q., Yang, X., Yao, Y., Zeng, W., Peng, X., 2022. CuFSDAF: an enhanced flexible spatiotemporal data fusion algorithm parallelized using graphics processing units. IEEE Trans. Geosci. Rem. Sens. 60. https://doi.org/10.1109/TGRS.2021.3080384.

Ge, Y., Li, Y., Chen, J., Sun, K., Li, D., Han, Q., 2020. A learning-enhanced two-pair spatiotemporal reflectance fusion model for gf-2 and gf-1 wfv satellite data. Sensors 20. https://doi.org/10.3390/s20061789.

Gevaert, C.M., García-Haro, F.J., 2015. A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion. Remote Sens. Environ. 156, 34–44. https://doi.org/10.1016/j.rse.2014.09.012.

Ghamisi, P., Rasti, B., Yokoya, N., Wang, Q., Hofle, B., Bruzzone, L., Bovolo, F., Chi, M., Anders, K., Gloaguen, R., Atkinson, P.M., Benediktsson, J.A., 2019. Multisource and multitemporal data fusion in remote sensing: a comprehensive review of the state of the art. IEEE Geosci. Remote Sens. Mag. 7, 6–39. https://doi.org/10.1109/MGRS.2018.2890023.

Ghassemian, H., 2016. A review of remote sensing image fusion methods. Inf. Fusion 32, 75–89. https://doi.org/10.1016/j.inffus.2016.03.003.
Ghosh, R., Gupta, P.K., Tolpekin, V., Srivastav, S.K., 2020. An enhanced spatiotemporal fusion method – implications for coal fire monitoring using satellite imagery. Int. J. Appl. Earth Obs. Geoinf. 88, 102056. https://doi.org/10.1016/j.jag.2020.102056.
Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in Neural Information Processing Systems. IEEE. https://doi.org/10.1109/ICCVW.2019.00369.
Guo, D., Shi, W., Hao, M., Zhu, X., 2020. Fsdaf 2.0: improving the performance of retrieving land cover changes and preserving spatial details. Remote Sens. Environ. 248, 111973. https://doi.org/10.1016/j.rse.2020.111973.
Hazaymeh, K., Hassan, Q.K., 2015. Spatiotemporal image-fusion model for enhancing the temporal resolution of Landsat-8 surface reflectance images using MODIS images. J. Appl. Remote Sens. 9, 096095. https://doi.org/10.1117/1.jrs.9.096095.
He, C., Zhang, Z., Xiong, D., Du, J., Liao, M., 2017. Spatio-temporal series remote sensing image prediction based on multi-dictionary Bayesian fusion. ISPRS Int. J. Geo-Inf. 6. https://doi.org/10.3390/ijgi6110374.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proc. IEEE Conf. Comput. Vis. pattern Recognit 770–778. https://doi.org/10.1002/chin.200650130.
Hilker, T., Wulder, M.A., Coops, N.C., Linke, J., McDermid, G., Masek, J.G., Gao, F., White, J.C., 2009a. A new data fusion model for high spatial- and temporal-resolution mapping of forest disturbance based on Landsat and MODIS. Remote Sens. Environ. 113, 1613–1627. https://doi.org/10.1016/j.rse.2009.03.007.
Hilker, T., Wulder, M.A., Coops, N.C., Seitz, N., White, J.C., Gao, F., Masek, J.G., Stenhouse, G., 2009b. Generation of dense time series synthetic Landsat data through data blending with MODIS using a spatial and temporal adaptive reflectance fusion model. Remote Sens. Environ. 113, 1988–1999. https://doi.org/10.1016/j.rse.2009.05.011.
Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313, 504–507.
Hochreiter, S., Schmidhuber, J., 1996. LSTM can solve hard long time lag problems. Adv. Neural Inf. Process. Syst. 473–479.
Hou, S., Sun, W., Guo, B., Li, C., Li, X., Shao, Y., Zhang, J., 2020. Adaptive-SFSDAF for spatiotemporal image fusion that selectively uses class abundance change information. Rem. Sens. 12, 1–23. https://doi.org/10.3390/rs12233979.
Houborg, R., McCabe, M.F., Gao, F., 2016. A spatio-temporal enhancement method for medium resolution LAI (STEM-LAI). Int. J. Appl. Earth Obs. Geoinf. 47, 15–29. https://doi.org/10.1016/j.jag.2015.11.013.
Htitiou, A., Boudhar, A., Benabdelouahab, T., 2021. Deep learning-based spatiotemporal fusion approach for producing high-resolution NDVI time-series datasets. Can. J. Rem. Sens. 47, 182–197. https://doi.org/10.1080/07038992.2020.1865141.
Huang, B., Wang, J., Song, H., Fu, D., Wong, K., 2013. Generating high spatiotemporal resolution land surface temperature for urban heat island monitoring. Geosci. Rem. Sens. Lett. IEEE 10, 1011–1015. https://doi.org/10.1109/LGRS.2012.2227930.
Huang, B., Zhang, H., 2014. Spatio-temporal reflectance fusion via unmixing: accounting for both phenological and land-cover changes. Int. J. Rem. Sens. 35, 6213–6233. https://doi.org/10.1080/01431161.2014.951097.
Hutengs, C., Vohland, M., 2016. Downscaling land surface temperatures at regional scales with random forest regression. Remote Sens. Environ. 178, 127–141. https://doi.org/10.1016/j.rse.2016.03.006.
Jia, D., Cheng, C., Shen, S., Ning, L., 2022. Multitask deep learning framework for spatiotemporal fusion of NDVI. IEEE Trans. Geosci. Rem. Sens. 60. https://doi.org/10.1109/TGRS.2021.3140144.
Jia, D., Cheng, C., Song, C., Shen, S., Ning, L., Zhang, T., 2021. A hybrid deep learning-based spatiotemporal fusion method for combining satellite images with different resolutions. Rem. Sens. 13, 1–33. https://doi.org/10.3390/rs13040645.
Jia, D., Song, C., Cheng, C., Shen, S., Ning, L., Hui, C., 2020. A novel deep learning-based spatiotemporal fusion method for combining satellite images with different resolutions using a two-stream convolutional neural network. Rem. Sens. 12. https://doi.org/10.3390/rs12040698.
Jiang, J., Zhang, Q., Yao, X., Tian, Y., Zhu, Y., Cao, W., Cheng, T., Cheng, T., 2020. HISTIF: a new spatiotemporal image fusion method for high-resolution monitoring of crops at the subfield level. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 13, 4607–4626. https://doi.org/10.1109/JSTARS.2020.3016135.
