

基于深度学习的多源遥感图像时空融合研究综述

孔维泉^{1,2}, 钱育蓉^{1,2,3}, 白璐^{1,2}, 范迎迎^{1,2}, 骆森³, 公维军⁴

1. 新疆大学 计算机科学与技术学院, 乌鲁木齐 830046;

2. 丝路多语言认知计算国际合作联合实验室, 乌鲁木齐 830046;

3. 新疆大学 软件学院 新疆维吾尔自治区大数据与智能软件工程研究中心, 乌鲁木齐 830091;

4. 河西学院 信息技术与传媒学院, 张掖 734000

摘要: 受限于扫描幅宽和像素大小之间的权衡关系, 遥感图像空间分辨率和时间分辨率存在相互制约, 使得单一卫星传感器难以获取高时空分辨率遥感影像。因此, 多源遥感图像时空融合技术成为解决这一难题的关键途径。然而, 受限于理想物理模型和复杂现实场景的差距, 传统融合方法在处理非匀质融合场景时效果不理想。深度学习的兴起带来了如卷积神经网络、扩散模型等前沿时空融合技术。为及时掌握本领域的发展前沿, 本文重点分析了基于深度学习的多源遥感图像时空融合前沿技术。首先, 系统梳理了传统遥感图像时空融合方法的基本原理, 并深入分析了其优势与局限性; 其次, 按照算法原理和特点, 将前沿深度学习技术分为5类进行整理, 并全面总结了其在时空融合任务中的应用进展; 然后, 归纳并介绍了用于遥感图像时空融合的开源数据集、深度学习算法和常用的性能评估指标, 并通过在 CIA、DX 和 SW 数据集上的定量与定性实验, 开展了时空融合算法验证与对比分析。最后, 综合各类时空融合算法的发展与实验分析结果, 总结了现阶段遥感图像时空融合领域面临的主要挑战和未来潜在的发展方向。

关键词: 时空融合, 多源遥感图像, 深度学习, 算法原理, 技术分类, 数据集

中图分类号: TP751/TP391/TP183/P2

引用格式: 孔维泉, 钱育蓉, 白璐, 范迎迎, 骆森, 公维军. 2026. 基于深度学习的多源遥感图像时空融合研究综述. 遥感学报, 30(6): 1544–1563

Kong W Q, Qian Y R, Bai L, Fan Y Y, Luo S and Gong W J. 2026. Review of deep learning-based spatiotemporal fusion of multisource remote sensing images. National Remote Sensing Bulletin, 30(6): 1544–1563 [DOI: 10.11834/jrs.20265179]

1 引言

伴随着遥感技术的迅速发展, 现有星载传感器众多, 如中分辨率成像光谱仪 MODIS (moderate resolution imaging spectroradiometer)、Landsat、Sentinel 和高分 (GF) 系列等, 受传感器硬件技术和预算的限制, 遥感图像的空间分辨率和时间分辨率存在相互制约的关系 (黄波和姜晓璐, 2021)。例如, 空间分辨率为 30 m 的 Landsat 多光谱数据需要经历至少 16 d 的重访周期 (Roy 等, 2014)。所以类 Landsat 影像难以及时捕捉扰动事件

(Gao 等, 2006)。MODIS 具有较高的时间分辨率 (1–2 d), 但空间分辨率较低 (250–1000 m) (Justice 等, 1998)。类 MODIS 图像在监测异质景观中的生态系统和地面覆盖变化时效果有限, 与 Landsat 图像呈现出很强的互补性。而在土地利用和土地覆盖变化监测、植物物候监测、环境动态监测、农业评估等实际应用中, 单卫星传感器所获得的特征较为单一, 不能满足实际需求, 需要具有高时空分辨率的遥感图像 (李云飞等, 2022)。因此, 多源遥感图像时空融合 STF (Spatio-Temporal Fusion) 技术受到研究人员的广

