Remote Sensing of Environment 334 (2026) 115214

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Spatial-X fusion for multi-source satellite imageries

Jiang He[a, b], Liupeng Lin[c], Zhuo Zheng[d], Qiangqiang Yuan[e,*], Jie Li[e], Liangpei Zhang[f], Xiao xiang Zhu[a, b,**]

[a] Chair of Data Science in Earth Observation, Technical University of Munich, 80333, Munich, Germany
[b] Munich Center for Machine Learning, 80333, Munich, Germany
[c] School of Resource and Environmental Sciences, Wuhan University, 430079, Wuhan, China
[d] Department of Computer Science, Stanford University, Stanford 94305, USA
[e] School of Geodesy and Geomatics, Wuhan University, 430079, Wuhan, China
[f] State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, 430079, Wuhan, China

HIGHLIGHTS

- A unified spatial-X fusion framework is proposed.
- Multiple-degradation model-driven deep unfolding for high-performing CNN.
- Spatial-X intrinsic interaction prior is proposed to capture multimodal dependencies.
- Comprehensive validation on four major fusion tasks shows the superiority of SpaXFus.
- The systematic analysis of spatial-X fusion's benefits for downstream applications.

ARTICLE INFO

Edited by Dr Jing M. Chen

Keywords:
Model-driven
Generalized framework
Spatial-channel information
Image fusion
Remote sensing

ABSTRACT

Multi-source remote sensing data can highlight different types of information based on user needs, resulting in large volumes of data and significant challenges. Hardware and environmental constraints create mutual dependencies between information types, particularly between spatial data and other types, limiting the development of high-precision applications. Traditional methods are task-specific, leading to many algorithms without a unified solution, which greatly increases the computational and deployment costs of image fusion. In this paper, we summarize four remote sensing fusion tasks, including pan-sharpening, hyperspectral-multispectral fusion, spatio-temporal fusion, and polarimetric SAR fusion. By defining the spectral, temporal, and polarimetric information, as X, we propose the concept of generalized spatial-channel fusion, referred to as Spatial-X fusion. Then, we design an end-to-end network SpaXFus, a generalized spatial-channel fusion framework through a model-driven unfolding approach that exploits spatial-X intrinsic interactions to capture internal dependencies and self-interactions. Comprehensive experimental results demonstrate the superiority of SpaXFus, e.g., SpaXFus can achieve four remote sensing image fusion tasks with superior performance (across all fusion tasks, spectral distortion decreases by 25.48 %, while spatial details improve by 7.5 %) and shows huge improvements across multiple types of downstream applications, including vegetation index generation, fine-grained image classification, change detection, and SAR vegetation extraction.

1. Introduction

Remote sensing has emerged as a vital tool for understanding the Earth's surface, providing invaluable data for diverse applications such as environmental monitoring (Fu et al., 2022), urban planning (Benedek et al., 2011), disaster management (Zhu et al., 2010), and resource exploration (Hong et al., 2024). With the increasing availability of time series data from various sensors, such as multispectral (MS), hyperspectral (HS), and polarimetric synthetic aperture radar (SAR) imagery, the need to integrate multiple types of information has become a pressing challenge in remote sensing. Due to the limitations imposed by sensor

* Corresponding author.
** Corresponding author at: Chair of Data Science in Earth Observation, Technical University of Munich, 80333, Munich, Germany.
Email addresses: qqyuan@sgg.whu.edu.cn (Q. Yuan), xiaoxiang.zhu@tum.de (X.x. Zhu).

https://doi.org/10.1016/j.rse.2025.115214
Received 19 September 2025; Received in revised form 16 December 2025; Accepted 20 December 2025

hardware and the imaging conditions, various constraints exist among different data sources. For example, MS imagery strikes a balance between spectral and spatial resolution, HS imagery provides finer spectral information but often at the cost of spatial resolution, temporal data with high temporal resolution is typically accompanied by spatial degradation, and polarimetric SAR (PolSAR) data, while offering valuable polarimetric insights, is constrained by both polarization and spatial resolution. Consequently, no single modality can fully capture the complexity of the observed scene. In practical applications, the data acquired are often of low quality due to these inherent limitations. Consequently, there is a pressing need for the development of effective image fusion techniques that can integrate multi-source information, address these constraints and ultimately improve data quality.

Image fusion plays a critical role in maximizing the utility of remote sensing data by combining complementary information from different sensors. Over the years, numerous fusion techniques have been developed to address specific fusion tasks, such as pan-sharpening, hyperspectral-multispectral fusion (HMFusion), spatio-temporal fusion (STFusion), and PolSAR fusion. For instance, pan-sharpening aims to enhance the spatial resolution of MS images by integrating them with high-resolution panchromatic imagery, while hyperspectral-multispectral image fusion seeks to combine the rich spectral content of HS data with the high spatial resolution of MS data. Likewise, spatio-temporal fusion enables the acquisition of high-spatial-resolution (HR) time series by integrating data captured at different times, and PolSAR fusion refines the radiative characteristics of land cover by combining spatial information with polarimetric characteristics. Despite the specific goals of these tasks, they share a common theme: fusing spatial information with additional spectral, temporal, or polarimetric information to generate a more informative and accurate representation of the scene.

Although a wide range of fusion techniques have been developed for these tasks, most existing methods are task-specific and lack generalizability across different fusion scenarios. Traditional pan-sharpening approaches, such as component substitution (CS) (Chen et al., 2024; Garzelli et al., 2007), multiresolution analysis (MRA) (Liu, 2000), and variational optimization models (VO) (Ballester et al., 2006), are widely used to improve the spatial resolution of MS images. However, these methods struggle to achieve a good balance between spatial blurring and spectral distortions, and are gradually being surpassed by deep learning-based algorithms. By addressing the different stages of the pan-sharpening process, deep learning has significantly enhanced algorithm performance through various approaches, including convolutional neural networks (CNNs) (Masi et al., 2016), residual learning (Wei et al., 2017), multi-scale convolutions (Wang et al., 2021), generative adversarial networks (GANs) (Ma et al., 2020), conditional GANs (Zhou et al., 2022), attention-based CNNs (Zheng et al., 2020), transformers (Zhang et al., 2024a), model-driven CNNs (He et al., 2022), and diffusion models (Meng et al., 2023).

Similarly, hyperspectral-multispectral image fusion methods are usually inspired by pan-sharpening techniques, such as CS-based (Yokoya et al., 2011; Choi et al., 2010), MRA-based (Nunez et al., 1999), and deep learning-based methods (Yang et al., 2018). Considering the different spectral information in MS and HS images, researchers have also developed Bayesian and statistical models (Sui et al., 2019), sparse representation (Lanaras et al., 2015), and low-rank regularization (Dian et al., 2024), but these often require extensive computational resources and large training datasets, limiting their applicability to other fusion tasks, such as spatio-temporal or PolSAR fusion.

Spatio-temporal fusion algorithms can be categorized into weight function-based methods (Hilker et al., 2009), unmixing-based methods (Zhukov et al., 1999), learning-based methods (Huang and Song, 2012), Bayesian-based methods (Li et al., 2013), and hybrid methods (Gevaert and GarcÃ-a-Haro, 2015). Weight function-based methods assign importance to different input images based on their spatial and temporal characteristics, thereby optimizing the fusion process (Zhu et al., 2010). Unmixing-based methods analyze the spectral mixtures to extract underlying components, facilitating the reconstruction of HR images (Xu et al., 2015). In contrast, learning-based methods leverage machine learning techniques to capture complex mappings between images (Song and Huang, 2013). Bayesian-based methods utilize probabilistic frameworks to address uncertainty, effectively integrating multiple types of information (Shen et al., 2016). Finally, hybrid methods combine various strategies, drawing on the strengths of different approaches to improve fusion performance (Zhu et al., 2016).

Recognizing the importance of HR polarimetric information, Pastina et al. (2001) introduced polarimetric information into PolSAR image super-resolution using SPECAN techniques, which is the earliest attempt to improve the spatial resolution of SAR data. Then, polarimetric component decomposition-based methods are utilized to further improve the injection of polarimetric information, such as 2D-PBWE (Suwa and Iwamoto, 2006), projection onto convex sets algorithm (Jiong and Jian, 2007), coherent target decomposition (Zou et al., 2008), and polarimetric spatial correlation (SRPSC) (Zhang et al., 2011). Although they can extend traditional bandwidth extrapolation from SAR to PolSAR images, they inadequately utilize polarimetric data and sometimes suffer from grid effects. With the rapid development of deep learning, Lin et al. developed a series of CNN-based PolSAR fusion algorithms, including deep CNNs and residual learning (Shen et al., 2020).

In remote sensing, data often contain spatial information in terms of spatial resolution, but depending on the specific applications, other types of information are also included in the form of additional channels, such as spectral, temporal, and polarimetric data. The challenge of fusing disparate types of information is significantly compounded when dealing with high-dimensional datasets. While these data sources provide substantial information, their inherent complexity complicates the preservation and integration of all relevant features during the fusion process. Additionally, existing fusion techniques often need to take into account the characteristics involved in different tasks and develop solutions accordingly. For instance, in the context of spatial-spectral fusion techniques, such as pan-sharpening and HMFusion, the interdependence between spatial and spectral information further complicates the pursuit of balance. In contrast, in spatio-temporal fusion, irregular acquisition intervals or rapidly changing environments hinder the effective integration of spatial and temporal information. Traditional methods that rely on temporal regularization or statistical models often fall short in capturing the complete dynamics of the scene. Similarly, PolSAR data, characterized by multiple polarimetric channels, present unique challenges for fusion, necessitating careful control of the interaction between spatial and polarimetric information to retain essential details. However, task-specific image fusion approaches require distinct parameter tuning, leading to increased computational and deployment costs. Moreover, they overlook the shared characteristics across multiple tasks, thereby limiting the generalization capability of the algorithm.

Given these challenges, there is a pressing need for a more versatile and generalized spatial-channel fusion framework capable of accommodating a range of fusion tasks within a unified approach. It aims to integrate spatial information with various types of information in the form of channels, collectively denoted as $\boldsymbol{X}$, which may encompass multispectral, hyperspectral, temporal, or polarimetric information. Thus, generalized spatial-channel fusion can be represented as ***Spatial-X fusion***, which encapsulates the core concept of the proposed framework: spatial information serves as the main component, while X represents the diverse modalities that are fused with the spatial information.

In this paper, we propose a generalized spatial-channel fusion framework, *i.e.* **Spa**tial-**X** **Fus**ion, which originates from the model-driven solutions to multiple fusion tasks with spatial-X intrinsic interaction prior (**SpaXFus**). The SpaXFus enables a more precise representation of the interactions between spatial and X information, resulting in superior fusion performance. Additionally, the spatial-X intrinsic interaction

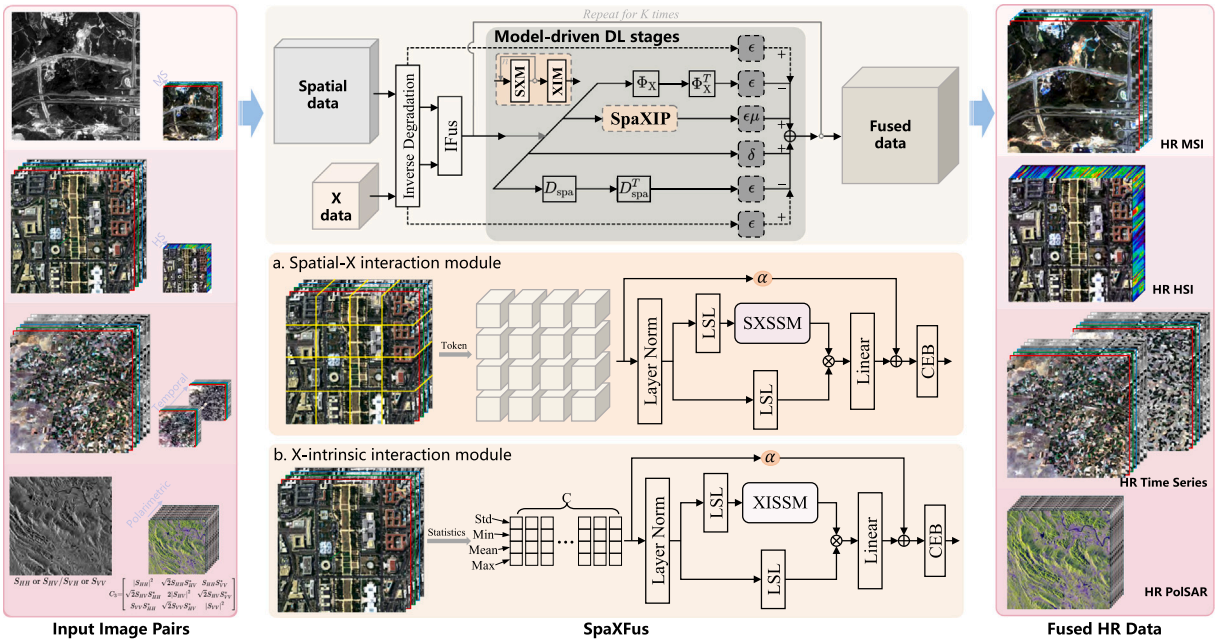

Fig. 1. Overview of SpaXFus: The proposed SpaXFus framework addresses four types of remote sensing image fusion tasks, with its core built upon a model-driven deep learning stage. Specifically, a K-stage optimization-based solution is unfolded into CNNs. At the heart of SpaXFus is the Spatial-X Intrinsic Interaction Prior (SpaXIP), which comprises n Spatial-X Interaction Modules (SXMs) and an X-Intrinsic Interaction Module (XIM). These modules utilize the Spatial-X State-Space Model (SXSSM) and the X-Intrinsic State-Space Model (XISSM) to effectively capture long- and short-range relationships in spatial and X domains.

prior facilitates the integration of information across multiple channels by incorporating two key components: the Spatial-X Interaction Module and the X Intrinsic Interaction Module. The former enables cross-modal learning by integrating spatial data with other modalities, while the latter effectively captures internal relationships within X information. Together, these components enhance fusion performance, leading to more robust and informative representations. The proposed framework enables the integration of spatial information with spectral, temporal, or polarimetric data, improving the quality of the fused imagery, as shown in Fig. 1.

The key contributions of this research are summarized as follows:

- We introduce a novel, generalized spatial-channel fusion framework, Spatial-X fusion, where X involves multispectral, hyperspectral, temporal, and polarimetric information, making it applicable to a wide range of remote sensing applications, including precision agriculture, change detection, and vegetation monitoring.
- The proposed SpaXFus establishes a data-driven unified paradigm for remote sensing image fusion, which addresses the fragmentation of existing image fusion tasks and enhances the model generalization.
- To capture the global dependencies and local interactions among spatial and X domain, the spatial-X intrinsic interaction prior is proposed to effectively explore the internal relationships within spatial-X information and facilitate the adaptive fusion of features with diverse characteristics.
- We demonstrate the versatility of SpaXFus through four spatial-channel fusion tasks, including pan-sharpening, hyperspectral-multispectral fusion, spatio-temporal fusion, and PolSAR fusion, showing its superiority across various datasets.
- For the first time, this work comprehensively explores the impact of image fusion on downstream applications, including vegetation index production generation, fine-grained image classification, change detection, and SAR vegetation extraction.

The rest of this paper is organized as follows. Section 2 provides a review of related works on four types of fusion tasks. Section 3 describes the mathematical formulation of spatial-X fusion. Section 4 introduces the proposed SpaXFus framework. Section 5 presents the experimental results, including comparisons with existing fusion algorithms across four fusion tasks and evaluations of improvements in downstream applications. Finally, Section 6 concludes the paper by summarizing the key findings and discussing potential directions for future research.

2. Related works

2.1. Pan-sharpening

Over the past few decades, many methods have been proposed to achieve pan-sharpening. The main algorithms for traditional pan-sharpening can be divided into four categories:

(1) Component substitution-based methods. Methods based on component substitution aim at replacing the low-resolution spatial component of MS images with PAN images. Moreover, the spatial components are always extracted by methods based on intensity-hue-saturation (IHS) (Loncan et al., 2015), Brovey transformation (Gillespie et al., 1987), principal component analysis (PCA) (Kwarteng and Chavez, 1989), and Gram-Schmidt transformation (GS) (Aiazzi et al., 2007). *(2) Multi-resolution analysis-based methods.* In this type of method, MS and PAN images are decomposed into various resolutions, and the spatial details in PAN images are injected into the same-level MS features. Laplacian pyramids (Burt and Adelson, 1987), curvelet (Aiazzi et al., 2006), wavelets (Ranchin and Wald, 2000), and contourlet transformations (Do and Vetterli, 2005) are some classical decompositions in this group. *(3) Hybrid methods.* These methods try to combine the strengths of both component substitution and multi-resolution analysis methods. The main idea is to improve the spatial details of the fused image at multiple scales. Substitute Wavelet Intensity (SWI) (González-Audícana et al., 2004), Additive Wavelet Luminance Proportional (AWLP) (Otazu et al., 2005), and GS-Wavelet (Javan et al., 2021) are all hybrid methods. *(4) Optimization-based methods.* Considering the spatial and spectral degradation in remote sensing imaging, the variational optimization-based methods regard pan-sharpening as an inverse problem and build different cost functions to search for the best estimation of the ideal

high-resolution MS images. The optimization-based methods that fuse PAN and MS images by different constraints include P+XS (Ballester et al., 2006), Total Variation (TV) (Palsson et al., 2013), $l_{1/2}$ gradient prior (Zeng et al., 2016), filter estimation (Xiao et al., 2023), and local gradient constraints (Fu et al., 2019).