Jiang, K., Wang, Z., Yi, P., Wang, G., Lu, T., Jiang, J., 2019. Edge-enhanced GAN for remote sensing image superresolution. IEEE Trans. Geosci. Rem. Sens. 57, 5799–5812. https://doi.org/10.1109/TGRS.2019.2902431.
Jiang, X., Huang, B., 2022. Unmixing-based spatiotemporal image fusion accounting for complex land cover changes. IEEE Trans. Geosci. Rem. Sens. 60.
Kattenborn, T., Leitloff, J., Schiefer, F., Hinz, S., 2021. Review on convolutional neural networks (CNN) in vegetation remote sensing. ISPRS J. Photogrammetry Remote Sens. 173, 24–49. https://doi.org/10.1016/j.isprsjprs.2020.12.010.
Ke, Y., Im, J., Park, S., Gong, H., 2017. Spatiotemporal downscaling approaches for monitoring 8-day 30 m actual evapotranspiration. ISPRS J. Photogrammetry Remote Sens. 126, 79–93. https://doi.org/10.1016/j.isprsjprs.2017.02.006.
Ke, Y., Im, J., Park, S., Gong, H., 2016. Downscaling of MODIS One kilometer evapotranspiration using Landsat-8 data and machine learning approaches. Rem. Sens. 8, 1–26. https://doi.org/10.3390/rs8030215.
Kwan, C., Budavari, B., Gao, F., Zhu, X., 2018. A hybrid color mapping approach to fusing MODIS and Landsat images for forward prediction. Rem. Sens. 10, 1–19. https://doi.org/10.3390/rs10040520.
Lei, D., Ran, G., Zhang, L., Li, W., 2022. A spatiotemporal fusion method based on multiscale feature extraction and spatial channel attention mechanism. Rem. Sens. 14, 461. https://doi.org/10.3390/rs14030461.
Li, A., Bo, Y., Zhu, Y., Guo, P., Bi, J., He, Y., 2013. Blending multi-resolution satellite sea surface temperature (SST) products using Bayesian maximum entropy method. Remote Sens. Environ. 135, 52–63. https://doi.org/10.1016/j.rse.2013.03.021.
Li, D., Li, Y., Yang, W., Ge, Y., Han, Q., Ma, L., Chen, Y., Li, X., 2018. An enhanced single-pair learning-based reflectance fusion algorithm with spatiotemporally extended training samples. Rem. Sens. 10. https://doi.org/10.3390/rs10081207.
Li, J., Li, Y., He, L., Chen, J., Plaza, A., 2020. Spatio-temporal fusion for remote sensing data: an overview and new benchmark. Sci. China Inf. Sci. 63, 1–17. https://doi.org/10.1007/s11432-019-2785-y.
Li, S., Kang, X., Fang, L., Hu, J., Yin, H., 2017. Pixel-level image fusion: a survey of the state of the art. Inf. Fusion 33, 100–112. https://doi.org/10.1016/j.inffus.2016.05.004.
Li, W., Cao, D., Peng, Y., Yang, C., 2021a. Msnet: a multi-stream fusion network for remote sensing spatiotemporal fusion based on transformer and convolution. Rem. Sens. 13. https://doi.org/10.3390/rs13183724.
Li, W., Fu, H., Yu, L., Gong, P., Feng, D., Li, C., Clinton, N., 2016. Stacked Autoencoder-based deep learning for remote-sensing image classification: a case study of African land-cover mapping. Int. J. Rem. Sens. 37, 5632–5646. https://doi.org/10.1080/01431161.2016.1246775.
Li, W., Yang, C., Peng, Y., Du, J., 2022. A pseudo-siamese deep convolutional neural network for spatiotemporal satellite image fusion. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 15, 1205–1220. https://doi.org/10.1109/JSTARS.2022.3143464.
Li, W., Yang, C., Peng, Y., Zhang, X., 2021b. A multi-cooperative deep convolutional neural network for spatiotemporal satellite image fusion. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 14, 10174–10188. https://doi.org/10.1109/JSTARS.2021.3113163.
Li, W., Zhang, X., Peng, Y., Dong, M., 2021c. Spatiotemporal fusion of remote sensing images using a convolutional neural network with attention and multiscale mechanisms. Int. J. Rem. Sens. 42, 1973–1993. https://doi.org/10.1080/01431161.2020.1809742.
Li, W., Zhang, X., Peng, Y., Dong, M., 2020. DMNet: a network architecture using dilated convolution and multiscale mechanisms for spatiotemporal fusion of remote sensing images. IEEE Sensor. J. 20, 12190–12202. https://doi.org/10.1109/JSEN.2020.3000249.
Li, X., Foody, G.M., Boyd, D.S., Ge, Y., Zhang, Y., Du, Y., Ling, F., 2020. SFSDAF: an enhanced FSDAF that incorporates sub-pixel class fraction change information for spatio-temporal image fusion. Remote Sens. Environ. 237, 111537. https://doi.org/10.1016/j.rse.2019.111537.
Li, X., Ling, F., Foody, G.M., Ge, Y., Zhang, Y., Du, Y., 2017. Generating a series of fine spatial and temporal resolution land cover maps by fusing coarse spatial resolution remotely sensed images and fine spatial resolution land cover maps. Remote Sens. Environ. 196, 293–311. https://doi.org/10.1016/j.rse.2017.05.011.
Li, Y., Li, J., He, L., Chen, J., Plaza, A., 2020. A new sensor bias-driven spatio-temporal fusion model based on convolutional neural networks. Sci. China Inf. Sci. 63. https://doi.org/10.1007/s11432-019-2805-y.
Li, Y., Ren, Y., Gao, W., Jia, J., Tao, S., Liu, X., 2022. An enhanced spatiotemporal fusion method – implications for DNN based time-series LAI estimation by using

Sentinel-2 and MODIS. Field Crop. Res. 279, 108452. https://doi.org/10.1016/j.fcr.2022.108452.
Liao, C., Wang, J., Pritchard, I., Liu, J., Shang, J., 2017. A spatio-temporal data fusion model for generating NDVI time series in heterogeneous regions. Rem. Sens. 9, 1–28. https://doi.org/10.3390/rs9111125.
Liao, L., Song, J., Wang, Jindi, Xiao, Z., Wang, Jian, 2016. Bayesian method for building frequent landsat-like NDVI datasets by integrating MODIS and landsat NDVI. Rem. Sens. 8. https://doi.org/10.3390/rs8060452.