收稿日期: 2025-05-27; 预印本: 2026-03-31

基金项目: 国家自然科学基金 (编号: 62266043, 62566021); 新疆维吾尔自治区杰出青年科学基金 (编号: 2023D01E01); 新疆维吾尔自治区青年拔尖人才项目 (编号: 2023TSYCCX0043); 天山创新团队计划 (编号: 2023D14012); 新疆维吾尔自治区重点研发专项 (编号: 2023B01029-1, 2023B01029-2); 新疆维吾尔自治区研究生科研创新项目 (编号: XJ2025G089, XJ2025G036); 新疆维吾尔自治区自然科学基金 (编号: 2022D01B123); 甘肃省科技计划基础研究计划自然科学基金 (编号: 25JRRG037)

第一作者简介: 孔维泉, 研究方向为多源遥感图像时空融合。E-mail: kwq_312218@163.com

通信作者简介: 钱育蓉, 研究方向为网络计算和遥感图像处理。E-mail: qyr@xju.edu.cn

泛关注。表1呈现了在遥感图像STF领域中，用于获取高空间分辨率和高时间分辨率图像的常见传感器类型及其特点。

表1 STF常用卫星数据及参数

卫星	传感器	空间分辨率/m	时间分辨率/d
Landsat 5	TM	30	16
Landsat 7	ETM+	30	16
Landsat 8	OLI	30	16
	TIRS		
GF-1	PMS	8—16	4
	WFV		
GF-2	PMS	4	5
GF-6	PMS	8—16	4
	WFV		
Sentinel-2	MSI	10	5
Sentinel-3	OLCI	300	2
Terra/Aqua	MODIS	250—1000	1

STF旨在结合来自不同传感器的数据，通过融合低空间、高时间分辨率图像与高空间、低时间分辨率图像，消除单一信息源所具有的不确定性和不完整性（Zhang等，2021a），生成高空间和高时间分辨率兼并的遥感图像，以解决单一卫星传

感器时空分辨率难以权衡的问题（Wang等，2023；Zhang等，2021a）。高时空分辨率图像用于后续的解译、监测、跟踪等任务（王志浩和钱运涛，2024），融合过程如图1所示。

深度学习的快速发展推动了多源遥感图像时空融合方法的不断创新，但相关综述文献仍较为稀缺，未能紧跟技术前沿。本文对涉及时空融合综述的文献进行了整理与汇总，如表2所示。多数文献主要讨论了全色锐化 Pan-sharpening (Panchromatic sharpening)、高光谱—全色锐化 (Hyperspectral Pan-sharpening)、高光谱—多光谱融合 (Hyperspectral-Multispectral Fusion)、高光谱—激光雷达 (Hyperspectral-LiDAR) 等多源数据融合方法，以及传统时空融合方法（Li等，2020a，2022a；Ghamisi等，2019；Chen等，2023），而对前沿深度学习方法的系统综述仍然相对匮乏。这种全面综述的缺乏，阻碍了基于深度学习的STF方法的进一步发展。为此，本文调查、整理并深入分析了近年来基于深度学习的遥感图像STF前沿方法与成果，并对各类深度学习融合方法进行了对比分析，探讨了未来的研究趋势和发展前景，最终形成综述。此外，本文还整理了STF领域的数据集及评估指标，具体贡献可归纳为3点。