However, these methods struggle to achieve a good balance between spatial blurring and spectral distortions, and are gradually surpassed by deep learning-based algorithms (He et al., 2023; Zhong et al., 2016). Masi et al. (2016) regarded the pan-sharpening task as a particular form of image super-resolution and utilized a three-layer convolutional neural network (PNN) to address pan-sharpening. As the deeper networks achieve a more robust learning ability, residual learning is employed to improve the depth of CNNs and achieve better performance (Shao and Cai, 2018). Wei et al. (2017) introduced a global residual skip to enhance the spatial details. Yang et al. (2017) employed high-pass filters before ResNet to extract better textures. To further improve the modeling capability of CNNs, there have been many works, including pyramid networks (Zhang et al., 2019), adaptive weights (Liu et al., 2020a), attention-based CNNs (Guan and Lam, 2022), the gradient prior (Zhang and Ma, 2021), two-stream networks (Liu et al., 2020c), the deep unrolling (Cao et al., 2022), generative adversarial networks (GANs) (Liu et al., 2020b), and diffusion models (Zhong et al., 2024). Other approaches have focused on improving loss functions and integrating more binding constraints, such as spatial and spectral consistency loss (Luo et al., 2020) and self-attention mechanisms with sparse constraints (Qu et al., 2020). Moreover, some methods have integrated prior knowledge (Ni et al., 2022) and meta-learning (Wang et al., 2022) to enhance performance.

2.2. Hyperspectral-multispectral fusion

Early hyperspectral-multispectral image fusion methods were largely inspired by pan-sharpening techniques and can be classified into three main categories: CS-based approaches (Aiazzi et al., 2002), MRA-based methods (Yokoya et al., 2011), and deep learning-based approaches (Xu et al., 2020a). These methods primarily framed the fusion task as a band assignment problem, which involves determining which high-spatial-resolution MS band should be used to enhance the spatial resolution of a given HS band. In early works on this topic (Zhang and He, 2007), this issue was addressed by manually defining association rules tailored to the specific sensors used in image acquisition. However, no algorithmic solutions were proposed to generalize this process. More recently, the band assignment problem has been explored in depth, leading to more systematic approaches (Picone et al., 2017; Simoes et al., 2014).

Another category of HS-MS fusion methods is based on low-rank approximation, where spectral signatures are assumed to lie in a low-dimensional subspace. This subspace is represented by a matrix or tensor with a rank much lower than the original data dimensions. Algorithms such as vertex component analysis (Wei et al., 2016) and truncated Singular Value Decomposition (SVD) (Wei et al., 2015) are commonly used to learn the spectral basis from HS data. Low-rank tensor models, which exploit local low-rank structures in hyperspectral images, have been further developed through Tucker tensor decomposition (Li et al., 2018a; Lanaras et al., 2015) and Canonical Polyadic tensor decomposition (Xu et al., 2020b). For example, Zhou et al. (2017) applied local low-rank assumptions to perform hyperspectral super-resolution.

Additionally, some HS-MS fusion methods are based on sparse representation. These methods assume that the spectral basis forms an over-complete dictionary, where spectral signatures are represented as linear combinations of a few dictionary atoms, ensuring sparsity (Dong et al., 2016). Techniques such as K-SVD dictionary learning (Li et al., 2018b) are used to construct this dictionary from HS images. The coefficient estimation is regularized using sparse priors, with sparse coding techniques employed to optimize the solution. Studies such as Dian and Li (2019) and Li et al. (2018b) used coupled sparse matrix factorization to achieve high-resolution hyperspectral imaging, while Akhtar et al. (2015) employed Bayesian sparse coding for improved performance. Sparse tensor methods, including non-local sparse tensor approaches (Dian et al., 2024), have also been developed to extend these concepts into the tensor domain.

Recently, deep learning-based techniques have also had a profound impact on HMFusion (Deng et al., 2023). Deep learning-based fusion methods typically aim to learn nonlinear mapping functions between high-resolution target HS images and observed HS and MS image pairs (Xie et al., 2022). Approaches such as CNNs extract spatial and spectral features, facilitating more accurate fusion (Hu et al., 2022). For example, Yang et al. (2018) used a two-branch CNN to capture spatial neighborhood features and spectral information, while Xu et al. (Xu et al., 2020a) enhanced performance using mechanisms like skip connections. Other advanced models include SSR-NET (Zhang et al., 2020) for spatial-spectral reconstruction and deep blind fusion techniques (Jia et al., 2023), which adjust for unknown sensor characteristics, thereby improving image quality across various datasets.

2.3. Spatio-temporal fusion

Spatio-temporal fusion is a critical approach for obtaining high spatio-temporal resolution Earth observation data. Currently, a variety of spatio-temporal fusion methods have been developed, which can be categorized into four main types: spatial weighting, spatial unmixing, hybrid, and learning-based approaches (Wang et al., 2023). Spatial weighting methods, such as STARFM (Gao et al., 2006) and its enhanced variant, ESTARFM (Zhu et al., 2010), utilize spectral, temporal, and spatial data from nearby pixels, with adjustments in methods like Fit-FC (Wang and Atkinson, 2018) and Agri-Fuse to account for seasonal and phenological variations (Gu et al., 2023). Spatial unmixing-based methods, such as the multiresolution technique and the STDFA (Wu et al., 2012), rely on accurate proportion estimation while assuming minimal land cover changes.

Hybrid methods leverage the strengths of both spatial weighting and spatial unmixing techniques. For example, Zhu et al. (2016) proposed FSDAF by integrating the core principles of spatial unmixing and STARFM into a unified framework. Furthermore, Li et al. (2020b) incorporated sub-pixel land cover change information into FSDAF, developing SFSDAF to better address changes in highly heterogeneous regions. Subsequently, FSDAF 2.0 introduced change detection algorithms to enhance its capacity to manage pixels experiencing land cover transitions (Guo et al., 2020).

Learning-based methods establish nonlinear mappings between images of differing resolutions (Huang and Song, 2012). For example, the spatio-temporal temperature fusion network (STTFN) offers significant potential to explicitly model the relationships between multi-source data through nonlinear approaches (Yin et al., 2021). Song et al. (2018) employed deep convolutional neural networks (CNNs) to model the relationship between Landsat and MODIS images, while Liu et al. (2019) developed a two-stream CNN (StfNet) to capture temporal dependencies within image sequences. To address geometric registration issues, Qin et al. (2022) used multiscale features, and Zhang et al. (2024b) introduced an efficient cross-paired wavelet-based network (ECPW-STFN) requiring fewer inputs. Additionally, GAN-based approaches were introduced to resolve reference image selection issues (Chen et al., 2020), and Transformer models, with their capacity for long-range feature extraction, have also been applied to spatio-temporal fusion tasks (Chen et al., 2022).

2.4. PolSAR fusion

PolSAR image fusion seeks to enhance the spatial resolution of PolSAR images by integrating them with HR single-polarimetric SAR (SinSAR) images (or dual-polarimetric SAR). Early methods for enhancing SAR resolution mainly relied on frequency-domain techniques to

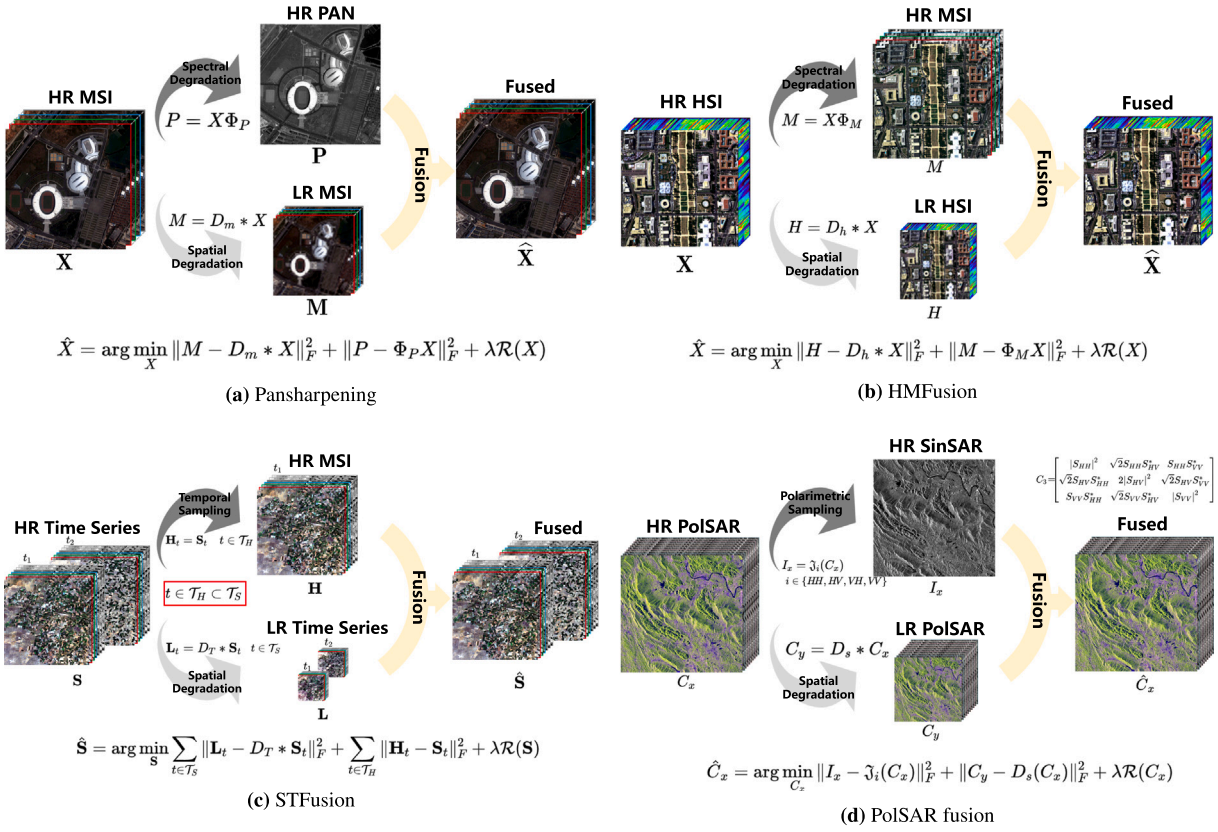

Fig. 2. The problem formulations of four fusion tasks.

improve spatial clarity in PolSAR images. For example, Pastina et al. (2001) first introduced polarimetric information into PolSAR super-resolution through single-channel spectral analysis, while Suwa and Iwamoto (2006) proposed a two-dimensional bandwidth extrapolation technique, extending traditional SAR bandwidth extrapolation to PolSAR images. Although these methods succeeded in enhancing spatial resolution, they did not account for relationships between different polarizations.

Later approaches began utilizing prior image information to further enhance resolution. Jiong and Jian (2007) employed the POCS algorithm to extract information from low-resolution polarimetric SAR channels, generating higher-resolution images through fusion, though this approach compromised original polarimetric and phase information (Zhang et al., 2011). Similarly, Zou et al. (2008) introduced a super-resolution method using target decomposition and quadrant-based pixel weighting to enhance central pixels, though this often introduced grid artifacts. To address this issue, a super-resolution method based on polarimetric spatial correlation was developed, which used pixel-to-pixel polarimetric correlations to initialize subpixel values and iteratively refine them to create a high-resolution PolSAR image (Zhang et al., 2011). However, variations across polarimetric decompositions reduced accuracy in some cases.

Recently, deep learning has shown potential in PolSAR super-resolution, though its application remains limited. The first deep learning-based multichannel PolSAR super-resolution method (MSSR) allowed simultaneous processing of PolSAR channels but did not fully preserve key polarimetric and numerical characteristics (Lin et al., 2019, 2023). To address these limitations, Shen et al. (2020) proposed an approach using complex blocks, transposed convolution, and PReLU to retain these properties. Additionally, Lin et al. (2021a) introduced a fusion network (PSFN) that combines low-resolution PolSAR and high-resolution SinSAR data, later evolving into FDFNet, which incorporates SAR super-resolution and polarimetric decomposition attention to better preserve polarimetric information (Lin et al., 2021b).

3. Spatial-X fusion

Although deep learning algorithms have achieved remarkable progress in multi-source satellite image fusion, their black-box nature and limited generalization significantly constrain the reliability of results in downstream tasks. To address this, this section begins with formulations for four fusion tasks and presents a unified framework for the Spatial-X fusion.

3.1. Formulations of four remote sensing fusion tasks

3.1.1. Pan-sharpening

In pansharpening, the objective is to fuse a high-resolution panchromatic image $\mathbf{P} \in \mathbb{R}^{w\times h\times 1}$ with a low-resolution multispectral image $\mathbf{M} \in \mathbb{R}^{w'\times h'\times c}$ to recover a high-resolution multispectral image $\mathbf{X} \in \mathbb{R}^{w\times h\times c}$ with spatial dimensions w and h, and C spectral bands, where $w' < w$ and $h' < h$. Fig. 2(a) illustrates the degradation and reconstruction of pansharpening task.

The multispectral image $\mathbf{M}$ is generated by applying a spatial degradation operator D_m, which can be modeled as:

$$\mathbf{M} = D_m * \mathbf{X} \tag{1}$$

where $D_m \in \mathbb{R}^{w'h'\times wh}$ represents spatial downsampling in multispectral images, reducing the spatial resolution of $\mathbf{X}$, typically involving downsampling or Gaussian blurring. $*$ is the convolution operator. The degradation ratio r is definited as $r = w/h = w'/h'$, and usually $r = 4$. Similarly, the panchromatic image $\mathbf{P}$ is obtained by applying a spectral transformation Φ_P, which combines the bands of $\mathbf{X}$ into a single

panchromatic band, without spatial degradation:

$$\mathbf{P} = \mathbf{X}\Phi_P \tag{2}$$

where $\Phi_P \in \mathbb{R}^{c\times 1}$ represents the spectral response of the panchromatic sensor, which captures the spectral characteristics of the multispectral image as a weighted average of its original spectral bands.

Pansharpening is to estimate $\mathbf{X}$ from $\mathbf{P}$ and $\mathbf{M}$, yielding an approximation $\hat{\mathbf{X}}$ that best preserves the spatial characteristics from $\mathbf{P}$ and the spectral characteristics from $\mathbf{M}$. Mathematically, the pansharpened image $\hat{\mathbf{X}}$ should satisfy an energy minimization criterion, balancing spatial and spectral fidelity:

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} \|\mathbf{M} - D_m * \mathbf{X}\|_F^2 + \|\mathbf{P} - \Phi_P\mathbf{X}\|_F^2 + \lambda\mathcal{R}(\mathbf{X}) \tag{3}$$

where $\|\cdot\|_F$ is the Frobenius norm and is usually the ℓ_2 norm. $\|\mathbf{P} - \Phi_P\mathbf{X}\|^2$ represents the spectral fidelity to the panchromatic image, ensuring $\hat{\mathbf{X}}$ captures the high spatial detail of $\mathbf{P}$, while $\|\mathbf{M} - D_m * \mathbf{X}\|^2$ maintains spectral consistency with $\mathbf{M}$, ensuring the spectral characteristics are faithfully preserved. $\mathcal{R}(\mathbf{X})$ denotes the regularizer that imposes prior knowledge, $\|\cdot\|^2$ refers to the Euclidean norm of data-fidelity terms, and λ is a trade-off parameter between the regularizer and data-fidelity terms.

3.1.2. Hyperspectral-multispectral fusion

As for HMFusion, let $\mathbf{X} \in \mathbb{R}^{w\times h\times C}$ represent the ideal HR hyperspectral images, with spatial dimensions w and h, and C spectral bands. Observed HR multispectral and LR hyperspectral images are denoted as $\mathbf{M} \in \mathbb{R}^{w\times h\times c}$ and $\mathbf{H} \in \mathbb{R}^{w'\times h'\times c}$, respectively, where $c \ll C$. HMFusion aims at recovering HR hyperspectral images from HR multispectral and LR hyperspectral image pairs, as shown in Fig. 2(b).

The observed LR hyperspectral image $\mathbf{H}$ originates from $\mathbf{X}$ through spatial degradation, given by:

$$\mathbf{H} = D_h * \mathbf{X} \tag{4}$$

where $D_h \in \mathbb{R}^{w'h'\times wh}$ denotes the spatial degradation in hyperspectral images, typically involving downsampling or Gaussian blurring. Generally, the degradation ratio $r = 4$. The HR multispectral image $\mathbf{M}$ results from spectral degradation of $\mathbf{X}$, modeled as:

$$\mathbf{M} = \mathbf{X}\Phi_M \tag{5}$$

where $\Phi_M \in \mathbb{R}^{C\times c}$ is the spectral response matrix of the multispectral sensor, representing spectral downsampling.

The goal of HMFusion is to reconstruct an estimate $\hat{\mathbf{X}}$ of the high-resolution hyperspectral image $\mathbf{X}$, which aligns with both the spatial fidelity of $\mathbf{M}$ and the spectral fidelity of H. This task can be formulated as an optimization problem by minimizing an energy function that ensures both spatial and spectral consistency:

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} \|\mathbf{H} - D_h * \mathbf{X}\|_F^2 + \|\mathbf{M} - \Phi_M\mathbf{X}\|_F^2 + \lambda\mathcal{R}(\mathbf{X}) \tag{6}$$

where the first term, $\|\mathbf{M} - \Phi_M\mathbf{X}\|_F^2$, enforces spatial consistency with the HR multispectral image $\mathbf{M}$, while the second term, $\|\mathbf{H} - D_h * \mathbf{X}\|_F^2$, maintains spectral consistency with the LR hyperspectral image $\mathbf{H}$. This framework enables the recovery of an accurate HR hyperspectral image $\hat{\mathbf{X}}$ that captures both high spatial resolution and rich spectral information.