Lin, A., Chen, B., Xu, J., Zhang, Z., Lu, G., Zhang, D., 2022. DS-TransUNet: dual swin transformer U-net for medical image segmentation. IEEE Trans. Instrum. Meas. 71, 1–13. https://doi.org/10.1109/TIM.2022.3178991.
Liu, H., Weng, Q., 2012. Enhancing temporal resolution of satellite imagery for public health studies: a case study of West Nile Virus outbreak in Los Angeles in 2007. Remote Sens. Environ. 117, 57–71. https://doi.org/10.1016/j.rse.2011.06.023.
Liu, Maolin, Ke, Y., Yin, Q., Chen, X., Im, J., 2019. Comparison of five spatio-temporal satellite image fusion models over landscapes with various spatial heterogeneity and temporal variation. Rem. Sens. 11. https://doi.org/10.3390/rs11222612.
Liu, M., Liu, X., Wu, L., Zou, X., Jiang, T., Zhao, B., 2018. A modified spatiotemporal fusion algorithm using phenological information for predicting reflectance of paddy rice in southern China. Rem. Sens. 10. https://doi.org/10.3390/rs10050772.
Liu, Meng, Yang, W., Zhu, X., Chen, J., Chen, X., Yang, L., Helmer, E.H., 2019. An Improved Flexible Spatiotemporal DAta Fusion (IFSDAF) method for producing high spatiotemporal resolution normalized difference vegetation index time series. Remote Sens. Environ. 227, 74–89. https://doi.org/10.1016/j.rse.2019.03.012.
Liu, Q., Zhou, F., Hang, R., Yuan, X., 2017. Bidirectional-convolutional LSTM based spectral-spatial feature learning for hyperspectral image classification. Rem. Sens. 9, 1330. https://doi.org/10.3390/rs9121330.
Liu, W., Zeng, Y., Li, S., Huang, W., 2020. Spectral unmixing based spatiotemporal downscaling fusion approach. Int. J. Appl. Earth Obs. Geoinf. 88, 102054. https://doi.org/10.1016/j.jag.2020.102054.
Liu, W., Zeng, Y., Li, S., Pi, X., Huang, W., 2019. An improved spatiotemporal fusion approach based on multiple endmember spectral mixture analysis. Sensors 19. https://doi.org/10.3390/s19112443.
Liu, X., Deng, C., Chanussot, J., Hong, D., Zhao, B., 2019. StfNet: a two-stream convolutional neural network for spatiotemporal image fusion. IEEE Trans. Geosci. Rem. Sens. 57, 6552–6564. https://doi.org/10.1109/TGRS.2019.2907310.
Liu, X., Deng, C., Wang, S., Huang, G., Zhao, B., Lauren, P., 2016. Fast and Accurate Spatiotemporal Fusion Based upon Extreme Learning Machine, vol. 13. pp. 2039–2043.
Liu, Y., Zhang, Yao, Wang, Y., Hou, F., Yuan, J., Tian, J., Zhang, Yang, Shi, Z., Fan, J., He, Z., 2021. A Survey of Visual Transformers 1–23.
Lu, M., Chen, J., Tang, H., Rao, Y., Yang, P., Wu, W., 2016. Land cover change detection by integrating object-based data blending model of Landsat and MODIS. Remote Sens. Environ. 184, 374–386. https://doi.org/10.1016/j.rse.2016.07.028.
Luong, M.T., Pham, H., Manning, C.D., 2015. Effective approaches to attention-based neural machine translation. Conf. Proc. - EMNLP 2015 Conf. Empir. Methods Nat. Lang. Process. 1412–1421. https://doi.org/10.18653/v1/d15-1166.
Luppino, L.T., Hansen, M.A., Kampffmeyer, M., Bianchi, F.M., Moser, G., Jenssen, R., Anfinsen, S.N., 2022. Code-aligned autoencoders for unsupervised change detection in multimodal remote sensing images. IEEE Transact. Neural Networks Learn. Syst. https://doi.org/10.1109/TNNLS.2022.3172183.
Ma, J., Zhang, W., Marinoni, A., Gao, L., Zhang, B., 2018. An improved spatial and temporal reflectance unmixing model to synthesize time series of landsat-like images. Rem. Sens. 10. https://doi.org/10.3390/rs10091388.
Ma, Y., Wei, J., Tang, W., Tang, R., 2021. Explicit and stepwise models for spatiotemporal fusion of remote sensing images with deep neural networks. Int. J. Appl. Earth Obs. Geoinf. 105, 102611. https://doi.org/10.1016/j.jag.2021.102611.
Malleswara Rao, J., Rao, C.V., Senthil, K., Lakshmi, B., Dadhwal, V.K., 2015. Spatiotemporal data fusion using temporal high-pass modulation and edge primitives. IEEE Trans. Geosci. Rem. Sens. 53, 5853–5860. https://doi.org/10.1109/TGRS.2015.2422712.
Mao, X., Li, Q., Xie, H., Lau, R.Y.K., Wang, Z., Smolley, S.P., 2017. Least squares generative adversarial networks. BProceedings IEEE Int. Conf. Comput. Vis 2794–2802. https://doi.org/10.1080/0142569900110108.
Markham, B.L., 1984. Characterization of the Landsat Sensors' Spatial Responses. pp. 864–875.
Maselli, F., Chiesi, M., Pieri, M., 2019. A new method to enhance the spatial features of multitemporal NDVI image series. IEEE Trans. Geosci. Rem. Sens. 57, 4967–4979. https://doi.org/10.1109/TGRS.2019.2894850.
Meng, J., Du, X., Wu, B., 2013. Generation of high spatial and temporal resolution NDVI and its application in crop biomass estimation. Int. J. Digit. Earth 6, 203–218. https://doi.org/10.1080/17538947.2011.623189.
Mirza, M., Osindero, S., 2014. Conditional Generative Adversarial Nets 1–7.
Mizuochi, H., Hiyama, T., Ohta, T., Fujioka, Y., Kambatuku, J.R., Iijima, M., Nasahara, K.N., 2017. Development and evaluation of a lookup-table-based approach to data fusion for seasonal wetlands monitoring: an integrated use of AMSR series, MODIS, and Landsat. Remote Sens. Environ. 199, 370–388. https://doi.org/10.1016/j.rse.2017.07.026.
Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., 2009. Preferred reporting Items for systematic reviews and meta-analyses :the PRISMA statement. Ann. Intern. Med. 151, 264–269.
Moosavi, V., Talebi, A., Mokhtari, M.H., Shamsi, S.R.F., Niazi, Y., 2015. A wavelet-artificial intelligence fusion approach (WAIFA) for blending Landsat and MODIS surface temperature. Remote Sens. Environ. 169, 243–254. https://doi.org/10.1016/j.rse.2015.08.015.