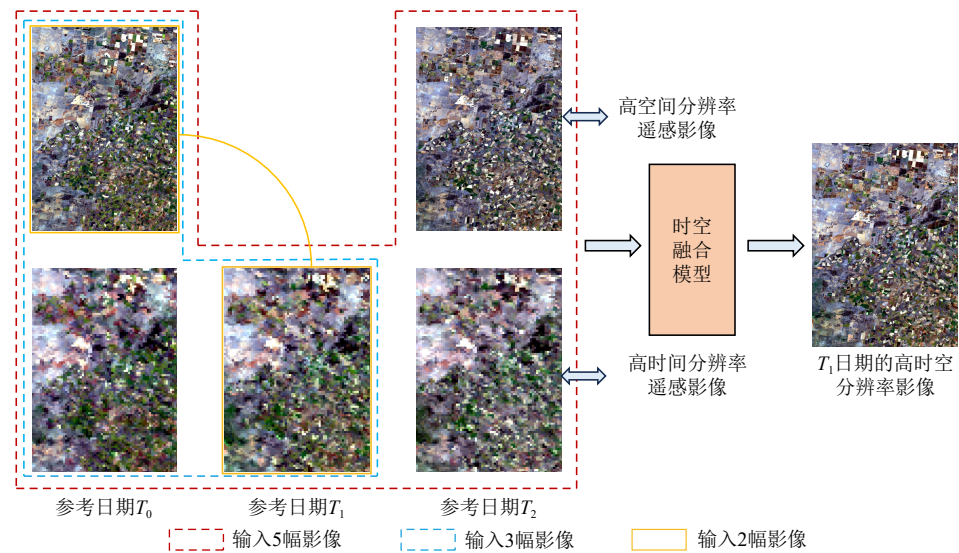


图1 不同输入数量下的遥感影像时空融合方法示意图

Fig. 1 Schematic diagram of remote-sensing image spatiotemporal-fusion methods with different input counts

(1) 对传统STF方法的原理、优势和局限性进行了综述和对比。(2) 对基于深度学习的遥感图像STF前沿方法进行系统分类、梳理和总结，并在 CIA、DX 和 SW 数据集上对深度学习方法进行对比实验分析。(3) 综述了当前遥感图像STF研究中存在的问题和下一步发展方向，为未来研究方向提供参考。

表2 国内外遥感图像STF相关综述汇总
Table 2 Summary of domestic and international reviews on STF of remote sensing images

论文名称	来源期刊	IF(2026年)
Comparison of spatiotemporal fusion models: A review (Chen 等, 2015)	Remote Sensing	4.1
多源卫星遥感影像时空融合研究的现状及展望 (黄波和赵涌泉, 2017)	《测绘学报》	5.17
Spatiotemporal fusion of multisource remote sensing data: Literature survey, taxonomy, principles, applications, and future directions (Zhu 等, 2018)	Remote Sensing	4.1
遥感数据时空融合研究进展及展望 (董文全和蒙继华, 2018)	《自然资源遥感》	3.73
Spatiotemporal image fusion in remote sensing (Belgiu 和 Stein, 2019)	Remote Sensing	4.1
Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art (Ghamisi 等, 2019)	IEEE Geoscience and Remote Sensing Magazine	16.4
Spatio-temporal fusion for remote sensing data: an overview and new benchmark (Li 等, 2020a)	Science China-Information Sciences	7.6
Image fusion meets deep learning: A survey and perspective (Zhang 等, 2021a)	Information Fusion	15.5
Spatially continuous and high-resolution land surface temperature product generation: A review of reconstruction and spatiotemporal fusion techniques (Wu 等, 2021)	IEEE Geoscience and Remote Sensing Magazine	16.4
Deep learning in multimodal remote sensing data fusion: A comprehensive review (Li 等, 2022a)	International Journal of Applied Earth Observation and Geoinformation	8.6
遥感地表温度产品时空融合方法研究综述 (李冉 等, 2022)	《遥感学报》	4.79
遥感图像时空融合综述 (杨广奇 等, 2022)	《计算机工程与应用》	5.17
多源图像融合方法的研究综述 (张丽霞 等, 2022)	《计算机工程与科学》	2.54
Spatio-temporal fusion methods for spectral remote sensing: A comprehensive technical review and comparative analysis (Swain 等, 2024)	Tropical Ecology	1.7
Spatiotemporal fusion for spectral remote sensing: A statistical analysis and review (Chen 等, 2023)	Journal of King Saud University-Computer and Information Sciences	6.1
Review of pixel-level remote sensing image fusion based on deep learning (Wang 等, 2023)	Information Fusion	15.5
Comprehensive analysis of temporal-spatial fusion from 1991 to 2023 using bibliometric tools (Cui 等, 2024b)	Atmosphere	2.3
数据融合视角下的遥感参量空间降尺度 (景映红 等, 2024)	《武汉大学学报(信息科学版)》	4.75
Recent advances in deep learning-based spatiotemporal fusion methods for remote sensing images (Lian 等, 2025)	Sensors	3.5