3.1.3. Spatio-temporal fusion

In spatio-temporal fusion, the goal is to recover HR time series data at all times from a limited set of LR image time series and sparse HR observations, as shown in Fig. 2(c). We assume that the LR time series data $\mathbf{L}_t$ are degraded from HR time series data $\mathbf{S}_t$ via spatial downsampling in the time sequence $\mathcal{T}_S$. Specifically, the LR image is obtained by applying spatial degradation D_T, such that:

$$\mathbf{L}_t = D_T * \mathbf{S}_t \tag{7}$$

Additionally, the HR observations $\mathbf{H}_t$ are available at specific time points $t \in \mathcal{T}_H \subset \mathcal{T}_S$, where the high-resolution image is directly observed, i.e., $\mathbf{H}_t = \mathbf{S}_t$ for $t \in \mathcal{T}_H$.

Given the degradation model, our objective is to reconstruct the high-resolution images $\mathbf{S}_t$ for all time $t \in \mathcal{T}_S$, including both the observed and unobserved time points. This is formulated as an optimization problem, where the goal is to minimize the error between the reconstructed images and the observed data, subject to spatio-temporal consistency constraints:

$$\hat{\mathbf{S}} = \arg\min_{\mathbf{S}} \sum_{t\in\mathcal{T}_S} \|\mathbf{L}_t - D_T * \mathbf{S}_t\|_F^2 + \sum_{t\in\mathcal{T}_H} \|\mathbf{H}_t - \mathbf{S}_t\|_F^2 + \lambda\mathcal{R}(\mathbf{S}) \tag{8}$$

Here, the first term ensures temporal consistency between the reconstructed and LR images, the second term enforces consistency with the HR observations, and the third term $\mathcal{R}(\mathbf{S})$ is a regularization function that promotes temporal and spatial priors in the solution. The weight λ balances the contributions of the data fidelity and regularization terms.

3.1.4. PolSAR fusion

In PolSAR fusion tasks, HR PolSAR data contain rich spatial and polarimetric information, typically represented by a scattering matrix S and covariance matrix C_x, as shown in Fig. 2(d). The S matrix for PolSAR data is:

$$S = \begin{bmatrix} S_{HH} & S_{HV} \\ S_{VH} & S_{VV} \end{bmatrix} \tag{9}$$

where S_{HH} and S_{VV} denote co-polarized channels (horizontal-horizontal and vertical-vertical), and S_{HV} and S_{VH} represent cross-polarized channels (horizontal-vertical and vertical-horizontal). In contrast, SinSAR data capture only one of these channels, providing limited polarimetric information compared to PolSAR data. When the reciprocity condition holds and system noise is disregarded, the backscattering matrix can be transformed into a Lexicographic covariance matrix C_3 (Lee and Pottier, 2017):

$$C_3 = \begin{bmatrix} |S_{HH}|^2 & S_{HH}S_{HV}^* & S_{HH}S_{VV}^* \\ S_{HV}S_{HH}^* & |S_{HV}|^2 & S_{HV}S_{VV}^* \\ S_{VV}S_{HH}^* & S_{VV}S_{HV}^* & |S_{VV}|^2 \end{bmatrix} \tag{10}$$

where $*$ denotes the complex conjugate. To facilitate analysis and computation, C_3 is often vectorized into a real-valued 1×9 vector, C_{value}, by separating each matrix element into its real and imaginary components:

$$C_{\text{value}} = [R_{11}, R_{12}, I_{12}, R_{13}, I_{13}, R_{22}, R_{23}, I_{23}, R_{33}] \tag{11}$$

where R_{ij} and I_{ij} represent the real and imaginary parts of the elements in C_3, respectively. Thus, $C_x \in \mathbb{R}^{w\times h\times 9}$.

Due to system limitations or operational constraints, a scene of HR PolSAR data C_x will degenerate into the observational LR PolSAR images $C_y \in \mathbb{R}^{w'\times h'\times 9}$. Among them, both HR PolSAR image and LR PolSAR image are three-dimensional matrices composed of C_{value}, which represents the value of one pixel of C_x and C_y. Let D_s denote the spatial downsampling operator. Thus, the relationship between HR and LR PolSAR data can be expressed as:

$$C_y = D_s(C_x) \tag{12}$$

Since HR SinSAR data is a specific polarimetric channel extracted from C_x through a mapping $\mathfrak{J}_i(C_x)$, where i represents a chosen polarimetric channel (e.g., HH or HV/VH or VV):

$$I_x = \mathfrak{J}_i(C_x), \quad i \in \{HH, HV, VH, VV\} \tag{13}$$

where $I_x \in \mathbb{R}^{w\times h\times 1}$ denotes the observed HR SinSAR data. To reconstruct HR PolSAR data C_x from the available observations I_x and C_y,

the PolSAR fusion can be built as an optimization model:

$$\hat{C}_x = \arg\min_{C_x} \|I_x - \Im_i(C_x)\|_F^2 + \|C_y - D_s(C_x)\|_F^2 + \lambda\mathcal{R}(C_x) \tag{14}$$

where $\|\Im_i(C_x) - I_x\|_F^2$ ensures consistency with the HR spatial details, $\|D_s(C_x) - C_y\|_F^2$ enforces agreement with the fully polarimetric information, $\mathcal{R}(C_x)$ is a regularization term promoting spatial and polarimetric priors, and λ is a weighting factor balancing observation fidelity with regularization.

3.2. Definition of spatial-X fusion

In this paper, we aim to propose a generalized fusion framework for these tasks involving spatial and X degradation. Generally, spatial degradation includes all downsampling operations related to spatial resolution, and we can unify them into a single spatial downsampling operation matrix D_{spa}. Thus, the unified expression for spatial degradation can be written as:

$$\mathbf{A} = D_{\text{spa}} * \mathbf{X} \tag{15}$$

where $\mathbf{A}$ denotes the observed LR remote sensing image with C_X channels, $\mathbf{X}$ means the ideal HR remote sensing image, and D_{spa} is a unified matrix representing spatial downsampling, and its specific form depends on the task. In different tasks, spatial downsampling can involve different downsampling operations, such as D_m in pan-sharpening, D_h in HMFusion, D_T in STFusion, and D_s in PolSAR fusion.

Besides the spatial information, different remote sensing images tend to provide various types of information, including multispectral, hyperspectral, temporal, and polarimetric information. To build a universal framework, we propose a novel concept called X information, which encompasses all the information in remote sensing images except the spatial domain. Thus, X degradation is proposed as:

$$\mathbf{B} = \mathbf{X}\Phi_{\text{X}} \tag{16}$$

where $\mathbf{B}$ denotes the observed HR remote sensing image with poor X information (with $c_{\mathbf{B}}$ channels), Φ_{X} represents the X degradation with different degradation operations for various tasks.

In pan-sharpening, Φ_{X} is in the form of $\Phi_P \in \mathbb{R}^{c\times 1}$, which means the spectral response functions of panchromatic sensors, degenerate the multispectral image to a panchromatic image.

In HMFusion, Φ_{X} equals $\Phi_m \in \mathbb{R}^{C\times c}$, which means the spectral transformation from hyperspectral domain into multispectral domain, degenerating the hyperspectral image to a multispectral image.

In ST fusion, Φ_{X} can be regarded as $\Phi_t \in \mathbb{R}^{\mathcal{T}_S\times\mathcal{T}_H}$, which is defined as a sampling matrix, selecting specific remote sensing images in time phases $\mathcal{T}_H$ from time set $\mathcal{T}_S$. For instance, the time set $\mathcal{T}_S = t_1, t_2, t_3$ and we select t_1, t_2 as $\mathcal{T}_H$. The Φ_t is formulated as:

$$\Phi_t = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \tag{17}$$

In PolSAR fusion, $\Phi_{\text{X}} = \Phi_s \in \mathbb{R}^{4\times 1}$, which can also be regarded as a sampling matrix, choosing the different polarimetric channels for different applications. If HH mode is chosen, Φ_s can be written as:

$$\Phi_s = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{18}$$

With spatial degradation and X degradation, spatial-X fusion can be defined as integrating HR spatial details in $\mathbf{B}$ with fine X information $\mathbf{A}$ and recovering HR X information $\mathbf{X}$, which is an optimization problem:

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} \|\mathbf{A} - D_{\text{spa}} * \mathbf{X}\|_F^2 + \|\mathbf{B} - \mathbf{X}\Phi_{\text{X}}\|_F^2 + \lambda\mathcal{R}(\mathbf{X}) \tag{19}$$

where $\mathcal{R}(\mathbf{X})$ is a regularizer that imposes prior knowledge, $\|\cdot\|_F$ represents the Frobenius norm of data-fidelity terms, $F = 2$ in this paper, and λ is a trade-off parameter between the regularizer and data-fidelity terms.

4. Methodology

In this part, we first derive the iterative solution to the spatial-X fusion problem based on optimization model. Then, we leverage a model-driven solution flow to guide the construction of the deep learning network, and finally introduce the spatial-X intrinsic interaction prior to explore the latent knowledge within the spatial-X domain.

4.1. The optimization-based solution

While the energy function ensures consistency between the reconstructed and observed images and incorporates image priors, solving this optimization problem directly is challenging due to the various priors involved. To decouple the regularization and data-fidelity terms in Eq. (19) and facilitate minimization, we apply a variable-splitting technique by introducing an auxiliary variable $\mathbf{Z}$, thus reformulating the optimization problem with a constraint $\mathbf{Z} = \mathbf{X}$:

$$\begin{aligned} \hat{\mathbf{X}} = \arg\min_{\mathbf{X}} &\frac{1}{2}\left\|\mathbf{A} - D_{\text{spa}}\mathbf{X}\right\|_2^2 + \frac{1}{2}\left\|\mathbf{B} - \mathbf{X}\Phi_{\text{X}}\right\|_2^2 + \lambda\mathcal{R}(\mathbf{Z}) \\ &\text{s.t. } \mathbf{Z} = \mathbf{X} \end{aligned} \tag{20}$$

Applying the half-quadratic splitting method yields a modified cost function:

$$\begin{aligned} \mathcal{L}_\mu(\mathbf{X}, \mathbf{Z}) = &\frac{1}{2}\left\|\mathbf{A} - D_{\text{spa}}\mathbf{X}\right\|_2^2 + \frac{1}{2}\left\|\mathbf{B} - \mathbf{X}\Phi_{\text{X}}\right\|_2^2 \\ &+ \frac{\mu}{2}\|\mathbf{Z} - \mathbf{X}\|_2^2 + \lambda\mathcal{R}(\mathbf{Z}) \end{aligned} \tag{21}$$

where μ serves as a penalty parameter. This reformulation allows us to decompose the original optimization problem into two subproblems that can be solved independently and efficiently,

$$\begin{cases} \hat{\mathbf{X}} = \arg\min_{\mathbf{X}} \frac{1}{2}\left\|\mathbf{A} - D_{\text{spa}}\mathbf{X}\right\|_2^2 + \frac{1}{2}\left\|\mathbf{B} - \mathbf{X}\Phi_{\text{X}}\right\|_2^2 + \frac{\mu}{2}\|\mathbf{Z} - \mathbf{X}\|_2^2 \\ \hat{\mathbf{Z}} = \arg\min_{\mathbf{Z}} \frac{1}{2}\|Z - \mathbf{X}\|_2^2 + \frac{\lambda}{\mu}\mathcal{R}(\mathbf{Z}) \end{cases} \tag{22}$$

In the $\mathbf{X}$-subproblem, an approximate solution is iteratively updated using the gradient descent algorithm:

$$\begin{aligned} \hat{\mathbf{X}}_{k+1} &= \mathbf{X}_k - \epsilon(\mathbf{X}_k\Phi_{\text{X}}\Phi_{\text{X}}^T + D_{\text{spa}}^T D_{\text{spa}}\mathbf{X}_k - \mathbf{B}\Phi_{\text{X}}^T - D_{\text{spa}}^T\mathbf{A} + \mu\mathbf{X}_k - \mu\mathbf{Z}_k) \\ &= \delta\mathbf{X}_k - \epsilon\mathbf{X}_k\Phi_{\text{X}}\Phi_{\text{X}}^T - \epsilon D_{\text{spa}}^T D_{\text{spa}}\mathbf{X}_k + \epsilon\mathbf{B}\Phi_{\text{X}}^T + \epsilon D_{\text{spa}}^T\mathbf{A} + \epsilon\mu\mathbf{Z}_k \end{aligned} \tag{23}$$

where ϵ is the optimization stride, and $\delta = 1 - \epsilon\mu$. The $\mathbf{Z}$-subproblem, which incorporates prior knowledge, is addressed using proximal operators:

$$\hat{\mathbf{Z}}_k = \text{Prox}(\mathbf{X}_k) = \arg\min_{\mathbf{Z}} \frac{1}{2}\left\|\mathbf{Z} - \mathbf{X}_k\right\|_2^2 + \frac{\lambda}{\mu}\mathcal{R}(\mathbf{Z}) \tag{24}$$

By iteratively updating these two subproblems, we can efficiently solve Eq. (19) via alternating updates for $\mathbf{X}$ and $\mathbf{Z}$.

4.2. Model-driven deep learning architecture

Building on Eqs. (23) and (24), we propose a model-driven deep learning architecture for spatial-X fusion, where all operators are unrolled into convolutional layers. The architecture comprises K stages, corresponding to K iterations in the optimization-based solution. At each stage, the proposed SpaXFus framework takes the HR data $\mathbf{B}$, LR data $\mathbf{A}$, and the previous output $\mathbf{X}_k$ as inputs, generating $\mathbf{X}_{k+1}$ as the updated output.

Within SpaXFus, end-to-end CNNs are leveraged to learn six key components derived from Eq. (23): the memory term $\delta\mathbf{X}_k$, which preserves

information from previous iterations and enables SpaXFus to progressively optimize the output by integrating multi-stage information; the degradation and restoration terms within the spatial and X domains, which align with the data flow of the optimization-based solution and handle the degradation and subsequent restoration; the reconstruction terms derived directly from inputs **A** and **B**, which ensure fidelity to the original data; and the implicit image priors embedded in the update of $\mathbf{Z}_k$, which provide regularization and consistency. These components are integrated to update $\mathbf{X}_{k+1}$, following Eq. (29), where $\mathbf{X}_{k+1}$ is computed as a combination of memory, degradation, restoration, and prior terms:

$$\mathbf{M}_k = \delta \cdot \mathbf{X}_k \tag{25}$$

$$\mathbf{C}_k = \epsilon \cdot \mathbf{X}_k \Phi_{\mathrm{X}} \Phi_{\mathrm{X}}^T + \epsilon \cdot D_{\mathrm{spa}}^T D_{\mathrm{spa}} \mathbf{X}_k \tag{26}$$

$$\mathbf{R}_k = \epsilon \cdot \mathbf{B} \Phi_{\mathrm{X}}^T + \epsilon \cdot D_{\mathrm{spa}}^T \mathbf{A} \tag{27}$$

$$\mathbf{E}_k = \epsilon\mu \cdot \mathrm{Prox}(\mathbf{X}_k) \tag{28}$$

$$\mathbf{X}_{k+1} = \mathbf{M}_k - \mathbf{C}_k + \mathbf{R}_k + \mathbf{E}_k \tag{29}$$

where $\mathbf{M}_k$ incorporates memory, $\mathbf{C}_k$ and $\mathbf{R}_k$ model degradation and restoration, and $\mathbf{E}_k$ captures implicit priors.

Following model-driven approach, the network architecture reflects the outlined data flow as shown in Fig. 1. In each stage, channel attention mechanisms are employed to adaptively learn hyperparameters δ, ϵ, and μ, allowing for stage-specific and channel-sensitive weight adjustments. Firstly, the memory term $\mathbf{M}_k$ is transformed into a learnable module:

$$\mathbf{M}_k = \mathrm{CAM}_\delta(\mathbf{X}_k), \tag{30}$$

where CAM_δ denotes a channel attention module responsible for parameter δ.

In Eqs. (26) and (27), Φ_{X} represents the degradation operator for X-domain information. This operator is implemented as a point-wise convolution (1×1 convolution) in SpaXFus, with Φ_{X}^T realized as its inverse operation (a $c_{\mathbf{B}} \times C_{\mathbf{X}}$ point-wise convolution). For spatial degradation D_{spa}, stride convolution is employed to approximate the image degradation process, eliminating the need for predefined point spread functions. The inverse degradation, D_{spa}^T, is modeled using a learnable deconvolution operation:

$$\begin{aligned}\mathbf{C}_k = {} & \mathrm{CAM}_\epsilon\left(\mathrm{PConv}^{-1}\left(\mathrm{PConv}\left(\mathbf{X}_k\right)\right)\right) \\ & + \mathrm{CAM}_\epsilon\left(\mathrm{DConv}\left(\mathrm{SConv}\left(\mathbf{X}_k\right)\right)\right)\end{aligned} \tag{31}$$

$$\mathbf{R}_k = \mathrm{CAM}_\epsilon\left(\mathrm{PConv}^{-1}\left(\mathbf{B}\right)\right) + \mathrm{CAM}_\epsilon\left(\mathrm{DConv}\left(\mathbf{A}\right)\right) \tag{32}$$

where PConv and PConv^{-1} are point-wise convolution and its inverse, SConv is stride convolution for spatial downsampling, and DConv is the corresponding deconvolution.

To address the proximal operator in Eq. (28), we introduce the Spatial-X Intrinsic Interaction Prior (SpaXIP). SpaXIP captures both local-global spatial dependencies and intrinsic X-domain interactions, enriching the modeling of implicit image priors:

$$\mathbf{E}_k = \mathrm{CAM}_{\epsilon\mu}\left(\mathrm{SpaXIP}(\mathbf{X}_k)\right), \tag{33}$$

where SpaXIP represents the spatial-X intrinsic interaction network.