Moreno-Martínez, Á., Izquierdo-Verdiguier, E., Maneta, M.P., Camps-Valls, G., Robinson, N., Muñoz-Marí, J., Sedano, F., Clinton, N., Running, S.W., 2020. Multispectral high resolution sensor fusion for smoothing and gap-filling in the cloud. Remote Sens. Environ. 247, 111901. https://doi.org/10.1016/j.rse.2020.111901.
Mustafa, A.A., Singh, M., Sahoo, R.N., Ahmed, N., Khanna, M., Sarangi, A., 2011. Land suitability analysis for different crops: a multi criteria decision making approach using remote sensing and gis. Water Technol. 3, 61–84.
Peng, K., Wang, Q., Tang, Y., Tong, X., Atkinson, P.M., 2022. Geographically weighted spatial unmixing for spatiotemporal fusion. IEEE Trans. Geosci. Rem. Sens. 60. https://doi.org/10.1109/TGRS.2021.3115136.
Peng, M., Zhang, L., Sun, X., Cen, Y., Zhao, X., 2022. A synchronous long time-series completion method using 3-D fully convolutional neural networks. Geosci. Rem. Sens. Lett. IEEE 19. https://doi.org/10.1109/LGRS.2021.3055847.
Peng, M., Zhang, L., Sun, X., Cen, Y., Zhao, X., 2020. A fast three-dimensional convolutional neural network-based spatiotemporal fusion method (STF3DCNN) using a spatial-temporal-spectral dataset. Rem. Sens. 12, 1–20. https://doi.org/10.3390/rs12233888.
Peng, Y., Li, W., Luo, X., Du, J., Zhang, X., Gan, Y., Gao, X., 2022. Spatiotemporal reflectance fusion via tensor sparse representation. IEEE Trans. Geosci. Rem. Sens. 60. https://doi.org/10.1109/TGRS.2021.3091157.
Peng, Y., Li, W., Luo, X., Du, J., Zhang, X., Gan, Y., Gao, X., 2012. Spatiotemporal reflectance fusion via sparse representation. IEEE Trans. Geosci. Rem. Sens. 50, 3707–3716. https://doi.org/10.1109/TGRS.2021.3091157.
Pettorelli, N., Laurance, W.F., O'Brien, T.G., Wegmann, M., Nagendra, H., Turner, W., 2014. Satellite remote sensing for applied ecologists: opportunities and challenges. J. Appl. Ecol. 51, 839–848. https://doi.org/10.1111/1365-2664.12261.
Ping, B., Meng, Y., Su, F., 2018. An enhanced linear spatio-temporal fusion method for blending Landsat and MODIS data to synthesize Landsat-like imagery. Rem. Sens. 10. https://doi.org/10.3390/rs10060881.
Qiu, Y., Zhou, J., Chen, J., Chen, X., 2021. Spatiotemporal fusion method to simultaneously generate full-length normalized difference vegetation index time series (SSFIT). Int. J. Appl. Earth Obs. Geoinf. 100, 102333. https://doi.org/10.1016/j.jag.2021.102333.
Quan, J., Zhan, W., Ma, T., Du, Y., Guo, Z., Qin, B., 2018. An integrated model for generating hourly Landsat-like land surface temperatures over heterogeneous landscapes. Remote Sens. Environ. 206, 403–423. https://doi.org/10.1016/j.rse.2017.12.003.
Rabbi, J., Ray, N., Schubert, M., Chowdhury, S., Chao, D., 2020. Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network. Rem. Sens. 12, 1–25. https://doi.org/10.3390/RS12091432.
Rao, Y., Zhu, X., Chen, J., Wang, J., 2015. An improved method for producing high spatial-resolution NDVI time series datasets with multi-temporal MODIS NDVI data and Landsat TM/ETM+ images. Rem. Sens. 7, 7865–7891. https://doi.org/10.3390/rs70607865.

Ren, Y., Zhu, C., Xiao, S., 2018. Small object detection in optical remote sensing images via modified Faster R-CNN. Appl. Sci. 8. https://doi.org/10.3390/app8050813.
Roy, D.P., Ju, J., Lewis, P., Schaaf, C., Gao, F., Hansen, M., Lindquist, E., 2008. Multi-temporal MODIS-Landsat data fusion for relative radiometric normalization, gap filling, and prediction of Landsat data. Remote Sens. Environ. 112, 3112–3130. https://doi.org/10.1016/j.rse.2008.03.009.
Sadeh, Y., Zhu, X., Dunkerley, D., Walker, J.P., Zhang, Y., Rozenstein, O., Manivasagam, V.S., Chenu, K., 2020. Sentinel-2 and planetscope data fusion into daily 3 M images for leaf area index monitoring. Int. Geosci. Remote Sens. Symp. 5274–5277. https://doi.org/10.1109/IGARSS39084.2020.9324336.
Sdraka, M., Papoutsis, I., Psomas, B., Vlachos, K., Ioannidis, K., Karantzalos, K., Gialampoukidis, I., Vrochidis, S., 2022. Deep learning for downscaling remote sensing images: fusion and super-resolution. IEEE Geosci. Remote Sens. Mag. https://doi.org/10.1109/MGRS.2022.3171836.
Settle, J.J., Drake, N.A., 1993. Linear mixing and the estimation of ground cover proportions. Int. J. Rem. Sens. 14, 1159–1177. https://doi.org/10.1080/01431169308904402.
Shang, C., Li, Xinyan, Yin, Z., Li, Xiaodong, Wang, L., Zhang, Y., Du, Y., Ling, F., 2022. Spatiotemporal reflectance fusion using a generative adversarial network. IEEE Trans. Geosci. Rem. Sens. 60. https://doi.org/10.1109/TGRS.2021.3065418.
Shao, Z., Cai, J., 2018. Remote sensing image fusion with deep convolutional neural network. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 11, 1656–1669. https://doi.org/10.1109/JSTARS.2018.2805923.
Shao, Z., Cai, J., Fu, P., Hu, L., Liu, T., 2019. Deep learning-based fusion of Landsat-8 and Sentinel-2 images for a harmonized surface reflectance product. Remote Sens. Environ. 235, 111425. https://doi.org/10.1016/j.rse.2019.111425.