注: IF为期刊影响因子。

2 相关工作

随着计算机技术的发展, 深度学习技术在遥感领域受到越来越多的关注, 尤其是卷积神经网络 CNN (Convolutional Neural Network) 的出现, 极大的促进了深度学习在遥感领域的发展, 并涌现了一系列基于深度学习的多源遥感图像 STF 技术。根据 Wang 等 (2023) 和 Zhu 等 (2018) 的研究, STF 方法可分为传统方法和基于深度学习的方法。在不同的假设和应用场景下, 传统 STF 方法可以分为 5 类 (Li 等, 2020a; Chen 等, 2015; Swain 等, 2024; Zhang 等, 2024b): 解混 (unmixing) 方法、权函数 (weight function) 方法、贝叶斯 (Bayesian) 方法、字典学习 (dictionary learning) 方法和混合 (hybrids) 方法。根据前沿研究, 本文将前沿深度

学习方法分为 6 类: CNN、Transformer、生成对抗网络 GAN (Generative Adversarial Network)、循环神经网络 RNN (Recurrent Neural Network)、小波变换 (wavelet transform) 和扩散模型 (diffusion model), 并对每种方法的结构、原理和特点进行了综述。图 2 展示了传统 STF 方法以及常见的 CNN、Transformer 和 GAN 方法的原理结构。

2.1 传统 STF 方法

2.1.1 光谱解混方法

光谱解混方法基于线性光谱混合模型, 在亚像素尺度上提取端元 (即地物纯像元) 及对应的丰度 (即各端元的比例) (Belgiu 和 Stein, 2019)。其核心思想是假设低空间分辨率图像中的像素是

高空间分辨率端元的线性组合, 将低分辨率影像

的混合光谱分解成多个端元的贡献，并结合高分辨率数据推测端元在时空上的变化，从而实现高质量的时空融合。代表性方法包括：多传感器多分辨率技术 MMT (Multi-sensor Multi-resolution Technique) (Zhukov 等, 1999)、基于对象的空间解混模型 OBSUM (Object-Based Spatial Unmixing Model) (Guo 等, 2024) 和光谱自相关时空融合模型 FIRST (spatiotemporal fusion incorrpoting spectral

autocorrelation) (Liu 等, 2022) 等方法。但光谱解混存在不可避免的共线性和块效应问题，限制了解混性能 (Wu 等, 2015; Wang 等, 2021)。因此，一些研究引入新的约束条件以优化解混模型 (Wu 等, 2015)，并提出块去除策略以缓解块效应问题 (Wang 等, 2021)。尽管该方法在一定程度上受到上述问题的限制，但其在建模粗细分辨率影像的空间异构关系方面仍然发挥着重要作用。