Before the first stage, initial outputs $\mathbf{X}_1^A$ and $\mathbf{X}_1^B$ are reconstructed from **A** and **B** through inverse degradation streams:

$$\mathbf{X}_1^A = \mathrm{DConv}\left(\mathbf{A}\right) \tag{34}$$

$$\mathbf{X}_1^B = \mathrm{PConv}^{-1}\left(\mathbf{B}\right) \tag{35}$$

The LR data **A** is upsampled to match the spatial resolution of **B**, while **B** is expanded to the channel dimension of **A**. These initial reconstructions are fused to initialize $\mathbf{X}_1$:

$$\mathbf{X}_1 = \mathrm{IFus}\left(\mathbf{X}_1^{\mathbf{A}}, \mathbf{X}_1^{\mathbf{B}}\right) \tag{36}$$

where IFus is a weighted fusion module for fusing multi-source information.

4.3. Spatial-X intrinsic interaction prior

In image fusion tasks, it is crucial to extract as much spatial and X information from the images as possible and capture the relationships between them. As stated in Tobler's First Law of Geography, "everything is related to everything else" (Tobler, 1970), highlighting the interconnectedness of all phenomena on Earth. Accordingly, effectively reconstructing HR remote sensing data from low-quality images requires leveraging both global dependencies and local interactions. With the advancement of deep learning, numerous algorithms have emerged to capture internal interactions, including attention mechanisms and transformers. Among these, the structured state space model (SSM) has garnered widespread attention from researchers due to its accurate learning capabilities and efficient computational speed (Gu et al., 2022).

4.3.1. Preliminaries

The SSM draws inspiration from continuous dynamical systems. This approach models sequences $x(t)$ by projecting them through a hidden state $h(t) \in \mathbb{R}^N$, defined by:

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t), \\ y(t) &= Ch(t) + Dx(t), \end{aligned} \tag{37}$$

where A controls state transitions, B and C handle input and output projections, respectively, and D represents the memory weight of the previous state.

To adapt the model for discrete settings, a timescale parameter Δ is introduced. Using zero-order hold assumption, the continuous matrices A and B are converted into discrete forms $\overline{A}$ and $\overline{B}$:

$$\begin{aligned} \overline{A} &= \exp(\Delta A), \\ \overline{B} &= (\Delta A)^{-1}(\exp(\Delta A) - I)\Delta B. \end{aligned} \tag{38}$$

The system then operates in discrete time, with:

$$\begin{aligned} h_t &= \overline{A}h_{t-1} + \overline{B}x_t, \\ y_t &= Ch_t + Dx_t, \end{aligned} \tag{39}$$

where Δ controls step size for temporal modeling.

Finally, Structured State Space for Sequence Models (S4) calculates outputs using a convolutional operation:

$$\begin{aligned} K &= \left(CB, C\overline{A}B, \ldots, C\overline{A}^{N-1}B\right), \\ y &= x * K, \end{aligned} \tag{40}$$

where K represents the convolution kernel. By combining state-space dynamics and signal processing techniques, S4 efficiently models long-range dependencies in sequences (Guo et al., 2025).

4.3.2. Overall architecture

In SpaXFus, SpaXIP is employed to explore the latent prior knowledge in **X** following Eq. (33). The overview of SpaXIP is shown in Fig. 3, consisting of global dependency modeling and channel interaction enhancement.

In SpaXIP, we proposed a spatial-X interaction module to explore the global relationship across both spatial and X domains, encompassing shallow feature extraction and spatial-X SSM. Furthermore, the statistical characteristics of different channels were analyzed in X-intrinsic interaction module, thereby enhancing local interactions among channels. To address the challenge of fully capturing diverse types of X information, a channel enhancement block was incorporated after each module, ensuring that the network retains and reinforces X information in deeper layers.

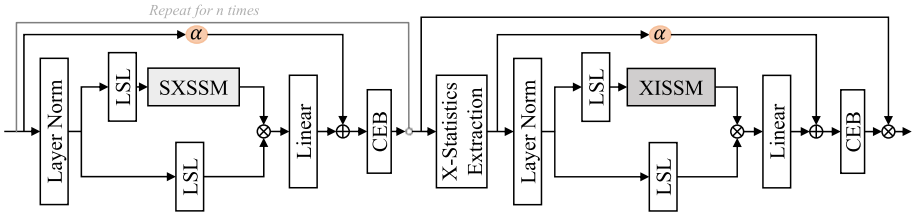

Fig. 3. Overview of spatial-X intrinsic interaction network.

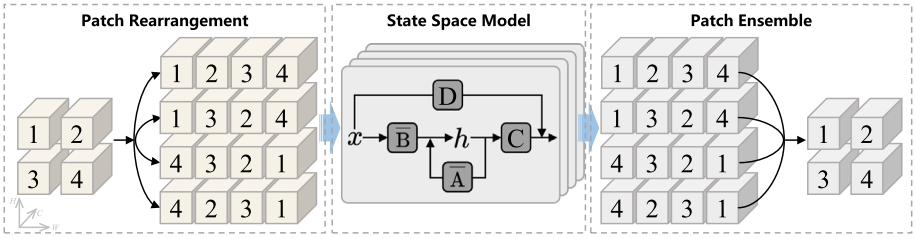

Fig. 4. The framework of spatial-X state space model, where A controls state transitions, B and C handle input and output projections, respectively, D presents the memory weight of the previous state, and arrow shows the dataflow.

4.3.3. Spatial-X interaction module

As illustrated in Fig. 3, there are n spatial-X interaction modules in SpaXIP. At each model-driven DL stage, given the updated $\mathbf{X}_k$, we first apply a LayerNorm layer to enhance the model's stability. Subsequently, two branches, consisting of the Linear layer followed by SiLU activation function, are used to extract deeper features.

$$\begin{aligned}\mathbf{F}_k^{l1} &= \mathrm{LSL}(\mathrm{LN}(\mathbf{F}_k^{l-1}))\\ \mathbf{F}_k^{l2} &= \mathrm{LSL}(\mathrm{LN}(\mathbf{F}_k^{l-1}))\end{aligned} \tag{41}$$

where $k \le K$ denotes the k-th model-driven DL stage, $l \le n$ denotes the l-th spatial-X interaction module, $\mathbf{F}_k^{l-1}$ is the output from the previous spatial-X interaction module, and $\mathbf{F}_k^0$ corresponds to $\mathbf{X}_k$. The feature $\mathbf{F}_k^{l1}$ is then fed into SXSSM to capture global dependencies across both spatial and channel domains.

$$\mathbf{W}_k^l = \mathrm{SXSSM}(\mathbf{F}_k^{l1}) \tag{42}$$

The architecture of SXSSM is illustrated in Fig. 4. In this part, the input features are cropped into patches and rearranged into multiple sequences using four distinct scanning patterns. The global dependencies within each sequence are computed based on Eqs. (38) and (39) in the SSM. Finally, the global dependency weights are aggregated using a patch ensemble approach.

The global dependency weights $\mathbf{W}_k^l$ are used to refine the spatial-X features $\mathbf{F}_k^{l2}$ via a Linear layer and selective memory connection:

$$\mathbf{F}_k^{l3} = \mathrm{Linear}(\mathbf{W}_k^l \circ \mathbf{F}_k^{l2}) + \alpha \mathbf{F}_k^{l-1} \tag{43}$$

where α represents a learnable parameter and $\circ$ denotes Hadamard product. To further enhance the model's ability to capture X information, a Channel Enhance Block (CEB) is applied:

$$\mathbf{F}_k^l = \mathrm{CEB}(\mathbf{F}_k^{l3}) \tag{44}$$

As a result, we obtain the output of the l-th spatial-X interaction module, $\mathbf{F}_k^l$. The CEB is composed of a residual channel attention, integrated with a LayerNorm layer.

4.3.4. X-intrinsic interaction module

In addition to capturing global dependencies across spatial and X domains, X-intrinsic interaction plays a crucial role in remote sensing image fusion. To address this, we introduce an X-intrinsic interaction module at the end of the SpaXIP framework.

Initially, we extract statistical features such as the maximum, average, minimum, and standard deviation. These features effectively characterize the data distribution across different channels, facilitating the exploration of X-intrinsic interactions. The extracted features are then fed into further feature extraction processes.

$$\begin{aligned}\mathbf{J}_k^1 &= \mathrm{LSL}(\mathrm{LN}(\mathrm{XSE}(\mathbf{F}_k^n)))\\ \mathbf{J}_k^2 &= \mathrm{LSL}(\mathrm{LN}(\mathrm{XSE}(\mathbf{F}_k^n)))\end{aligned} \tag{45}$$

where $\mathbf{F}_k^n$ is the output of the final spatial-X interaction module. The feature $\mathbf{J}_k^1$ is then passed into XISSM to capture interactions among channels.

$$\mathbf{W}_k^{\mathrm{X}} = \mathrm{XISSM}(\mathbf{J}_k^1) \tag{46}$$

The architecture of XISSM is shown in Fig. 5. In this part, similar channels are grouped, and the channel order is rearranged to form multiple sequences through channel permutation. The X-intrinsic interactions of each sequence are calculated using SSM. Finally, the multiple X-intrinsic interactions are integrated in channel ensemble.

Then, X-intrinsic interactions $\mathbf{W}_k^{\mathrm{X}}$ are applied to $\mathbf{J}_k^2$, followed by a Linear layer and selective memory connection, before being passed into CEB:

$$\mathbf{J}_k = \mathrm{CEB}(\mathrm{Linear}(\mathbf{W}_k^{\mathrm{X}} \circ \mathbf{J}_k^2) + \alpha \mathrm{XSE}(\mathbf{F}_k^n)) \tag{47}$$

Finally, the fusion of results from both SXM and XIM yields the final output, $\mathbf{Z}_k$, incorporating prior knowledge:

$$\mathbf{Z}_k = \mathrm{Prok}(\mathbf{X}_k) = \mathrm{Conv}(\mathbf{J}_k) \circ \mathbf{F}_k^n \tag{48}$$

where $\mathrm{Conv}(\cdot)$ represents a 1D convolutional layer, which ensembles $\mathbf{J}_k \in \mathbb{R}^{C\times 4}$ into $\mathbb{R}^{C\times 1}$.

5. Experiments

In this section, the proposed SpaXFus method is applied to four typical remote sensing spatial-channel fusion tasks, including pan-sharpening, HMFusion, STFusion, and PolSAR fusion. Existing performance evaluations are often conducted by calculating quantitative

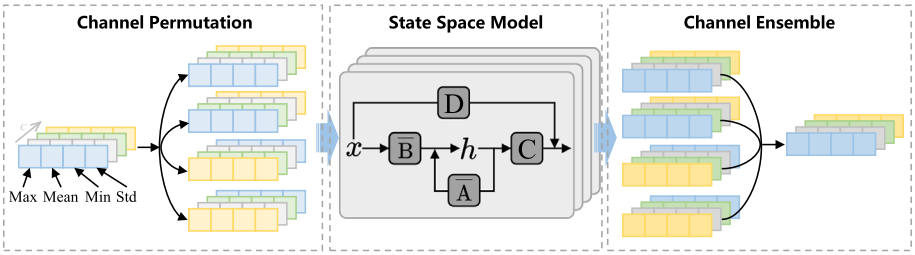

Fig. 5. The framework of X-intrinsic state space model.

metrics or through visual comparisons. However, considering that image fusion tasks are ultimately designed to serve downstream applications in real-world scenarios, this study introduces four corresponding downstream applications for each fusion task to verify whether the algorithms can genuinely integrate more useful information, including vegetation index product generation, fine-grained image classification, change detection, and SAR vegetation extraction, respectively. All quantitative results and visual comparisons are provided in the supplementary materials (Tables S1–S3 and Figures S1–S4).

5.1. Datasets

In this study, methods in each fusion task are evaluated using two datasets to assess the performance differences across different sensors.

Pan-sharpening: *(1) QB:* The QuickBird satellite captures a panchromatic (PAN) channel spanning 450–900 nm, alongside a multispectral image (MSI) comprising four bands within the visible to near-infrared (NIR) spectrum. The PAN channel has a spatial resolution of 61 cm, while the MS channels have a resolution of 2.44 m. The QB dataset used in this study consists of PAN images divided into 5120 patches of size 256×256, along with their corresponding MS patches. *(2) Gaofen2:* The Gaofen-2 satellite, equipped with dual PAN/MS cameras, collects images with a resolution of 0.81 m in the PAN channel and 3.24 m in four MS channels. The GF2 dataset includes 4122 patches, each sized 256×256, along with their corresponding MS patches.

HMFusion: *(1) CAVE:* The CAVE dataset,[1] consisting of 32 scenes (512×512 each), is widely used in HSI processing. It provides hyperspectral images (HSIs) spanning 400–700 nm in 31 bands with a spectral resolution of 10 nm, along with corresponding RGB images. Six scenes are used for testing, while the rest are used for training. Training data includes RGB patches cropped to 128×128 with a stride of 96, and corresponding HS patches downsampled by a factor of four. *(2) Sen2Chikusei:* The Chikusei dataset,[2] captured by the Headwall HyperspecVNIR-C sensor, covers agricultural and urban areas in Chikusei, Japan, with dimensions of 2517×2335. It contains 128 spectral bands (363–1018 nm) at a spatial resolution of 2.5 m. The image is divided into 2384 patches of size 128×128 as ground truth (GT), downsampled by four to create input LRHS data. Sentinel-2 data is simulated using Hysure (Simoes et al., 2014) on the Chikusei dataset.

STFusion: *(1) Land4Mod:* We use the AHB data proposed by Li et al. (2020a). All the high-resolution images (Landsat images with 30 m spatial resolution) are acquired by Landsat-8 OLI with six bands, including blue, green, red, NIR, short-wave infrared-1, and short-wave infrared-2. The low-resolution images (MODIS images with 500 m spatial resolution) are geometrically transformed based on the corresponding Landsat images. All Landsat images are upsampled to 480 m spatial resolution and cropped into 2340 patches of size 256×256, so the resolution ratio is 16 in this dataset. *(2) Sen2Pla:* DynamicEarthNet (Toker et al., 2022) is a daily multi-spectral satellite dataset for semantic change segmentation, consisting of monthly 10 m Sentinel-2 images and daily 3 m PlanetFusion images.[3] We selected 100 time pairs in 50 scenes with high-quality Sentinel-2 images to build the Sen2Pla dataset. All Sentinel-2 images are upsampled to 9 m, and PlanetFusion images are cropped into 5302 patches of size 96×96.

PolSAR Fusion: The training samples for PolSAR fusion are obtained from RadarSat-2 operating in high-resolution mode (8 m), comprising 23,232 HR PolSAR samples, each with a size of 40×40, along with their corresponding SinSAR. The HR PolSAR samples are subsequently downsampled by a factor of two to simulate the input LR PolSAR data. *(1) San Francisco:* A RadarSat-2 scene covering San Francisco, with a spatial size of 2400×2400, is selected, and the simulation process is applied to generate testing data. *(2) Quebec:* The SAR data covering Quebec is real-world data acquired in both standard mode (25 m) and high-resolution mode. The standard-mode data are upsampled to 16 m to maintain a resolution ratio of two.

For Pan-sharpening, Sen2Chikusei, and STFusion datasets, 90 % of them are allocated for training, and the remaining are used for testing.

5.2. Comparison methods

In pan-sharpening task, we select seven traditional methods for comparison, including BDSD (Garzelli et al., 2007), Adaptive Component Substitution with Partial Replacement (PRACS) (Choi et al., 2010), GSA (Aiazzi et al., 2007), ATWT-M3 (Ranchin and Wald, 2000), MTF-GLP-HPM (Aiazzi et al., 2006), AWLP (Otazu et al., 2005), and TV (Palsson et al., 2013), involving CS-based, MRA-based and VO-based algorithms. As for deep learning-based algorithms, PanNet (Yang et al., 2017), DRPNN (Wei et al., 2017), MSDCNN (Yuan et al., 2018), TFResNet (Liu et al., 2020c), TFMamba, SSDiff (Zhong et al., 2024), and MambaIR (Guo et al., 2025) are compared. TFMamba is an improved version that applies the Mamba module (Guo et al., 2025) to TFResNet to demonstrate the effectiveness of SSM. SSDiff as a diffusion-based method and MambaIR as an SSM-based method are two state-of-the-art algorithms.

In HMFusion, Hysure (Simoes et al., 2014), CNMF (Yokoya et al., 2011), FUSE (Wei et al., 2015), CSU (Lanaras et al., 2015), CSTF (Li et al., 2018b), NSSR (Dong et al., 2016), LTMR (Dian and Li, 2019), GTNN (Dian et al., 2024), PSRT (Deng et al., 2023), MSST (Jia et al., 2023), SSRNET (Zhang et al., 2020), Fusformer (Hu et al., 2022), PNXnet (He et al., 2022) and TFMamba are selected as comparison methods.

In STFusion, STARFM (Gao et al., 2006), FSDAF (Zhu et al., 2016), Fit-FC (Wang and Atkinson, 2018), StfNet (Liu et al., 2019), STTFN (Yin et al., 2021), MUSTFN (Qin et al., 2022), ECPW-STFN (Zhang et al., 2024b), and TFMamba are used to verify the superiority of the proposed SpaXFus. The first three are traditional algorithms and the remaining are based on deep learning.

In PolSAR fusion, we compared our SpaXFus with Bicubic, SRPSC (Zhang et al., 2011), MSSR, PSSR (Shen et al., 2020), MSPSRN (Lin et al., 2023), PSFN (Lin et al., 2022), and FDFNet (Lin et al., 2021b). Bicubic

[1] http://www.cs.columbia.edu/CAVE/databases/.
[2] https://naotoyokoya.com/Download.html.
[3] https://www.planet.com/pulse/planet-announces-powerful-new-products-at-planet-explore-2020/.