Shen, H., Huang, L., Zhang, L., Wu, P., Zeng, C., 2016a. Long-term and fine-scale satellite monitoring of the urban heat island effect by the fusion of multi-temporal and multi-sensor remote sensed data: a 26-year case study of the city of Wuhan in China. Remote Sens. Environ. 172, 109–125. https://doi.org/10.1016/j.rse.2015.11.005.
Shen, H., Meng, X., Zhang, L., 2016b. An integrated framework for the spatio-temporal-spectral fusion of remote sensing images. IEEE Trans. Geosci. Rem. Sens. 54, 7135–7148. https://doi.org/10.1109/TGRS.2016.2596290.
Shen, H., Wu, P., Liu, Y., Ai, T., Wang, Y., Liu, X., 2013. A spatial and temporal reflectance fusion model considering sensor observation differences. Int. J. Rem. Sens. 34, 4367–4383. https://doi.org/10.1080/01431161.2013.777488.
Shi, C., Wang, L., 2016. Linear spatial spectral mixture model. IEEE Trans. Geosci. Rem. Sens. 54, 3599–3611. https://doi.org/10.1109/TGRS.2016.2520399.
Shi, C., Wang, X., Zhang, M., Liang, X., Niu, L., Han, H., Zhu, X., 2019. A comprehensive and automated fusion method: the enhanced flexible spatiotemporal data fusion model for monitoring dynamic changes of land surface. Appl. Sci. 9, 1–19. https://doi.org/10.3390/app9183693.
Shi, W., Guo, D., Zhang, H., 2022. A reliable and adaptive spatiotemporal data fusion method for blending multi-spatiotemporal-resolution satellite images. Remote Sens. Environ. 268, 112770. https://doi.org/10.1016/j.rse.2021.112770.
Singh, Devendra, 2011. Generation and evaluation of gross primary productivity using Landsat data through blending with MODIS data. Int. J. Appl. Earth Obs. Geoinf. 13, 59–69. https://doi.org/10.1016/j.jag.2010.06.007.
Singh, D., 2011. Evaluation of long-term NDVI time series derived from landsat data through blending with MODIS data. Atmósfera 25, 43–63.
Song, B., Liu, P., Li, J., Wang, L., Zhang, L., He, G., Chen, L., Liu, J., 2022. MLFF-GAN: a multilevel feature fusion with GAN for spatiotemporal remote sensing images. IEEE Trans. Geosci. Rem. Sens. 60, 1–16. https://doi.org/10.1109/TGRS.2022.3169916.
Song, H., Huang, B., 2013. Spatiotemporal satellite image fusion through one-pair image learning. IEEE Trans. Geosci. Rem. Sens. 51, 1883–1896. https://doi.org/10.1109/TGRS.2012.2213095.
Song, H., Liu, Q., Wang, G., 2018. Spatiotemporal satellite image fusion. Using Deep Convolutional Neural Networks 11, 821–829.
Su, X., Li, J., Hua, Z., 2022. Transformer-based regression network for pansharpening remote sensing images. IEEE Trans. Geosci. Rem. Sens. 60. https://doi.org/10.1109/TGRS.2022.3152425.
Sun, H., Xiao, W., 2022. Similarity weight learning: a new spatial and temporal satellite image fusion framework. IEEE Trans. Geosci. Rem. Sens. 60, 1–17. https://doi.org/10.1109/tgrs.2022.3161070.
Sun, R., Chen, S., Su, H., Mi, C., Jin, N., 2019. The effect of NDVI time series density derived from spatiotemporal fusion of multisource remote sensing data on crop classification accuracy. ISPRS Int. J. Geo-Inf. 8, 502. https://doi.org/10.3390/ijgi8110502.
Sun, S., Mu, L., Wang, L., Liu, P., 2022. L-UNet: an LSTM network for remote sensing image change detection. Geosci. Rem. Sens. Lett. IEEE 19. https://doi.org/10.1109/LGRS.2020.3041530.
Sun, Y., Zhang, H., 2019. A two-stage spatiotemporal fusion method for remote sensing images. Photogramm. Eng. Rem. Sens. 85, 907–914. https://doi.org/10.14358/PERS.85.12.907.
Sun, Y., Zhang, H., Shi, W., 2019. A spatio-temporal fusion method for remote sensing data Using a linear injection model and local neighbourhood information. Int. J. Rem. Sens. 40, 2965–2985. https://doi.org/10.1080/01431161.2018.1538585.
Tan, Z., Di, L., Zhang, M., Guo, L., 2019. An enhanced deep convolutional model for spatiotemporal image fusion. Rem. Sens. 1–24. https://doi.org/10.3390/rs11242898.
Tan, Z., Gao, M., Li, X., Jiang, L., 2021. A flexible reference-insensitive spatiotemporal fusion model for remote sensing images using conditional generative adversarial network. IEEE Trans. Geosci. Rem. Sens. 1–13.
Tan, Z., Gao, M., Yuan, J., Jiang, L., Duan, H., 2022. A robust model for MODIS and landsat image fusion considering input noise. IEEE Trans. Geosci. Rem. Sens. 60. https://doi.org/10.1109/TGRS.2022.3145086.
Tan, Z., Yue, P., Di, L., Tang, J., 2018. Deriving high spatiotemporal remote sensing images using deep convolutional network. Rem. Sens. 10, 1–16. https://doi.org/10.3390/rs10071066.
Tang, J., Zeng, J., Zhang, L., Zhang, R., Li, J., Li, X., Zou, J., Zeng, Y., 2019. A modified flexible spatiotemporal data fusion model. Front. Earth Sci. 1–14.
Tao, Y., Xu, M., Zhong, Y., Cheng, Y., 2017. GAN-assisted two-stream neural network for high-resolution remote sensing image classification. Rem. Sens. 9, 1328. https://doi.org/10.3390/rs9121328.
Teo, T.A., Fu, Y.J., 2021. Spatiotemporal fusion of formosat-2 and landsat-8 satellite images: a comparison of "super resolution-then-blend" and "blend-then-super resolution" approaches. Rem. Sens. 13, 1–20. https://doi.org/10.3390/rs13040606.
Tian, F., Wang, Y., Fensholt, R., Wang, K., Zhang, L., Huang, Y., 2013. Mapping and evaluation of NDVI trends from synthetic time series obtained by blending landsat and MODIS data around a coalfield on the loess plateau. Rem. Sens. 5, 4255–4279. https://doi.org/10.3390/rs5094255.
Tian, H., Wang, P., Tansey, K., Zhang, J., Zhang, S., Li, H., 2021. An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the Guanzhong Plain, PR China. Agric. For. Meteorol. 310, 108629. https://doi.org/10.1016/j.agrformet.2021.108629.