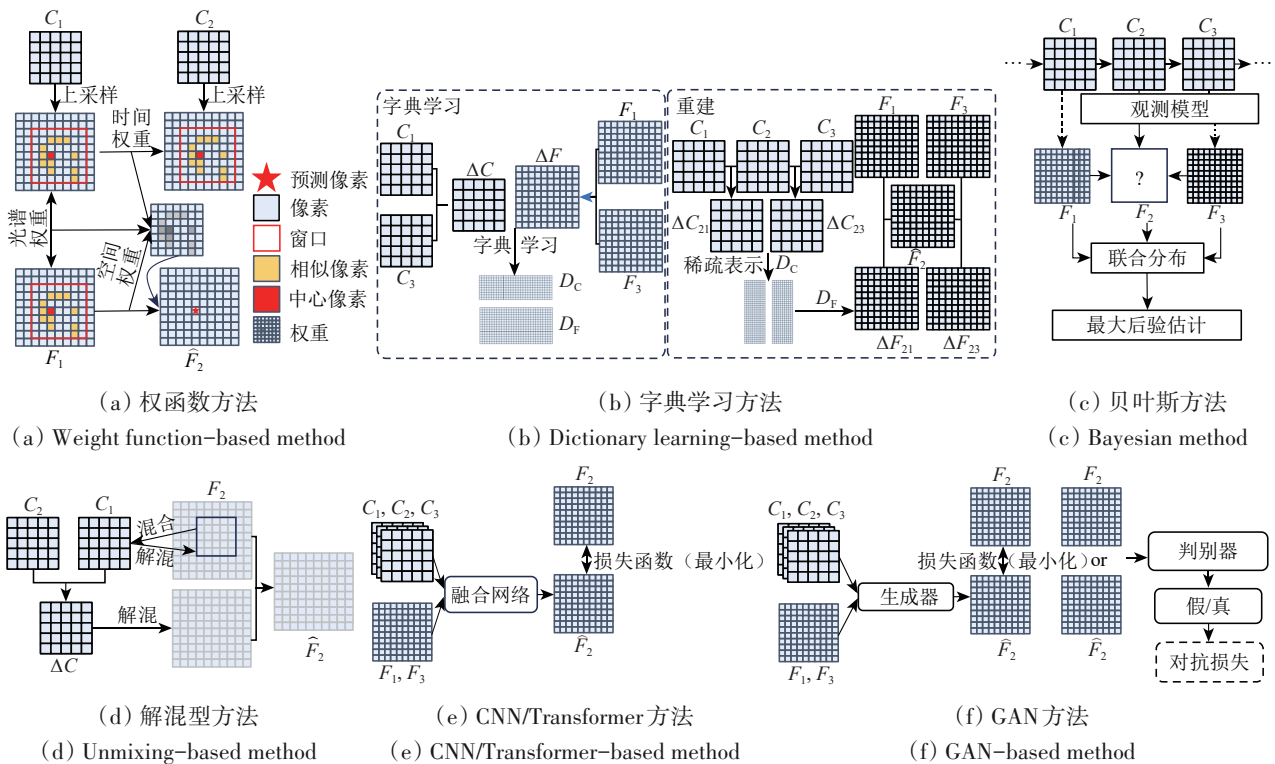


图2 常见STF方法的原理结构

Fig. 2 Principle structures of common STF methods

2.1.2 权函数方法

基于权函数的方法通过构造权重函数，对粗、细图像的对应像素进行加权，重建高时空图像，如时空自适应反射率融合模型 STARFM (Spatial and Temporal Adaptive Reflectance Fusion Model) (Gao 等, 2006)，以及基于 STARFM 改进的增强型时空自适应反射率融合模型 (ESTARFM) (Zhu 等, 2010) 和 STARFM-SI (STARFM under Simplified Input modality) (Wu 等, 2020)。但该类方法通常假设粗像元和精细像元的时间变化是一致的，利用粗图像的时间信息和细图像的空间信息来生成时空上的高分辨率地表反射率值。因此，该方法在均匀景观中表现良好，但在地物分布复杂的异

质区域，预测精度下降。

2.1.3 贝叶斯方法

基于贝叶斯估计的方法通常将 STF 视为最大后验概率问题，最大化预测日期下精细影像的条件概率。代表性方法包括：贝叶斯最大熵 BME (Bayesian Maximum Entropy) (Li 等, 2013)、归一化差异植被指数—贝叶斯时空融合模型 NDVI-BSFM (Normalized Difference Vegetation Index—Bayesian Spatiotemporal reflectance Fusion Model) (Liao 等, 2016) 和统一时空谱融合框架 STS (integrated framework for Spatio-Temporal-Spectral fusion) (Shen 等, 2016)。但该类方法主要对参考数据与预测数据之间的关系进行建模，难以应对土地类型变化，限

制了融合性能, 近年研究较少。

2.1.4 字典学习方法

字典学习方法通过稀疏表示建立粗—细分辨率图像间的映射关系, 并构造字典对学习捕获空间结构特征。然后, 通过相应的字典和粗图像的稀疏表示重构融合后的图像 (Song 和 Huang, 2013)。代表性方法包括: 基于稀疏表征的时空反射融合模型 SPSTFM (sparse representation-based spatiotemporal reflectance fusion model) (Huang 和 Song, 2012)。后续研究中, 基于非局部张量稀疏表示的时空融合半耦合字典学习方法 SCDNTSR (nonlocal tensor sparse representation-based semicoupled dictionary learning approach) (Peng 等, 2022) 扩展了张量稀疏表示, 并引入半耦合映射, 改进了字典学习方式, 进一步提高了融合效果。