Table 1
Quantitative comparisons of pan-sharpening on both QuickBird and Gaofen2 datasets. The best is in **bold**, and the second best is underlined.

Method	QuickBird					Gaofen2				
	CC	mPSNR	mSSIM	SAM	ERGAS	CC	mPSNR	mSSIM	SAM	ERGAS
BDSD	0.9351	41.8275	0.9431	2.2678	2.0736	0.9522	37.7822	0.9374	2.2495	2.6582
PRACS	0.9511	44.2434	0.9566	1.8110	1.4983	0.9186	38.2479	0.8999	2.1987	2.3423
GSA	0.9389	42.2266	0.9444	2.0716	1.9624	0.9557	38.5998	0.9338	2.0837	2.3930
ATWT-M3	0.9355	42.6250	0.9392	2.3004	1.8965	0.9650	41.6236	0.9580	2.2792	1.9615
MTF-GLP-HPM	0.9368	42.4101	0.9377	1.9326	1.5219	0.9576	39.9379	0.9447	1.8689	2.2295
AWLP	0.9239	41.5724	0.9382	2.0083	1.9605	0.9463	38.4166	0.9278	1.8262	2.4017
TV	0.9485	41.9099	0.9477	2.1589	1.9592	0.9838	40.7908	0.9756	2.5083	2.0329
PanNet	0.9436	43.7540	0.9638	1.8252	1.4921	0.9756	41.5536	0.9676	2.2735	1.6732
DRPNN	0.9302	42.5424	0.9443	2.0893	1.7648	0.9793	41.4807	0.9760	2.9592	1.8487
MSDCNN	0.9542	44.6686	0.9633	1.6680	1.4062	0.9792	41.6150	0.9763	2.7362	1.7741
TFResNet	0.9598	44.5490	0.9684	1.5521	1.3148	0.9832	41.7154	0.9784	2.5056	1.6961
TFMamba	0.9558	46.3847	0.9774	1.2494	1.0390	0.9827	42.9195	0.9799	2.0728	1.4775
SSDiff	0.9203	41.5058	0.9435	2.1681	1.7016	0.9733	41.9928	0.9657	1.9595	1.6910
MambaIR	0.9818	47.8015	0.9845	1.1090	0.8562	0.9850	43.6362	0.9813	1.7045	1.3017
SpaXFus	**0.9827**	**48.0588**	**0.9855**	**1.0693**	**0.8289**	**0.9853**	**44.1236**	**0.9839**	**1.6100**	**1.2063**

and SRPSC are two traditional methods and the others are deep learning-based methods. Moreover, MSSR and PSSR are two SAR image super-resolution methods.

5.3. Evaluation metrics

In this work, five quantitative metrics are utilized to evaluate the fusion performance from spatial and spectral domains for the first three tasks. Correlation Coefficient (CC), mean Peak Signal-to-Noise Ratio (mPSNR) in decibel units, and mean Structural SIMilarity (mSSIM) indicate spatial fidelity (Wang et al., 2004). Spectral Angle Mapper (SAM) in degrees evaluates the spectral distortion (Kruse et al., 1993). Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS) (Wald, 2000) is a comprehensive metric. The higher values for CC, mSSIM, and mPSNR indicate better image quality. On the contrary, lower RMSE, SAM, and ERGAS indicate less image distortion.

As for the PolSAR fusion, due to the unique radiation characteristics of SAR data, we perform Pauli decomposition of the coherency matrix to obtain three polarimetric components:

$$P_1 = \frac{S_{\mathrm{HH}} + S_{\mathrm{VV}}}{\sqrt{2}}, \quad P_2 = \frac{S_{\mathrm{HH}} - S_{\mathrm{VV}}}{\sqrt{2}}, \quad P_3 = \frac{2S_{\mathrm{HV}}}{\sqrt{2}} \tag{49}$$

where P_1, P_2, and P_3 denote the odd scattering, double scattering, and volume scattering mechanisms, respectively. Mean Absolute Error (MAE), Root Mean Square Error (RMSE), SAM, mPSNR, and ERGAS are calculated on (P_1, P_2, P_3) to assess PolSAR fusion performance. Additionally, the Riemannian distance d_R between the predicted and ground truth C_3 matrices is utilized, effectively capturing their geometric relationships while fully considering manifold properties.

5.4. Experimental setup

5.4.1. Implementation details

In this study, the Adamax optimization algorithm is employed to train SpaXFus. The initial learning rate is set to 0.001 and follows the OneCycleLR adjustment strategy, which dynamically changes the learning rate during training to improve performance. All CNN-based models are trained using the PyTorch framework in a Linux environment with 1 TB RAM and an Nvidia A40 GPU. Traditional methods are implemented in MATLAB on a system with an Intel Core i7-1355U CPU (1.70 GHz). In different fusion tasks, all the models have been retrained.

5.4.2. Experimental workflow for downstream applications

In the experiments of this paper, for each fusion task, we used datasets from two different sensors to construct a benchmark for evaluating fusion performance. To analyze the impact of fusion on downstream tasks, we then selected data from a single sensor for each task to assess downstream performance, thereby examining how fusion quality affects these tasks.

Specifically, in the pansharpening task, we used Gaofen-2 data to generate NDVI products using ENVI, and evaluated the impact of fusion on NDVI generation using metrics such as Coefficient of Determination (R^2), RMSE, and MAE. In HMFusion, we selected the Chikusei dataset, which provides refined classification labels (Zhu et al., 2026). We employed an SVM algorithm together with the official labels to produce full-scale ground truth labels, and then randomly sampled 10 % of the data to examine the influence of fusion on classification accuracy before and after fusion using Overall Accuracy (OA), Kappa coefficient, and F1-score.

In STFusion, we used the Sen2Pla dataset to evaluate the downstream change detection task. Since Sen2Pla is derived from DynamicEarthNet, we were able to obtain corresponding change detection labels for performance assessment. The indices include F1-score, Intersection over Union (IoU), Precision, and Recall. Finally, for the PolSAR fusion task, the RadarSat-2 images over San Francisco from the dataset provided by Liu et al. (2022) include vegetation extraction labels. We used the original fusion results, after registration and cropping, to assess the accuracy of downstream SAR vegetation extraction using Producer's Accuracy (PA), User's Accuracy (UA), and IoU.

5.5. Pan-sharpening

Pan-sharpening is a typical remote sensing image fusion task that involves spatial and spectral information. In this part, we compare the fusion performance and also compare the differences in Normalized Difference Vegetation Index (NDVI) products generated from the fused results on Gaofen2 dataset.

5.5.1. Comparisons on pan-sharpening

Table 1 reports the quantitative results. Fig. 6 shows the visualization comparisons. Traditional methods such as BDSD, PRACS, and GSA demonstrated good performance, particularly on the QuickBird dataset. However, they generally underperformed compared to deep learning-based methods across most evaluation metrics as shown in Fig. 6(r). On the Gaofen2 dataset, methods like PRACS and ATWT-M3 showed relatively better performance, especially in terms of CC, mPSNR (Wang et al., 2004), and mSSIM, but still fell short when compared to deep learning-based techniques.

Deep learning algorithms, including PanNet, DRPNN, MSDCNN, and TFResNet, delivered excellent results across both datasets, with superior mPSNR, mSSIM, and SAM scores. MSDCNN and TFResNet stood out on the QuickBird dataset, with high mPSNR and mSSIM values, and

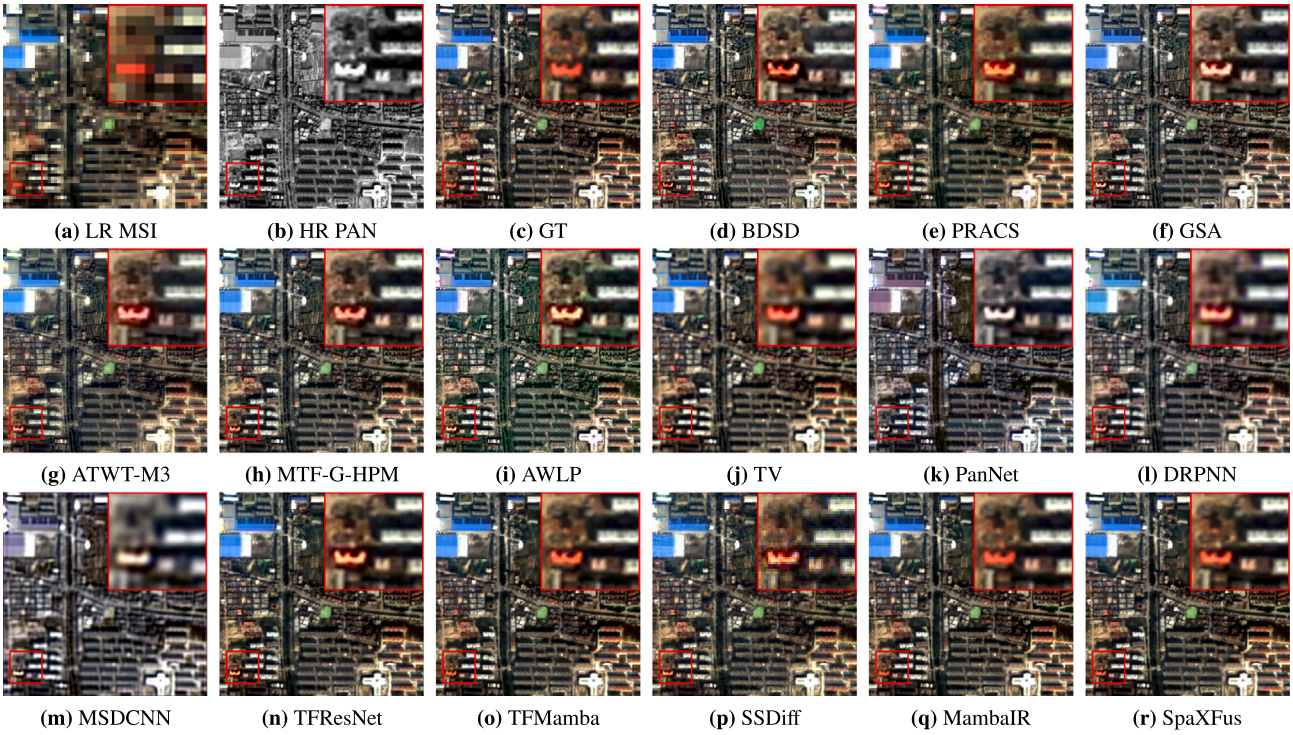

Fig. 6. Visualization comparisons of pan-sharpening on Quickbird dataset.

TFResNet also performed well in terms of SAM (Kruse et al., 1993) and ERGAS (Wald, 2000). SSDiff, a diffusion-based method, performed impressively on QuickBird, particularly in terms of mPSNR and mSSIM. On the Gaofen2 dataset, MambaIR, an SSM-based method, excelled in CC, mPSNR, and mSSIM.

The proposed SpaXFus algorithm outperformed all other methods on both datasets. SpaXFus achieved the highest CC, mPSNR, and mSSIM scores, with the lowest SAM and ERGAS values, as confirmed by visual inspection. Compared to MambaIR, SpaXFus showed a slight but consistent advantage in most metrics. These results, both quantitative and visual, demonstrate that SpaXFus is a highly effective algorithm for pan-sharpening tasks, surpassing existing state-of-the-art methods in both traditional and deep learning-based approaches.

5.5.2. Validation through NDVI product generation

Most pan-sharpening studies focus on comparing image quality metrics, often using RGB composites for visualization. This overlooks a key feature of remote sensing MS images: the inclusion of NIR bands, which are crucial for monitoring vegetation health. To address this, we evaluate fused MSIs through NDVI product generation, providing a more comprehensive assessment of fusion algorithms' ability to restore non-visible spectral bands.

Fig. S1 shows the NDVI results. Compared to NDVI from the original MS data, all methods improve spatial detail. Among traditional methods, CS-based algorithms outperform MRA-based ones but tend to underestimate NDVI in low-value (blue) regions due to poor NIR preservation. Deep learning methods, including DRPNN, TFResNet, and MambaIR, recover details well without notable underestimation or overestimation, while MSDCNN overestimates in blue regions. The proposed SpaXFus achieves comparable spatial details to CS-based methods while avoiding estimation errors, demonstrating strong recovery across spectral bands.

Fig. 7 provides quantitative comparisons with ground truth. BDSD, TV, TFResNet, and SSDiff underestimate low NDVI values, with TV, TFResNet, and SSDiff also overestimating high values. ATWT-M3 tends to overestimate overall. In contrast, SpaXFus achieves the highest R^2 and lowest error, confirming its ability to preserve spatial and spectral information effectively.

5.6. Hyperspectral-multispectral fusion

HMFusion can integrate spatial details with hyperspectral information, which can effectively reflect the subtle differences in surface objects. Thus, after HMFusion, we also introduce fine-grained image classification to verify the reliability of the fused information in this subsection.

5.6.1. Comparisons on HMFusion

Based on the quantitative results in Table 2 and the visual results in Fig. 8, SpaXFus outperforms all comparison methods on both datasets. On the CAVE dataset, it achieves the highest CC, mPSNR, and mSSIM values, as well as the lowest SAM and ERGAS scores, demonstrating its ability to produce high-quality fused images. This is also evident in the visual results, where SpaXFus delivers clearer and more detailed fused images than other methods. As shown in Fig. 9, all methods recover accurate spectra for fake pepper, but only SpaXFus works well on the real pepper location. Additionally, SpaXFus outperforms TFMamba, highlighting the effectiveness of the model-driven framework and Spatial-X interaction.

In the Sen2Chikusei dataset, SpaXFus again outperforms all other methods, achieving the highest CC, mPSNR, and mSSIM scores. The performance of Fusformer remains strong in terms of mPSNR, but SpaXFus achieves the best results across the board, including the lowest SAM and ERGAS values. Compared to TFMamba, SpaXFus achieves superior results in all metrics, particularly in SAM and ERGAS, indicating a more accurate fusion quality. Visually, SpaXFus consistently produces sharper and more precise images, with finer details preserved, particularly in regions of complex texture and high spatial resolution.

Overall, the results across both datasets demonstrate that SpaXFus not only excels in quantitative metrics but also offers significant improvements in visual image quality, outperforming existing state-of-the-art methods such as Fusformer and TFMamba in all major evaluation

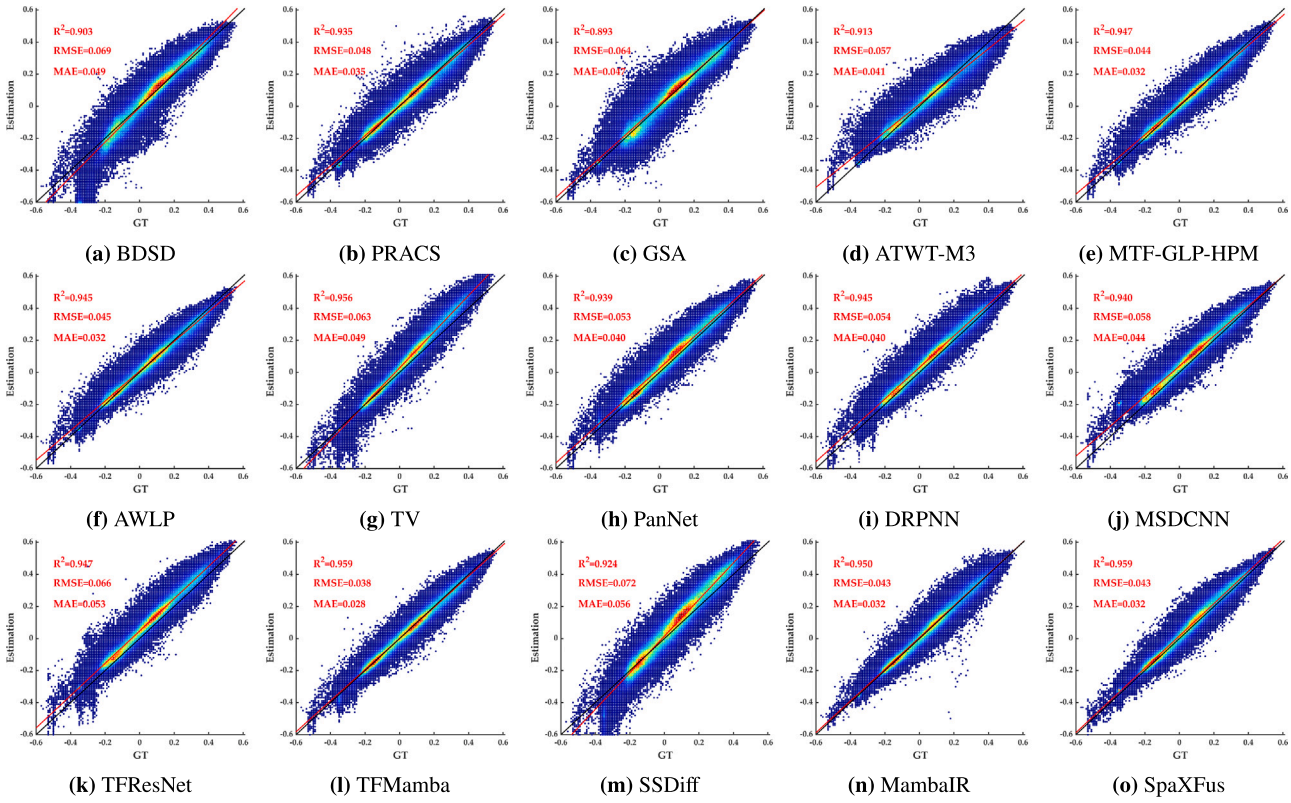

Fig. 7. Relations between the different estimated NDVIs and the GT. Black lines denote the ideal relationship $y = x$, and red lines illustrate the linear regression results. The color illustrates the density of samples. Goodness of fit R^2, RMSE, and MAE are displayed at the top left. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Quantitative comparisons of HMFusion on both CAVE and Sen2Chikusei datasets. The best is in **bold**, and the second best is underlined.