Tobler, W.R., 1970. A computer movie simulating urban growth in the detroit region. Econ. Geogr. 46, 234–240.
Udelhoven, T., 2012. Long term data fusion for a dense time series analysis with MODIS and Landsat imagery in an Australian Savanna. J. Appl. Remote Sens. 6, 063512. https://doi.org/10.1117/1.jrs.6.063512.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 1–11. https://doi.org/10.1109/2943.974352.
Walker, J., de Beurs, K., Wynne, R.H., 2015. Phenological response of an Arizona dryland forest to short-term climatic extremes. Rem. Sens. 7, 10832–10855. https://doi.org/10.3390/rs70810832.
Walker, J.J., De Beurs, K.M., Wynne, R.H., Gao, F., 2012. Evaluation of Landsat and MODIS data fusion products for analysis of dryland forest phenology. Remote Sens. Environ. 117, 381–393. https://doi.org/10.1016/j.rse.2011.10.014.
Wang, J., Huang, B., 2018. A spatiotemporal satellite image fusion model with autoregressive error correction (AREC). Int. J. Rem. Sens. 39, 6731–6756. https://doi.org/10.1080/01431161.2018.1466073.
Wang, J., Huang, B., 2017. A rigorously-weighted spatiotemporal fusion model with uncertainty analysis. Rem. Sens. 9. https://doi.org/10.3390/rs9100990.
Wang, J., Schmitz, O., Lu, M., Karssenberg, D., 2020. Thermal unmixing based downscaling for fine resolution diurnal land surface temperature analysis. ISPRS J. Photogrammetry Remote Sens. 161, 76–89. https://doi.org/10.1016/j.isprsjprs.2020.01.014.
Wang, L., Member, G.S., Li, R., Duan, C., 2022. Scheme Fine-Resolut. Rem. Sens. Imag. 19, 1–5.
Wang, L., Wang, X., Wang, Q., 2020. Using 250-m modis data for enhancing spatiotemporal fusion by sparse representation. Photogramm. Eng. Rem. Sens. 86, 383–392.

https://doi.org/10.14358/PERS.86.6.383.
Wang, P., Gao, F., Masek, J.G., 2014. Operational data fusion framework for building frequent landsat-like imagery. IEEE Trans. Geosci. Rem. Sens. 52, 7353–7365. https://doi.org/10.1109/TGRS.2014.2311445.
Wang, Q., Atkinson, P.M., 2018. Spatio-temporal fusion for daily Sentinel-2 images. Remote Sens. Environ. 204, 31–42. https://doi.org/10.1016/j.rse.2017.10.046.
Wang, Q., Blackburn, G.A., Onojeghuo, A.O., Dash, J., Zhou, L., Zhang, Y., Atkinson, P.M., 2017a. Fusion of landsat 8 OLI and sentinel-2 MSI data. IEEE Trans. Geosci. Rem. Sens. 55, 3885–3899. https://doi.org/10.1109/TGRS.2017.2683444.
Wang, Q., Ding, X., Tong, X., Atkinson, P.M., 2021a. Spatio-temporal spectral unmixing of time-series images. Remote Sens. Environ. 259, 112407. https://doi.org/10.1016/j.rse.2021.112407.
Wang, Q., Peng, K., Tang, Y., Tong, X., Atkinson, P.M., 2021b. Blocks-removed spatial unmixing for downscaling MODIS images. Remote Sens. Environ. 256, 112325. https://doi.org/10.1016/j.rse.2021.112325.
Wang, Q., Tang, Y., Tong, X., Atkinson, P.M., 2020. Virtual image pair-based spatio-temporal fusion. Remote Sens. Environ. 249, 112009. https://doi.org/10.1016/j.rse.2020.112009.
Wang, Q., Zhang, Y., Onojeghuo, A.O., Zhu, X., Atkinson, P.M., 2017b. Enhancing spatio-temporal fusion of MODIS and landsat data by incorporating 250 m MODIS data. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 10, 4116–4123. https://doi.org/10.1109/JSTARS.2017.2701643.
Wang, S., Wang, C., Zhang, C., Xue, J., Wang, P., Wang, X., Wang, W., Zhang, X., Li, W., Huang, G., Huo, Z., 2022. A classification-based spatiotemporal adaptive fusion model for the evaluation of remotely sensed evapotranspiration in heterogeneous irrigated agricultural area. Remote Sens. Environ. 273, 112962. https://doi.org/10.1016/j.rse.2022.112962.
Wang, Xiaofei, Wang, Xiaoyi, 2020. Spatiotemporal fusion of remote sensing image based on deep learning. 2020. J. Sens. https://doi.org/10.1155/2020/8873079.
Watts, J.D., Powell, S.L., Lawrence, R.L., Hilker, T., 2011. Improved classification of conservation tillage adoption using high temporal and synthetic satellite imagery. Remote Sens. Environ. 115, 66–75. https://doi.org/10.1016/j.rse.2010.08.005.
Wei, J., Tang, W., He, C., 2021. Enblending mosaicked remote sensing images with spatiotemporal fusion of convolutional neural networks. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 14, 5891–5902. https://doi.org/10.1109/JSTARS.2021.3082619.
Wei, J., Wang, L., Liu, P., Chen, X., Li, W., Zomaya, A.Y., 2017a. Spatiotemporal fusion of MODIS and landsat-7 reflectance images via compressed sensing. IEEE Trans. Geosci. Rem. Sens. 55, 7126–7139. https://doi.org/10.1109/TGRS.2017.2742529.
Wei, J., Wang, L., Liu, P., Song, W., 2017b. Spatiotemporal fusion of remote sensing images with structural sparsity and semi-coupled dictionary learning. Rem. Sens. 9, 1–16. https://doi.org/10.3390/rs9010021.
Weng, Q., Fu, P., Gao, F., 2014. Generating daily land surface temperature at Landsat resolution by fusing Landsat and MODIS data. Remote Sens. Environ. 145, 55–67. https://doi.org/10.1016/j.rse.2014.02.003.
Wu, B., Huang, B., Cao, K., Zhuo, G., 2017. Improving spatiotemporal reflectance fusion using image inpainting and steering kernel regression techniques. Int. J. Rem. Sens. 38, 706–727. https://doi.org/10.1080/01431161.2016.1271471.
Wu, B., Huang, B., Zhang, L., Member, S., 2015. An error-bound-regularized sparse coding for spatiotemporal reflectance fusion. IEEE Trans. Geosci. Rem. Sens. 1–13.