Method	CAVE					Sen2Chikusei				
	CC	mPSNR	mSSIM	SAM	ERGAS	CC	mPSNR	mSSIM	SAM	ERGAS
Hysure	0.9812	28.5002	0.8363	14.9918	6.1543	0.9913	44.5993	0.9890	1.5827	2.2661
CNMF	0.9871	28.1390	0.9218	7.3881	4.7168	0.9911	42.2850	0.9868	1.5419	2.2430
FUSE	0.9308	26.6617	0.8242	16.5955	12.6208	0.9924	41.1262	0.9844	1.4366	2.0336
CSU	0.9535	21.8613	0.7974	11.0717	9.8701	0.9889	33.2866	0.9363	1.6667	2.6124
CSTF	0.9574	22.4697	0.6063	19.3558	18.9112	0.7569	25.7209	0.6252	14.3557	16.2700
NSSR	0.9900	24.0462	0.8764	4.1425	4.6079	0.9732	37.4768	0.9404	2.7805	3.2582
LTMR	0.9799	24.0876	0.8491	6.8334	6.4586	0.9582	35.8172	0.9277	3.2708	4.5296
GTNN	0.9814	25.5633	0.8661	6.4658	6.2032	0.9650	37.2165	0.9431	2.9060	4.0843
PSRT	0.9980	38.4869	0.9774	2.8702	1.9882	0.9852	43.3565	0.9923	1.7362	3.2720
MSST	0.9991	38.1230	0.9900	2.5685	1.3155	0.9856	43.3058	0.9945	1.7036	2.7657
SSRNET	0.9980	36.7293	0.9838	4.5126	2.0905	0.9890	44.8197	0.9958	1.4280	2.3605
Fusformer	0.9984	**39.5142**	0.9882	2.2666	1.7862	0.9910	**48.7963**	0.9954	1.4013	2.4323
PNXnet	0.9981	38.4713	0.9798	3.4012	1.9742	0.9928	46.6941	0.9973	1.1841	1.8197
TFMamba	0.9978	36.9468	0.9836	3.2017	2.3183	0.9928	46.7363	0.9973	1.1806	1.8151
SpaXFus	**0.9997**	39.4677	**0.9939**	**1.8645**	**0.7663**	**0.9942**	48.6815	**0.9977**	**1.0486**	**1.6025**

criteria. These results validate SpaXFus as an effective and reliable method for the HMFusion task.

5.6.2. Validation through fine-grained image classification

Hyperspectral images provide detailed spectral information that aids in recognizing ground objects that are difficult to distinguish. To analyze how much HMFusion methods can enhance spectral representation, this study uses fine-grained image classification as a downstream validation. The algorithm employed is the classical Support Vector Machine (SVM), using the Chikusei dataset with classification labels. The dataset includes 19 classes, excluding the first class (water) in the selected study area, as shown in Fig. S2. In addition to common land cover classes, the study refines the classes of vegetation and bare soil, increasing classification difficulty.

Table S2 presents the fine-grained classification results with fused data, including OA, Kappa coefficient, and F1-score. Compared to classification using the original four-band Sen2 data, most fusion algorithms improve accuracy, except for CNMF and CSTF. Among traditional algorithms, CSU is most beneficial, while deep learning-based algorithms all significantly improve accuracy. TFMamba and SpaXFus, which consider internal dependencies of hyperspectral data, show substantial

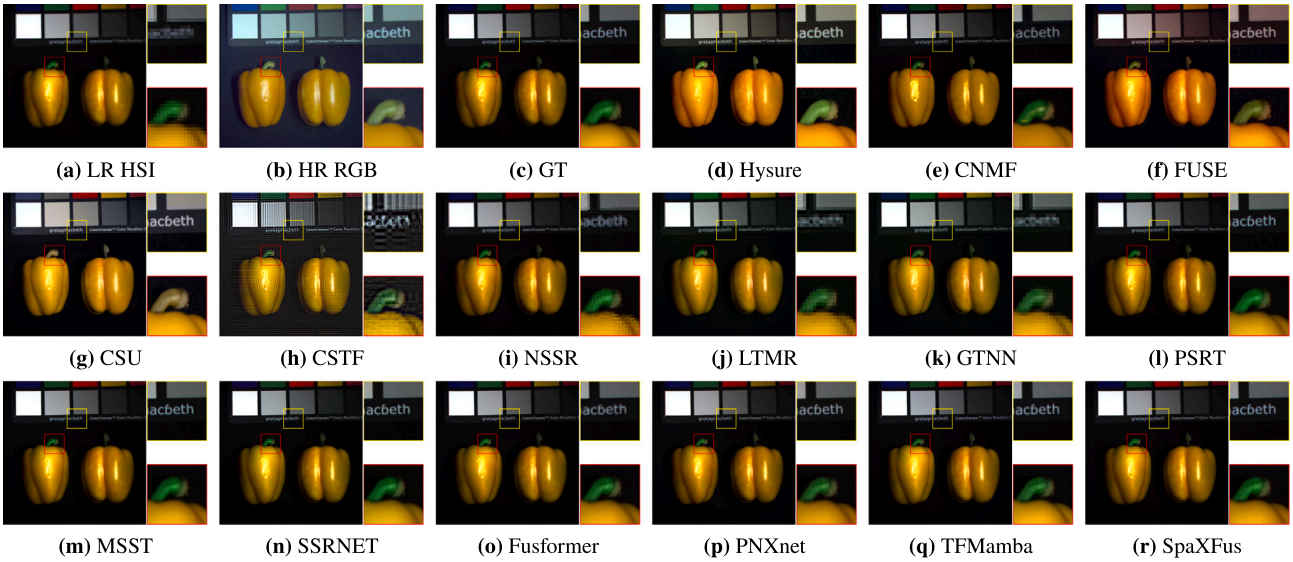

Fig. 8. Visualization comparisons of HMFusion on CAVE dataset.

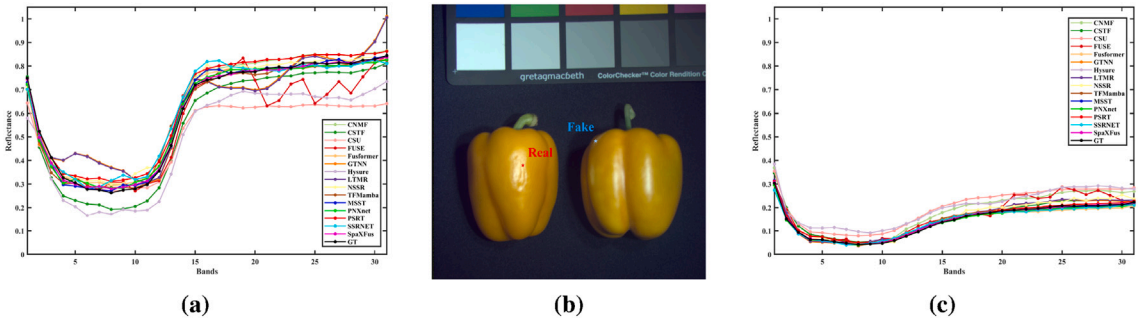

Fig. 9. Reflectances of "Peppers" images from the CAVE data set at the fake and real pepper locations. (a) Reflectances of results at the real pepper location. (b) Locations of samples for real and fake peppers. (c) Reflectances of results at the fake pepper location.

advantages. SpaXFus achieves the best F1-score and OA, indicating its superior integration of hyperspectral information.

Fig. 10 shows confusion matrices of classification, where the horizontal axis represents GT classes, and the vertical axis represents predicted classes. The darker the box, the higher the percentage, showing the proportion of GT samples classified into each class. Nearly all algorithms struggle with artificial grass classification (class 16), and CSTF fails to predict the first seven classes. However, Fusformer and SpaXFus perform well. For the challenging bare soil classes, SpaXFus shows significant advantages due to its ability to exploit X-intrinsic interaction, capturing spectral differences more effectively.

Visual comparisons in Fig. S2 show that CSTF suffers from poor accuracy due to its grid effect. In the zoomed-in region, high-brightness interference degrades the spectral information of the blue house in the original multispectral data, making it unrecognizable. MSST and SSRNET misclassify the blue house. Traditional algorithms generally fail to distinguish plastic houses from white roofs, except for FUSE, NSSR, and LTMR. Deep learning-based algorithms help identify white houses, and SpaXFus shows the closest match to the GT, supporting the quantitative evaluation.

5.7. Spatio-temporal fusion

STFusion can generate the HR data for different time points using existing data, which is crucial for monitoring land cover changes. To validate the authenticity of the temporal information generated by STFusion algorithms, we use change detection as a downstream application.

5.7.1. Comparisons on STFusion

The experimental results of the spatio-temporal fusion task on both the Land4Mod and Sen2Pla datasets are shown in Table 3. Traditional algorithms, such as STARFM, FSDAF, and Fit-FC, demonstrate lower performance across the evaluation metrics, especially on Sen2Pla dataset. Specifically, Fit-FC exhibits a lower CC value and higher ERGAS on both datasets, indicating its limited capability for accurate STFusion.

Among the deep learning-based methods, MUSTFN and ECPW-STFN stand out, with MUSTFN achieving high CC, mPSNR, and mSSIM on the Land4Mod dataset. However, SpaXFus outperforms all methods, achieving the best performance across almost all metrics on both the Land4Mod and Sen2Pla datasets. On the Land4Mod dataset, SpaXFus achieves the best CC, mPSNR, mSSIM, ERGAS, and especially the lowest SAM values, demonstrating its superiority in STFusion.

Furthermore, when visually compared on the Land4Mod dataset, as shown in Fig. 11, SpaXFus provides noticeably sharper and more detailed fusion results, preserving spatial and spectral features with higher clarity and accuracy. Other deep learning-based methods show blurring, while traditional methods show spectral distortion.

5.7.2. Validation through change detection

Given that DynamicEarthNet contains land cover across various time points, we chose the Sen2Pla dataset for change detection (Toker et al.,

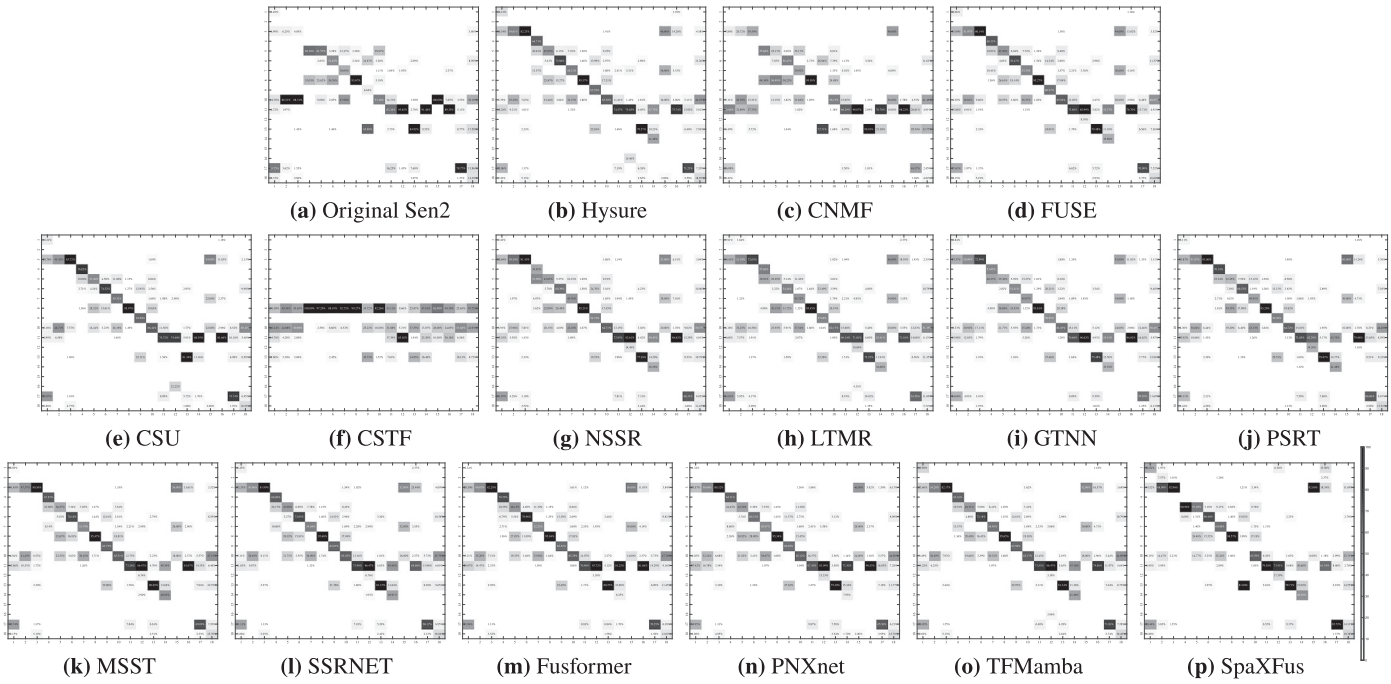

Fig. 10. Visualization of the confusion matrices for refined classification results based on the fused outputs from the Sen2Chikusei dataset. Rows represent the predicted labels, while columns denote the GT labels. The percentages within the matrix indicate the proportion of GT samples classified into each specific class, with only values exceeding 1 % being displayed. Classes 1 to 18 correspond to: Bare Soil (school), Bare Soil (park), Bare Soil (farmland), Natural Plants, Weeds in Farmland, Forest, Grass, Rice Field (grown), Rice Field (1st stage), Row Crops, Plastic House, Manmade (non-dark), Manmade (dark), Manmade (blue), Manmade (red), Manmade (grass), Asphalt, and Paved Ground, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Quantitative comparisons of spatio-temporal fusion on both Land4Mod (Li et al., 2020a) and Sen2Pla datasets. The best is in **bold**, and the second best is underlined.

Method	Land4Mod					Sen2Pla				
	CC	mPSNR	mSSIM	SAM	ERGAS	CC	mPSNR	mSSIM	SAM	ERGAS
STARFM	0.6089	29.7540	0.7177	4.2705	2.1178	0.3591	18.7404	0.7974	4.6189	20.9687
FSDAF	0.5874	29.2198	0.7579	4.3049	2.1601	0.3575	18.4040	0.7677	4.6675	21.0242
Fit-FC	0.4699	30.3249	0.8134	5.7226	3.4078	0.3272	23.5089	0.6836	8.0819	25.1615
STTFN	0.6042	27.0005	0.8326	5.7647	2.0684	0.6773	20.1320	0.8884	6.0790	17.3575
MUSTFN	0.7220	32.4079	0.8783	4.1098	1.0800	0.5127	20.4120	0.7383	5.8042	24.4401
ECPW-STFN	0.6148	29.1293	0.8469	5.4524	1.5184	0.7228	20.7511	**0.9121**	9.1036	**14.4571**
TFMamba	0.5843	30.1196	0.8532	4.7120	1.3726	0.6191	19.0777	0.8428	4.0303	26.2831
SpaXFus	**0.7898**	**34.1518**	**0.9018**	**3.3192**	**0.8489**	**0.9029**	**23.6275**	0.9061	**2.3893**	15.0765

2022). ChangeStar2 is used as the change detection method (Zheng et al., 2024). The change detection results are shown in Fig. S3, which presents land cover classifications at two time points and the ground truth (GT) for change detection. Change detection result of the original LR Sen2 image pair is shown in Fig. S3d. Due to the low resolution, it is difficult to extract meaningful semantic information, resulting in poor detection. Among the traditional algorithms, Fit-FC performs well and detects almost all changes. In deep learning algorithms, the results for large-scale changes are generally good. TFMamba and SpaXFus, which account for global dependencies, perform better. The proposed SpaXFus method yields the best results.

For a more accurate comparison, Table S1 presents the quantitative results of change detection, including IoU, Precision, Recall, and F1 score. The original Sentinel-2 data struggles to provide accurate semantic information, resulting in poor change detection performance. Fit-FC achieves the highest Recall, but its low Precision indicates significant false positives. TFMamba has the highest Precision but a low Recall, leading to substantial false negatives. Both algorithms have relatively low F1 scores. In contrast, the detection results based on SpaXFus fusion data achieve the highest IoU and F1 score, maintaining a good balance between Precision and Recall, indicating minimal false positives and false negatives. This also suggests that the spatio-temporal information integrated by SpaXFus is more reliable.

5.8. PolSAR fusion

PolSAR fusion can generate high spatial resolution fully-polarimetric SAR images, which contain rich polarimetric and scattering information, providing data for more accurate surface detection. To assess the reliability of the fused polarimetric information, this study introduces vegetation extraction as the downstream application for PolSAR fusion.

5.8.1. Comparisons on PolSAR fusion

The results of quantitative comparisons, presented in Table 4, demonstrate the effectiveness of SpaXFus in comparison with several traditional and deep learning-based methods. Bicubic and SRPSC, as traditional methods, exhibit relatively high errors in terms of MAE, RMSE,

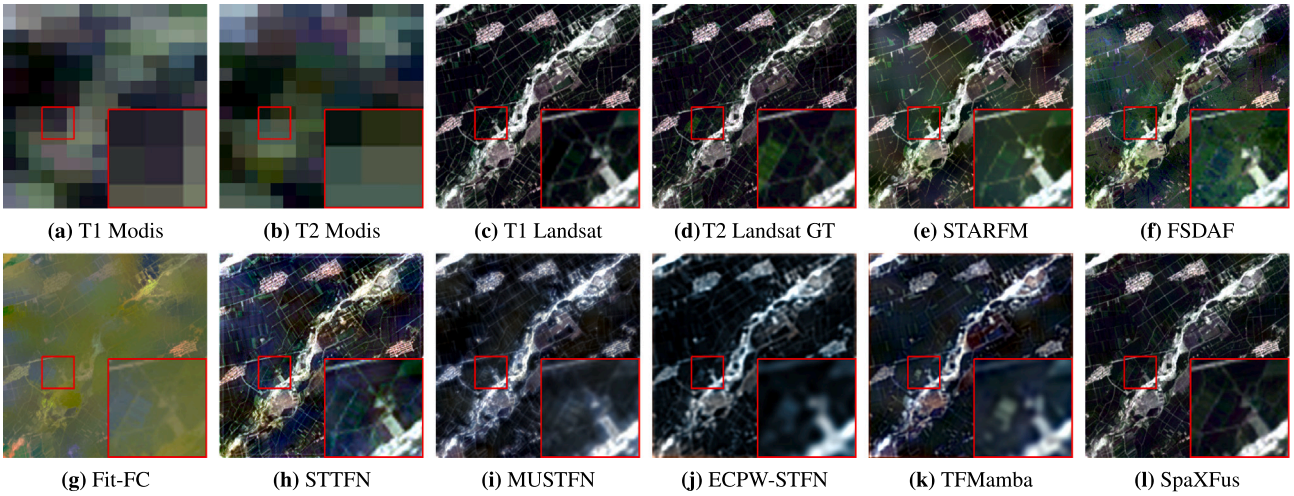

Fig. 11. Visualization comparisons of STFusion on Land4Mod dataset.

Table 4
Quantitative comparisons of PolSAR fusion on both San Francisco and Quebec areas. The best is in **bold**, and the second best is underlined.

Method	San Francisco						Quebec					
	MAE	RMSE	SAM	mPSNR	ERGAS	d_R	MAE	RMSE	SAM	mPSNR	ERGAS	d_R
Bicubic	0.0738	0.6077	4.4669	41.9463	44.1183	0.9240	0.1260	1.3416	10.7310	48.4529	54.1829	1.3978
SRPSC	0.0940	0.7326	6.0815	40.6827	51.2048	1.0853	0.1259	1.3637	10.9392	48.5932	54.6830	1.4028
MSSR	0.0733	0.5909	3.4067	41.8994	43.2933	0.6717	0.1280	1.3629	9.9977	47.6770	55.4798	1.4489
PSSR	0.0767	0.5993	5.9913	42.0909	49.0304	0.9416	0.1296	1.3071	12.8065	**50.5043**	50.3007	1.5961
MSPSRN	0.0739	0.5948	4.0591	42.2034	48.9310	0.8030	0.1161	1.2866	10.3558	49.7582	56.9625	1.5279
PSFN	0.0237	0.2002	3.9764	48.1620	9.3040	0.9624	0.0509	0.8120	11.1725	48.3388	33.2591	1.1419
FDFNet	0.0197	**0.1768**	3.3050	48.3027	7.7423	0.5520	0.0425	0.8126	7.4615	48.2313	31.6319	0.7646
SpaXFus	**0.0153**	0.1824	**3.2403**	**48.4783**	**7.7101**	**0.5139**	**0.0416**	**0.7963**	**7.3214**	48.8016	**30.9526**	**0.7520**

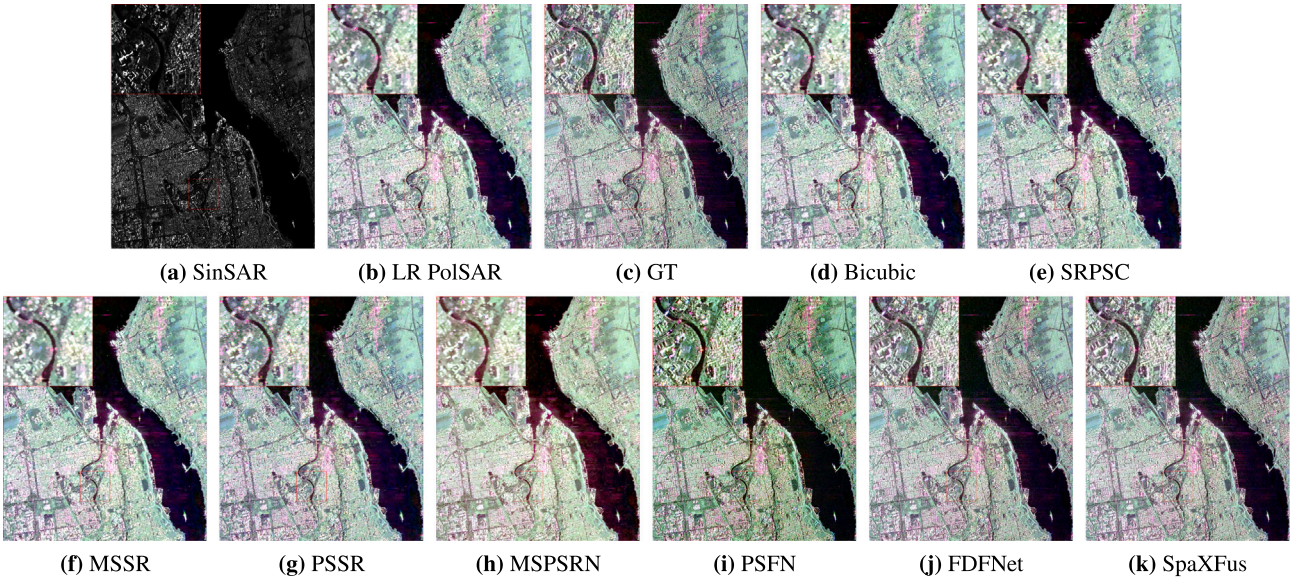

Fig. 12. PolSAR fusion results on Quebec dataset.

and SAM, particularly on the Quebec dataset, which is a real-world dataset. Deep learning-based methods, such as MSSR, PSSR, MSPSRN, PSFN, and FDFNet, show significant improvements, especially in terms of MAE and SAM, compared to the traditional methods.

In particular, SpaXFus outperforms all other methods across almost all evaluation metrics, including MAE, RMSE, SAM, mPSNR, and ERGAS, on both San Francisco and Quebec datasets. For example, on the Quebec dataset, SpaXFus achieves the lowest MAE and RMSE, indicating its superior ability to preserve fine details in the fused PolSAR images. Moreover, SpaXFus delivers the highest mPSNR and the lowest ERGAS, demonstrating its effectiveness in generating high-quality fused images that better match the real-world data. The d_R metric, which evaluates

the preservation of polarimetric information, also confirms the superiority of SpaXFus. On both the San Francisco and Quebec datasets, SpaXFus achieves the best d_R score, with values of 0.5139 and 0.7520, respectively, outperforming all other methods.

On the Quebec dataset, the visual results shown in Fig. 12 are consistent with the quantitative analysis, where SpaXFus produces more accurate and clearer images compared to other methods. The visual comparison reveals that SpaXFus is able to preserve key features, such as vegetation, buildings, and roads, much better than the competing methods. This is particularly evident in the finer details of the high-resolution fused images generated by SpaXFus versus those generated by methods like Bicubic or SRPSC, which tend to introduce more blurring and artifacts. This highlights SpaXFus as a promising method for PolSAR image processing and related remote sensing applications, especially in practical, real-world scenarios.

5.8.2. Validation through SAR vegetation extraction

Fig. S4 presents the results of vegetation extraction on HR PolSAR data generated by various fusion algorithms using the SVM algorithm. Observing the results from SinSAR, it can be seen that vegetation extraction from SAR requires multi-polarimetric scattering information to effectively differentiate vegetation from other land covers. When using PolSAR data, even simple Bicubic interpolation can extract the general vegetation, although there are still many false positives. Several super-resolution-based algorithms produce fragmented results with more false positives, while fusion-based algorithms yield results more similar to GT, particularly in the elongated region in the top-left corner, which is distinguished from the vegetation. Overall, the results from MSPSRN and SpaXFus are the best.

Table S3 presents the quantitative results of vegetation extraction, including Producer's Accuracy (PA), User's Accuracy (UA), and IoU. SinSAR-based results have high PA due to numerous false positives, nearly identifying all land covers as vegetation, which leads to low UA and IoU. SpaXFus achieves the highest UA, and its PA is the highest among all fusion algorithms, resulting in the highest IoU. This indicates that the polarimetric information fused by the proposed SpaXFus is more reliable and avoids distortion.

6. Conclusions

In this study, we summarize four existing remote sensing image fusion problems involving spatial degradation into a broader concept generalized spatial-channel fusion, termed spatial-X fusion. To address this, we propose a universal framework, SpaXFus, which integrates a model-driven unfolding framework with spatial-X intrinsic interaction. By leveraging degradation models, the algorithm demonstrates strong generalization capabilities while effectively capturing mutual dependencies and self-interactions in both the spatial and X domains. This results in broader applicability and superior performance in generalized spatial-channel fusion. Experimental results across multiple datasets and four different fusion tasks highlight the superiority of SpaXFus. Additionally, we build a benchmark where downstream applications are introduced to assess the effectiveness of the fused information. While the proposed method shows promising results, its dependency on data remains a limitation. SpaXFus must be retrained for each distinct task. Future work should focus on developing a spatial-channel fusion foundation model for few-shot or even zero-shot scenarios. Moreover, as this study indicates that fusion benefits downstream applications, achieving multi-level and multi-task collaboration will advance remote sensing image intelligent processing and understanding. Codes are released at https://github.com/zhu-xlab/SpaXFus.

CRediT authorship contribution statement

Jiang He: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Liupeng Lin:** Writing – review & editing, Visualization, Validation, Investigation, Data curation. **Zhuo Zheng:** Writing – review & editing, Visualization, Validation, Investigation, Data curation. **Qiangqiang Yuan:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Jie Li:** Resources, Data curation. **Liangpei Zhang:** Writing – review & editing, Supervision, Resources, Project administration, Conceptualization. **Xiao xiang Zhu:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Acknowledgments

This work was supported in part by the German Research Foundation (DFG GZ: ZH 498/18–1; Project number: 519016653); in part by the National Natural Science Foundation of China under Grant 42471414; and by the Munich Center for Machine Learning.

Appendix A. Supplementary data

Supplementary data for this article can be found online at doi:10.1016/j.rse.2025.115214.

Data availability

Data will be made available on request.

References

Aiazzi, B., Alparone, L., Baronti, S., Garzelli, A., 2002. Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis. IEEE Trans. Geosci. Remote Sens. 40, 2300–2312.

Aiazzi, B., Alparone, L., Baronti, S., Garzelli, A., Selva, M., 2006. Mtf-tailored multiscale fusion of high-resolution MS and pan imagery. Photogramm. Eng. Remote Sens. 72, 591–596.

Aiazzi, B., Baronti, S., Selva, M., 2007. Improving component substitution pansharpening through multivariate regression of MS + pan data. IEEE Trans. Geosci. Remote Sens. 45, 3230–3239.

Akhtar, N., Shafait, F., Mian, A., 2015. Bayesian sparse representation for hyperspectral image super resolution. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3631–3640, https://doi.org/10.1109/CVPR.2015.7298986

Ballester, C., Caselles, V., Igual, L., Verdera, J., Rougé, B., 2006. A variational model for p+ xs image fusion. Int. J. Comput. Vis. 69, 43–58.

Benedek, C., Descombes, X., Zerubia, J., 2011. Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics. IEEE Trans. Pattern Anal. Mach. Intell. 34, 33–50.

Burt, P.J., Adelson, E.H., 1987. The Laplacian pyramid as a compact image code. In: Readings in Computer Vision. Elsevier, pp. 671–679.

Cao, X., Chen, Y., Cao, W., 2022. Proximal Pannet: a model-based deep network for pansharpening. Proc. AAAI Conf. Artif. Intell. 36, 176–184. https://doi.org/10.1609/aaai.v36i1.19892

Chen, G., Jiao, P., Hu, Q., Xiao, L., Ye, Z., 2022. Swinstfm: remote sensing spatiotemporal fusion using swin transformer. IEEE Trans. Geosci. Remote Sens. 60, 1–18.

Chen, J., Wang, L., Feng, R., Liu, P., Han, W., Chen, X., 2020. Cyclegan-stf: spatiotemporal fusion via Cyclegan-based image generation. IEEE Trans. Geosci. Remote Sens. 59, 5851–5865.

Chen, L., Fu, Y., Gu, L., Yan, C., Harada, T., Huang, G., 2024. Frequency-aware feature fusion for dense image prediction. IEEE Trans. Pattern Anal. Mach. Intell. 46, 10763–10780.

Choi, J., Yu, K., Kim, Y., 2010. A new adaptive component-substitution-based satellite image fusion by using partial replacement. IEEE Trans. Geosci. Remote Sens. 49, 295–309.

Deng, S.-Q., Deng, L.-J., Wu, X., Ran, R., Hong, D., Vivone, G., 2023. PSRT: pyramid shuffle-and-reshuffle transformer for multispectral and hyperspectral image fusion. IEEE Trans. Geosci. Remote Sens. 61, 1–15.

Dian, R., Li, S., 2019. Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization. IEEE Trans. Image Process. 28, 5135–5146.

Dian, R., Liu, Y., Li, S., 2024. Hyperspectral image fusion via a novel generalized tensor nuclear norm regularization. IEEE Trans. Neural Netw. Learn. Syst..

Do, M.N., Vetterli, M., 2005. The contourlet transform: an efficient directional multiresolution image representation. IEEE Trans. Image Process. 14, 2091–2106.

Dong, W., Fu, F., Shi, G., Cao, X., Wu, J., Li, G., Li, X., 2016. Hyperspectral image super-resolution via non-negative structured sparse representation. IEEE Trans. Image Process. 25, 2337–2352.