Wu, J., Lin, L., Li, T., Cheng, Q., Zhang, C., Shen, H., 2022. Fusing Landsat 8 and Sentinel-2 data for 10-m dense time-series imagery using a degradation-term constrained deep network. Int. J. Appl. Earth Obs. Geoinf. 108. https://doi.org/10.1016/j.jag.2022.102738.
Wu, M., Huang, W., Niu, Z., Wang, C., 2015. Generating daily synthetic landsat imagery by combining landsat and MODIS data. Sensors 15, 24002–24025. https://doi.org/10.3390/s150924002.
Wu, M., Niu, Z., Wang, C., Li, W., 2012. Use of MODIS and Landsat time series data to generate high-resolution temporal synthetic Landsat data using a spatial and temporal reflectance fusion model. J. Appl. Remote Sens. 6, 063507. https://doi.org/10.1117/1.jrs.6.063507.
Wu, P., Shen, H., Ai, T., Liu, Y., 2013. Land-surface temperature retrieval at high spatial and temporal resolutions based on multi-sensor fusion. Int. J. Digit. Earth 6, 113–133. https://doi.org/10.1080/17538947.2013.783131.
Wu, P., Shen, H., Zhang, L., Göttsche, F.M., 2015. Integrated fusion of multi-scale polar-orbiting and geostationary satellite observations for the mapping of high spatial and temporal resolution land surface temperature. Remote Sens. Environ. 156, 169–181. https://doi.org/10.1016/j.rse.2014.09.013.
Xia, H., Chen, Y., Li, Y., Quan, J., 2019. Combining kernel-driven and fusion-based methods to generate daily high-spatial-resolution land surface temperatures. Remote Sens. Environ. 224, 259–274. https://doi.org/10.1016/j.rse.2019.02.006.
Xiao, G., Prasad Bavirisetti, D., Liu, G., Zhang, X., 2020. Decision-level image fusion. In: Image Fusion. pp. 149–170.
Xie, D., Gao, F., Sun, L., Anderson, M., 2018. Improving spatial-temporal data fusion by choosing optimal input image pairs. Rem. Sens. 10. https://doi.org/10.3390/rs10071142.
Xie, D., Zhang, J., Zhu, X., Pan, Y., Liu, H., Yuan, Z., Yun, Y., 2016. An improved STARFM with help of an unmixing-based method to generate high spatial and temporal resolution remote sensing data in complex heterogeneous regions. Sensors 16. https://doi.org/10.3390/s16020207.
Xiao, J., Suab, S.A., Chen, X., Singh, C.K., Singh, D., Aggarwal, A.K., Avtar, R., 2023. Enhancing assessment of corn growth performance using unmanned aerial vehicles (UAVs) and deep learning. Measurement 214, 112764.
Xie, X.L., Beni, G., 2011. A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell. 33, 865–866. https://doi.org/10.1109/TPAMI.2011.60.
Xu, C., Qu, J.J., Hao, X., Cosh, M.H., Prueger, J.H., Zhu, Z., Gutenberg, L., 2018. Downscaling of surface soil moisture retrieval by combining MODIS/Landsat and in situ measurements. Rem. Sens. 10. https://doi.org/10.3390/rs10020210.
Xu, X., Feng, Z., Cao, C., Li, M., Wu, J., Wu, Z., Shang, Y., Ye, S., 2021. An improved swin transformer-based model for remote sensing object detection and instance segmentation. Rem. Sens. 13. https://doi.org/10.3390/rs13234779.
Xu, Yong, Huang, B., Xu, Yuyue, Cao, K., Guo, C., Meng, D., 2015. Spatial and temporal image fusion via regularized spatial unmixing. Geosci. Rem. Sens. Lett. IEEE 12, 1362–1366. https://doi.org/10.1109/LGRS.2015.2402644.
Xue, J., Leung, Y., Fung, T., 2019. An unmixing-based Bayesian model for spatio-temporal satellite image fusion in heterogeneous landscapes. Rem. Sens. 11. https://doi.org/10.3390/rs11030324.
Xue, J., Leung, Y., Fung, T., 2017. A bayesian data fusion approach to spatio-temporal fusion of remotely sensed images. Rem. Sens. 9. https://doi.org/10.3390/rs9121310.
Yamashita, R., Nishio, M., Do, R.K.G., Togashi, K., 2018. Convolutional neural networks: an overview and application in radiology. Insights Imag. 9, 611–629. https://doi.org/10.1007/978-981-15-7078-0_3.
Yan, Y., Liu, X., Ou, J., Li, X., Wen, Y., 2018. Assimilating multi-source remotely sensed data into a light use efficiency model for net primary productivity estimation. Int. J. Appl. Earth Obs. Geoinf. 72, 11–25. https://doi.org/10.1016/j.jag.2018.05.013.
Yang, J., Yao, Y., Wei, Y., Zhang, Y., Jia, K., Zhang, X., 2020. A robust method for generating high-spatiotemporal-resolution surface reflectance. Rem. Sens. 12, 2312.
Yang, L., Song, J., Han, L., Wang, X., Wang, J., 2020. Reconstruction of high-temporal-and high-spatial-resolution reflectance datasets using difference construction and bayesian unmixing. Rem. Sens. 12, 1–26. https://doi.org/10.3390/rs12233952.
Yang, S., Gu, L., Li, X., Gao, F., Jiang, T., 2022. Fully automated classification method for crops based on spatiotemporal deep-learning fusion technology. IEEE Trans. Geosci. Rem. Sens. 60. https://doi.org/10.1109/TGRS.2021.3113014.
Yang, Z., Diao, C., Li, B., 2021. A robust hybrid deep learning model for spatiotemporal image fusion. Rem. Sens. 13. https://doi.org/10.3390/rs13245005.
Yin, G., Li, A., Jin, H., Bian, J., 2018. Spatiotemporal fusion through the best linear unbiased estimator to generate fine spatial resolution NDVI time series. Int. J. Rem. Sens. 39, 3287–3305. https://doi.org/10.1080/01431161.2018.1439202.
Yin, Z., Wu, P., Foody, G.M., Wu, Y., Liu, Z., Du, Y., Ling, F., 2021. Spatiotemporal fusion of land surface temperature based on a convolutional neural network. IEEE Trans. Geosci. Rem. Sens. 59, 1808–1822. https://doi.org/10.1109/TGRS.2020.2999943.
Ying, H., Leung, Y., Cao, F., Fung, T., Xue, J., 2018. Sparsity-based spatiotemporal fusion via adaptive multi-band constraints. Rem. Sens. 10, 1–18. https://doi.org/10.3390/rs10101646.
Zha, C., Min, W., Han, Q., Xiong, X., Wang, Q., Liu, Q., 2022. Multiple granularity spatiotemporal network for sea surface temperature prediction. Geosci. Rem. Sens. Lett. IEEE 19.
Zhai, H., Huang, F., Qi, H., 2020. Generating high resolution LAI based on a modified FSDAF model. Rem. Sens. 12. https://doi.org/10.3390/RS12010150.

Zhang, B., Zhang, L., Xie, D., Yin, X., Liu, C., Liu, G., 2016. Application of synthetic NDVI time series blended from landsat and MODIS data for grassland biomass estimation. Rem. Sens. 8, 1–21. https://doi.org/10.3390/rs8010010.
Zhang, H., Ma, J., 2021. GTP-PNet: a residual learning network based on gradient transformation prior for pansharpening. ISPRS J. Photogrammetry Remote Sens. 172, 223–239. https://doi.org/10.1016/j.isprsjprs.2020.12.014.
Zhang, Hongyan, Song, Y., Han, C., Zhang, L., 2021. Remote sensing image spatiotemporal fusion using a generative adversarial network. IEEE Trans. Geosci. Rem. Sens. 59, 4273–4286. https://doi.org/10.1109/TGRS.2020.3010530.
Zhang, Hua, Sun, Y., Shi, W., Guo, D., Zheng, N., 2021. An object-based spatiotemporal fusion model for remote sensing images. Eur. J. Remote Sens. 54, 86–101. https://doi.org/10.1080/22797254.2021.1879683.
Zhang, L., Yao, Y., Bei, X., Li, Y., Shang, K., Yang, J., Guo, X., Yu, R., Xie, Z., 2021. ERTFM: an effective model to fuse Chinese GF-1 and MODIS reflectance data for terrestrial latent heat flux estimation. Rem. Sens. 13, 1–23. https://doi.org/10.3390/rs13183703.
Zhang, M., Zeng, Y., Huang, W., Li, S., 2019. Combining spatiotemporal fusion and object-based image analysis for improving wetland mapping in complex and heterogeneous urban landscapes. Geocarto Int. 34, 1144–1161. https://doi.org/10.1080/10106049.2018.1474275.
Zhang, W., Li, A., Jin, H., Bian, J., Zhang, Z., Lei, G., Qin, Z., Huang, C., 2013. An enhanced spatial and temporal data fusion model for fusing landsat and modis surface reflectance to generate high temporal landsat-like data. Rem. Sens. 5, 5346–5368. https://doi.org/10.3390/rs5105346.
Zhang, W., Tang, P., Zhao, L., 2019. Remote sensing image scene classification using CNN-CapsNet. Rem. Sens. 11. https://doi.org/10.3390/rs11050494.
Zhang, Y., Foody, G.M., Ling, F., Li, X., Ge, Y., Du, Y., Atkinson, P.M., 2018. Spatial-temporal fraction map fusion with multi-scale remotely sensed images. Remote Sens. Environ. 213, 162–181. https://doi.org/10.1016/j.rse.2018.05.010.
Zhang, Y., Liu, J., Liang, S., Li, M., 2022. A new spatial – temporal depthwise separable convolutional fusion network for generating landsat 8-day surface reflectance time series over forest regions. Rem. Sens.
Zhao, H., Gallo, O., Frosio, I., Kautz, J., 2015. Loss Functions for Neural Networks for Image Processing 1–11.
Zhao, W., Rowlands, J.A., 2014. Amorphous Silicon Detectors, Comprehensive Biomedical Physics. Elsevier B.V. https://doi.org/10.1016/B978-0-444-53632-7.00620-1.
Zhao, Y., Huang, B., Song, H., 2018. A robust adaptive spatial and temporal image fusion model for complex land surface changes. Remote Sens. Environ. 208, 42–62. https://doi.org/10.1016/j.rse.2018.02.009.
Zheng, Y., Song, H., Sun, L., Wu, Z., Jeon, B., 2019. Spatiotemporal fusion of satellite images via very deep convolutional networks. Rem. Sens. 11. https://doi.org/10.3390/rs11222701.
Zhong, D., Zhou, F., 2019. Improvement of clustering methods for modelling abrupt land surface changes in satellite image fusions. Rem. Sens. 11. https://doi.org/10.3390/rs11151759.
Zhong, D., Zhou, F., 2018. A prediction smooth method for blending landsat and Moderate Resolution Imagine Spectroradiometer images. Rem. Sens. 10. https://doi.org/10.3390/rs10091371.
Zhou, F., Zhong, D., Peiman, R., 2020. Reconstruction of cloud-free sentinel-2 image time-series using an extended spatiotemporal image fusion approach. Rem. Sens. 12. https://doi.org/10.36745/IJCA.341.
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. Rem. Sens. 10. https://doi.org/10.3390/rs10040527.
Zhu, X., Chen, J., Gao, F., Chen, X., Masek, J.G., 2010. An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions. Remote Sens. Environ. 114, 2610–2623. https://doi.org/10.1016/j.rse.2010.05.032.
Zhu, X., Helmer, E.H., Gao, F., Liu, D., Chen, J., Lefsky, M.A., 2016. A flexible spatiotemporal method for fusing satellite images with different resolutions. Remote Sens. Environ. 172, 165–177. https://doi.org/10.1016/j.rse.2015.11.016.
Zhu, Y., Kang, E.L., Bo, Y., Zhang, J., Wang, Y., Tang, Q., 2019. Hierarchical bayesian model based on robust fixed rank filter for fusing MODIS SST and AMSR-E SST. Photogramm. Eng. Rem. Sens. 85, 119–131. https://doi.org/10.14358/PERS.85.2.119.
Zhukov, B., Oertel, D., Lanzl, F., Reinhäckel, G., 1999. Unmixing-based multisensor multiresolution image fusion. IEEE Trans. Geosci. Rem. Sens. 37, 1212–1226. https://doi.org/10.1109/36.763276.
Zurita-Milla, R., Clevers, J.G.P.W., Schaepman, M.E., 2008. Unmixing-based landsat TM and MERIS FR data fusion. Geosci. Rem. Sens. Lett. IEEE 5, 453–457. https://doi.org/10.1109/LGRS.2008.919685.
Zurita-Milla, R., Kaiser, G., Clevers, J.G.P.W., Schneider, W., Schaepman, M.E., 2009. Downscaling time series of MERIS full resolution data to monitor vegetation seasonal dynamics. Remote Sens. Environ. 113, 1874–1885. https://doi.org/10.1016/j.rse.2009.04.011.