Fu, X., Lin, Z., Huang, Y., Ding, X., 2019. A variational pan-sharpening with local gradient constraints. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10265–10274.
Fu, Y., Zhang, T., Wang, L., Huang, H., 2022. Coded hyperspectral image reconstruction using deep external and internal learning. IEEE Trans. Pattern Anal. Mach. Intell. 44, 3404–3420.
Gao, F., Masek, J., Schwaller, M., Hall, F., 2006. On the blending of the Landsat and MODIS surface reflectance: predicting daily Landsat surface reflectance. IEEE Trans. Geosci. Remote Sens. 44, 2207–2218.
Garzelli, A., Nencini, F., Capobianco, L., 2007. Optimal MMSE pan sharpening of very high resolution multispectral images. IEEE Trans. Geosci. Remote Sens. 46, 228–236.
Gevaert, C.M., GarcÃ-A-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
Gillespie, A.R., Kahle, A.B., Walker, R.E., 1987. Color enhancement of highly correlated images. II. Channel ratio and chromaticity transformation techniques. Remote Sens. Environ. 22, 343–365.
González-Audícana, M., Saleta, J.L., Catalán, R.G., García, R., 2004. Fusion of multispectral and panchromatic images using improved IHS and PCA mergers based on wavelet decomposition. IEEE Trans. Geosci. Remote Sens. 42, 1291–1299.
Gu, A., Goel, K., Ré, C., 2022. Efficiently modeling long sequences with structured state spaces. In: The International Conference on Learning Representations (ICLR).
Gu, Z., Chen, J., Chen, Y., Qiu, Y., Zhu, X., Chen, X., 2023. Agri-Fuse: a novel spatiotemporal fusion method designed for agricultural scenarios with diverse phenological changes. Remote Sens. Environ. 299, 113874.
Guan, P., Lam, E.Y., 2022. Multistage dual-attention guided fusion network for hyperspectral pansharpening. IEEE Trans. Geosci. Remote Sens. 60, 1–14. https://doi.org/10.1109/tgrs.2021.3114552
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.
Guo, H., Li, J., Dai, T., Ouyang, Z., Ren, X., Xia, S.-T., 2025. Mambair: a simple baseline for image restoration with state-space model. In: European Conference on Computer Vision. Springer, pp. 222–241.
He, J., Yuan, Q., Li, J., Xiao, Y., Zhang, L., 2023. A self-supervised remote sensing image fusion framework with dual-stage self-learning and spectral super-resolution injection. ISPRS J. Photogramm. Remote Sens. 204, 131–144.
He, J., Yuan, Q., Li, J., Zhang, L., 2022. A knowledge optimization-driven network with normalizer-free group resnet prior for remote sensing image pan-sharpening. IEEE Trans. Geosci. Remote Sens. 60, 1–16. https://doi.org/10.1109/TGRS.2022.3186916
Hilker, T., Wulder, M.A., Coops, N.C., Linke, J., McDermid, G., Masek, J.G., Gao, F., White, J.C., 2009. 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
Hong, D., Zhang, B., Li, X., Li, Y., Li, C., Yao, J., Yokoya, N., Li, H., Ghamisi, P., Jia, X., Plaza, A., Gamba, P., Benediktsson, J.A., Chanussot, J., 2024. Spectralgpt: Spectral Remote Sensing Foundation model. IEEE Trans. Pattern Anal. Mach. Intell. 46, 5227–5244.
Hu, J.-F., Huang, T.-Z., Deng, L.-J., Dou, H.-X., Hong, D., Vivone, G., 2022. Fusformer: a transformer-based fusion network for hyperspectral image super-resolution. IEEE Geosci. Remote Sens. Lett. 19, 1–5.
Huang, B., Song, H., 2012. Spatiotemporal reflectance fusion via sparse representation. IEEE Trans. Geosci. Remote Sens. 50, 3707–3716.
Javan, F.D., Samadzadegan, F., Mehravar, S., Toosi, A., Khatami, R., Stein, A., 2021. A review of image fusion techniques for pan-sharpening of high-resolution satellite imagery. ISPRS J. Photogramm. Remote Sens. 171, 101–117.
Jia, S., Min, Z., Fu, X., 2023. Multiscale spatial–spectral transformer network for hyperspectral and multispectral image fusion. Inf. Fusion 96, 117–129.
Jiong, C., Jian, Y., 2007. Super-resolution of polarimetric SAR images for ship detection. In: 2007 International Symposium on Microwave, Antenna, Propagation and EMC Technologies for Wireless Communications. IEEE, pp. 1499–1502.
Kruse, F.A., Lefkoff, A.B., Boardman, J.W., Heidebrecht, K.B., Shapiro, A.T., Barloon, P.J., Goetz, A.F.H., 1993. The spectral image processing system (SIPS)-interactive visualization and analysis of imaging spectrometer data. Remote Sens. Environ. 44, 145–163.
Kwarteng, P., Chavez, A., 1989. Extracting spectral contrast in Landsat thematic mapper image data using selective principal component analysis. Photogramm. Eng. Remote Sens. 55, 339–348.
Lanaras, C., Baltsavias, E., Schindler, K., 2015. Hyperspectral super-resolution by coupled spectral unmixing. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3586–3594.
Lee, J.-S., Pottier, E., 2017. Polarimetric Radar Imaging: from Basics to Applications. CRC Press.
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, J., Li, Y., He, L., Chen, J., Plaza, A., 2020a. Spatio-temporal fusion for remote sensing data: an overview and new benchmark. Sci. China Inf. Sci. 63, 1–17.
Li, S., Dian, R., Fang, L., Bioucas-Dias, J.M., 2018a. Fusing hyperspectral and multispectral images via coupled sparse tensor factorization. IEEE Trans. Image Process. 27, 4118–4130. https://doi.org/10.1109/TIP.2018.2836307
Li, S., Dian, R., Fang, L., Bioucas-Dias, J.M., 2018b. Fusing hyperspectral and multispectral images via coupled sparse tensor factorization. IEEE Trans. Image Process. 27, 4118–4130.
Li, X., Foody, G.M., Boyd, D.S., Ge, Y., Zhang, Y., Du, Y., Ling, F., 2020b. SFSDAF: an enhanced fsdaf that incorporates sub-pixel class fraction change information for spatio-temporal image fusion. Remote Sens. Environ. 237, 111537.
Lin, L., Li, J., Shen, H., 2023. Polsar image deep learning super-resolution model based on multi-scale attention mechanism. Natl. Remote Sens. Bull. 1–10. https://doi.org/10.11834/jrs.20233002
Lin, L., Li, J., Shen, H., Zhao, L., Yuan, Q., Li, X., 2021a. Low-resolution fully polarimetric SAR and high-resolution single-polarization SAR image fusion network. IEEE Trans. Geosci. Remote Sens. 60, 1–17.
Lin, L., Li, J., Shen, H., Zhao, L., Yuan, Q., Li, X., 2022. Low-resolution fully polarimetric SAR and high-resolution single-polarization SAR image fusion network. IEEE Trans. Geosci. Remote Sens. 60, 1–17. https://doi.org/10.1109/TGRS.2021.3121166
Lin, L., Li, J., Yuan, Q., Shen, H., 2019. Polarimetric SAR image super-resolution via deep convolutional neural network. In: IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 3205–3208.
Lin, L., Shen, H., Li, J., Yuan, Q., 2021b. FDFNet: a fusion network for generating high-resolution fully polsar images. IEEE Geosci. Remote Sens. Lett. 19, 1–5.
Liu, J.G., 2000. Smoothing filter-based intensity modulation: a spectral preserve image fusion technique for improving spatial details. Int. J. Remote Sens. 21, 3461–3472.
Liu, J., Feng, Y., Zhou, C., Zhang, C., 2020a. PWNET: an adaptive weight network for the fusion of panchromatic and multispectral images. Remote Sens. 12, 2804.
Liu, Q., Zhou, H., Xu, Q., Liu, X., Wang, Y., 2020b. PSGAN: a generative adversarial network for remote sensing image pan-sharpening. IEEE Trans. Geosci. Remote Sens. 59, 10227–10242.
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. Remote Sens. 57, 6552–6564.
Liu, X., Jiao, L., Liu, F., Zhang, D., Tang, X., 2022. Polsf: Polsar image datasets on SAN Francisco. In: International Conference on Intelligence Science. Springer, pp. 214–219.
Liu, X., Liu, Q., Wang, Y., 2020c. Remote sensing image fusion based on two-stream fusion network. Inf. Fusion 55, 1–15.
Loncan, L., De Almeida, L.B., Bioucas-Dias, J.M., Briottet, X., Chanussot, J., Dobigeon, N., Fabre, S., Liao, W., Licciardi, G.A., Simoes, M., et al., 2015. Hyperspectral pansharpening: a review. IEEE Geosci. Remote Sens. Mag. 3, 27–46.
Luo, S., Zhou, S., Feng, Y., Xie, J., 2020. Pansharpening via unsupervised convolutional neural networks. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 4295–4310.
Ma, J., Yu, W., Chen, C., Liang, P., Guo, X., Jiang, J., 2020. Pan-gan: an unsupervised pan-sharpening method for remote sensing image fusion. Inf. Fusion 62, 110–120.
Masi, G., Cozzolino, D., Verdoliva, L., Scarpa, G., 2016. Pansharpening by convolutional neural networks. Remote Sens. 8, 594.
Meng, Q., Shi, W., Li, S., Zhang, L., 2023. Pandiff: a novel pansharpening method based on denoising diffusion probabilistic model. IEEE Trans. Geosci. Remote Sens. 61, 1–17. https://doi.org/10.1109/TGRS.2023.3279864
Ni, J., Shao, Z., Zhang, Z., Hou, M., Zhou, J., Fang, L., Zhang, Y., 2022. LDP-Net: an unsupervised pansharpening network based on learnable degradation processes. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 5468–5479.
Nunez, J., Otazu, X., Fors, O., Prades, A., Pala, V., Arbiol, R., 1999. Multiresolution-based image fusion with additive wavelet decomposition. IEEE Trans. Geosci. Remote Sens. 37, 1204–1211.
Otazu, X., González-Audícana, M., Fors, O., Núñez, J., 2005. Introduction of sensor spectral response into image fusion methods. Application to wavelet-based methods. IEEE Trans. Geosci. Remote Sens. 43, 2376–2385.
Palsson, F., Sveinsson, J.R., Ulfarsson, M.O., 2013. A new pansharpening algorithm based on total variation. IEEE Geosci. Remote Sens. Lett. 11, 318–322.
Pastina, D., Lombardo, P., Farina, A., Daddi, P., 2001. Super-resolution of polarimetric SAR images of a ship. In: IGARSS 2001. Scanning the Present and Resolving the Future. Proceedings. IEEE 2001 International Geoscience and Remote Sensing Symposium (Cat. No. 01CH37217). IEEE, pp. 2343–2345.
Picone, D., Restaino, R., Vivone, G., Addesso, P., Dalla Mura, M., Chanussot, J., 2017. Band assignment approaches for hyperspectral sharpening. IEEE Geosci. Remote Sens. Lett. 14, 739–743. https://doi.org/10.1109/lgrs.2017.2677087
Qin, P., Huang, H., Tang, H., Wang, J., Liu, C., 2022. Mustfn: a spatiotemporal fusion method for multi-scale and multi-sensor remote sensing images based on a convolutional neural network. Int. J. Appl. Earth Obs. Geoinf. 115, 103113.
Qu, Y., Baghbaderani, R.K., Qi, H., Kwan, C., 2020. Unsupervised pansharpening based on self-attention mechanism. IEEE Trans. Geosci. Remote Sens. 59, 3192–3208.
Ranchin, T., Wald, L., 2000. Fusion of high spatial and spectral resolution images: the Arsis concept and its implementation. Photogramm. Eng. Remote Sens. 66, 49–61.
Shao, Z., Cai, J., 2018. Remote sensing image fusion with deep convolutional neural network. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11, 1656–1669.
Shen, H., Lin, L., Li, J., Yuan, Q., Zhao, L., 2020. A residual convolutional neural network for polarimetric SAR image super-resolution. ISPRS J. Photogramm. Remote Sens. 161, 90–108.
Shen, H., Meng, X., Zhang, L., 2016. An integrated framework for the spatio-temporal-spectral fusion of remote sensing images. IEEE Trans. Geosci. Remote Sens. 54, 7135–7148. https://doi.org/10.1109/TGRS.2016.2596290
Simoes, M., Bioucas-Dias, J., Almeida, L.B., Chanussot, J., 2014. A convex formulation for hyperspectral image superresolution via subspace-based regularization. IEEE Trans. Geosci. Remote Sens. 53, 3373–3388.
Song, H., Huang, B., 2013. Spatiotemporal satellite image fusion through one-pair image learning. IEEE Trans. Geosci. Remote Sens. 51, 1883–1896. https://doi.org/10.1109/TGRS.2012.2213095
Song, H., Liu, Q., Wang, G., Hang, R., Huang, B., 2018. Spatiotemporal satellite image fusion using deep convolutional neural networks. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11, 821–829.

Sui, L., Li, L., Li, J., Chen, N., Jiao, Y., 2019. Fusion of hyperspectral and multispectral images based on a Bayesian nonparametric approach. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12, 1205–1218. https://doi.org/10.1109/JSTARS.2019.2902847

Suwa, K., Iwamoto, M., 2006. A two-dimensional bandwidth extrapolation technique for polarimetric synthetic aperture radar images. IEEE Trans. Geosci. Remote Sens. 45, 45–54.

Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. Econ. Geogr. 46, 234–240.

Toker, A., Kondmann, L., Weber, M., Eisenberger, M., Camero, A., Hu, J., Hoderlein, A.P., Şenaras, Ç., Davis, T., Cremers, D., et al., 2022. Dynamicearthnet: daily multi-spectral satellite dataset for semantic change segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21158–21167.

Wald, L., 2000. Quality of high resolution synthesised images: is there a simple criterion? In: Third Conference Fusion of Earth Data: Merging Point Measurements, Raster Maps and Remotely Sensed Images. SEE/URISCA, pp. 99–103.

Wang, D., Zhang, P., Bai, Y., Li, Y., 2022. Metapan: unsupervised adaptation with meta-learning for multispectral pansharpening. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Wang, Q., Atkinson, P.M., 2018. Spatio-temporal fusion for Daily Sentinel-2 images. Remote Sens. Environ. 204, 31–42.

Wang, Q., Tang, Y., Ge, Y., Xie, H., Tong, X., Atkinson, P.M., 2023. A comprehensive review of spatial-temporal-spectral information reconstruction techniques. Sci. Remote Sens. 100102.

Wang, W., Zhou, Z., Liu, H., Xie, G., 2021. Msdrn: pansharpening of multispectral images via multi-scale deep residual network. Remote Sens. 13, 1200. https://doi.org/10.3390/rs13061200

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13, 600–612.

Wei, Q., Bioucas-Dias, J., Dobigeon, N., Tourneret, J.-Y., Chen, M., Godsill, S., 2016. Multiband image fusion based on spectral unmixing. IEEE Trans. Geosci. Remote Sens. 54, 7236–7249. https://doi.org/10.1109/tgrs.2016.2598784

Wei, Q., Dobigeon, N., Tourneret, J.-Y., 2015. Fast fusion of multi-band images based on solving a sylvester equation. IEEE Trans. Image Process. 24, 4109–4121.

Wei, Y., Yuan, Q., Shen, H., Zhang, L., 2017. Boosting the accuracy of multispectral image pansharpening by learning a deep residual network. IEEE Geosci. Remote Sens. Lett. 14, 1795–1799.

Wu, M., Niu, Z., Wang, C., Wu, C., Wang, L., 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.

Xiao, J.-L., Huang, T.-Z., Deng, L.-J., Wu, Z.-C., Wu, X., Vivone, G., 2023. Variational pansharpening based on coefficient estimation with nonlocal regression. IEEE Trans. Geosci. Remote Sens..

Xie, Q., Zhou, M., Zhao, Q., Xu, Z., Meng, D., 2022. MHF-Net: an interpretable deep network for multispectral and hyperspectral image fusion. IEEE Trans. Pattern Anal. Mach. Intell. 44, 1457–1473. https://doi.org/10.1109/TPAMI.2020.3015691

Xu, S., Amira, O., Liu, J., Zhang, C.-X., Zhang, J., Li, G., 2020a. HAM-MFN: hyperspectral and multispectral image multiscale fusion network with rap loss. IEEE Trans. Geosci. Remote Sens. 58, 4618–4628. https://doi.org/10.1109/TGRS.2020.2964777

Xu, Y., Huang, B., Xu, Y., Cao, K., Guo, C., Meng, D., 2015. Spatial and temporal image fusion via regularized spatial unmixing. IEEE Geosci. Remote Sens. Lett. 12, 1362–1366. https://doi.org/10.1109/LGRS.2015.2402644

Xu, Y., Wu, Z., Chanussot, J., Comon, P., Wei, Z., 2020b. Nonlocal coupled tensor CP decomposition for hyperspectral and multispectral image fusion. IEEE Trans. Geosci. Remote Sens. 58, 348–362. https://doi.org/10.1109/TGRS.2019.2936486

Yang, J., Fu, X., Hu, Y., Huang, Y., Ding, X., Paisley, J., 2017. Pannet: a deep network architecture for pan-sharpening. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5449–5457.

Yang, J., Zhao, Y.-Q., Chan, J.C.-W., 2018. Hyperspectral and multispectral image fusion via deep two-branches convolutional neural network. Remote Sens. 10, https://doi.org/10.3390/rs10050800

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. Remote Sens. 59, 1808–1822. https://doi.org/10.1109/TGRS.2020.2999943

Yokoya, N., Yairi, T., Iwasaki, A., 2011. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. IEEE Trans. Geosci. Remote Sens. 50, 528–537.

Yuan, Q., Wei, Y., Meng, X., Shen, H., Zhang, L., 2018. A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11, 978–989.

Zeng, D., Hu, Y., Huang, Y., Xu, Z., Ding, X., 2016. Pan-sharpening with structural consistency and 1/2 gradient prior. Remote Sens. Lett. 7, 1170–1179.

Zhang, F., Zhang, K., Sun, J., Wang, J., Bruzzone, L., 2024a. Drformer: learning disentangled representation for pan-sharpening via mutual information- based transformer. IEEE Trans. Geosci. Remote Sens. 62, 1–15. https://doi.org/10.1109/tgrs.2023.3339650

Zhang, H., Ma, J., 2021. GTP-PNet: a residual learning network based on gradient transformation prior for pansharpening. ISPRS J. Photogramm. Remote Sens. 172, 223–239.

Zhang, L., Zou, B., Hao, H., Zhang, Y., 2011. A novel super-resolution method of Polsar images based on target decomposition and polarimetric spatial correlation. Int. J. Remote Sens. 32, 4893–4913.

Zhang, X., Huang, W., Wang, Q., Li, X., 2020. SSR-NET: spatial–spectral reconstruction network for hyperspectral and multispectral image fusion. IEEE Trans. Geosci. Remote Sens. 59, 5953–5965.

Zhang, X., Li, S., Tan, Z., Li, X., 2024b. Enhanced wavelet based spatiotemporal fusion networks using cross-paired remote sensing images. ISPRS J. Photogramm. Remote Sens. 211, 281–297.

Zhang, Y., He, M., 2007. Multi-spectral and hyperspectral image fusion using 3-d wavelet transform. J. Electron. (China) 24, 218–224. https://doi.org/10.1007/s11767-005-0232-5

Zhang, Y., Liu, C., Sun, M., Ou, Y., 2019. Pan-sharpening using an efficient bidirectional pyramid network. IEEE Trans. Geosci. Remote Sens. 57, 5549–5563.

Zheng, Y., Li, J., Li, Y., Guo, J., Wu, X., Chanussot, J., 2020. Hyperspectral pansharpening using deep prior and dual attention residual network. IEEE Trans. Geosci. Remote Sens. 58, 8059–8076. https://doi.org/10.1109/tgrs.2020.2986313

Zheng, Z., Zhong, Y., Ma, A., Zhang, L., 2024. Single-temporal supervised learning for universal remote sensing change detection. Int. J. Comput. Vis. 1–21.

Zhong, J., Yang, B., Huang, G., Zhong, F., Chen, Z., 2016. Remote sensing image fusion with convolutional neural network. Sens. Imaging 17, 1–16.

Zhong, Y., Wu, X., Deng, L.-J., Cao, Z., Ssdiff: Spatial-spectral integrated diffusion model for remote sensing pansharpening, arXiv prepr. arXiv:2404.11537, 2024.

Zhou, H., Liu, Q., Weng, D., Wang, Y., 2022. Unsupervised cycle-consistent generative adversarial networks for pan sharpening. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

Zhou, Y., Feng, L., Hou, C., Kung, S.-Y., 2017. Hyperspectral and multispectral image fusion based on local low rank and coupled spectral unmixing. IEEE Trans. Geosci. Remote Sens. 55, 5997–6009. https://doi.org/10.1109/tgrs.2017.2718728

Zhu, H., Cao, P., Jiao, L., Li, X., Hou, B., Yi, X., Zhao, W., Ma, W., 2026. A progressive semi-distillation model for dual-source remote sensing image classification. IEEE Trans. Cybern. 56(1), 67–80.

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.

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.

Zhukov, B., Oertel, D., Lanzl, F., Reinhackel, G., 1999. Unmixing-based multisensor multiresolution image fusion. IEEE Trans. Geosci. Remote Sens. 37, 1212–1226.

Zou, B., Hao, H., Guo, X., 2008. Super-resolution of polarimetric SAR images based on target decomposition and polarimetric spatial correlation. In: IGARSS 2008-2008 IEEE International Geoscience and Remote Sensing Symposium. IEEE. pp. II–911.