2.1.5 混合方法

混合方法结合解混方法、贝叶斯方法、权函数方法以及深度学习方法等优势, 以提升融合效果与泛化能力, 但同时也增加了算法复杂性, 限制了其大规模应用。时空反射解混模型 STRUM (Spatial and Temporal Reflectance Unmixing Model) (Gevaert 和 García-Haro, 2015) 首次将 STARFM 和解混方法相结合, 并在解混阶段引入贝叶斯方法, 结合先验光谱信息约束解混过程。另一个广受关注的方法——灵活时空数据融合 FSDAF (flexible spatiotemporal data fusion) 方法 (Zhu 等, 2016) 结合解混方法和权函数方法, 获得了更精细的融合结果, 但在复杂地表动态监测中适用性较差。随后, 出现了诸多基于 FSDAF 的改进融合方法 (Guo 等, 2020; Li 等, 2020b; Shi 等, 2022; Liu 等, 2019, 2024; Xu 等, 2024), 这些方法充分考虑地表异质性特征、土地类型变化及类内可变性等问题, 在实际任务中表现出较强的稳健性。各传统方法的原理见图 2 (a) —图 2 (d)。

总体而言, 传统 STF 方法在均匀区域表现良好, 且具有计算高效、理论清晰和易于解释等优势。然而, 受限于物理模型与现实的差距 (如粗、细图像的非线性关系)、不合理的先验假设 (如线性解混、稀疏性) 以及对手工特征的依赖 (如字典) 等因素, 面对高度非线性关系或大规模多源数据时, 其性能仍受限。部分方法在非匀质区域融合性能下降, 难以跨时空尺度有效融合多源遥

感数据。

尽管传统 STF 方法存在一定局限, 但在多种场景中仍具有应用价值。权函数类方法在地物分布相对均匀的区域可快速得到较好结果; 解混方法能够在亚像素尺度刻画端元变化, 适合需要精细光谱分析的任务; 贝叶斯方法在有先验知识时可获得较高精度; 字典学习方法能够利用稀疏表示捕获空间结构; 混合方法如 FSDAF 系列则在复杂异质场景中表现稳健。因此, 在计算资源有限或缺乏大规模训练数据的情况下, 传统方法仍是实用的选择。

2.2 深度学习 STF 方法

2.2.1 CNN

深度学习的快速发展为 STF 技术提供了新范式, 其中基于 CNN 的模型凭借强大的空间特征提取能力与自适应多尺度信息融合机制, 能够精准捕捉地表动态变化并生成高精度的时空融合结果, 方法原理见图 2 (e)。深度卷积神经网络时空融合方法 STFCNN (SpatioTemporal Fusion using Deep Convolutional Neural Networks) (Song 等, 2018) 首次将深度学习方法应用于多源遥感图像 STF 任务中, 通过对输入粗图像和精细图像之间的非线性映射关系进行建模, 并利用超分辨率 CNN 克服粗、细图像的分辨率差异, 来提高非均质区域的预测精度, 见图 3 (a)。随后, 深度卷积时空融合网络 DCSTFN (Deep Convolutional Spatio-Temporal Fusion Network) (Tan 等, 2018) 通过整合卷积层和反卷积层, 构建了双流卷积网络, 直接建立从输入到输出的非线性映射, 提高了融合的准确性和鲁棒性, 见图 3 (b)。增强深度卷积时空融合网络 EDCSTFN (Enhanced Deep Convolutional Spatio-Temporal Fusion Network) (Tan 等, 2019) 设计了残差编码器以动态建模时空变化, 并提出复合损失函数, 解决了 DCSTFN 仅用 L2 损失导致的图像模糊问题, 显著提升了高频细节的保留能力, 见图 3 (c)。然而上述方法中的特征级融合会导致高频细节损失。为了解决这个问题, 多协同深度卷积神经网络 MCDNet (multicooperative deep convolutional neural network) (Li 等, 2021b) 采用超分辨网络、差分重建网络和协同训练网络进行协同训练, 并利用多尺度感知与扩张卷积增强特征提取能力, 扩大感受野, 显著提升了高频细节保留效

果, 见图3 (d)。后续研究提出的时空融合残差密集网络 STFRDN (Spatio-Temporal Fusion Residual Dense Network) (Erdem和Avdan, 2023)、双流卷积神经网络 STFNet (two-stream convolutional neural network) (Fu等, 2024)、两阶段粗到细时空融合方法 SIFnet (two-stage coarse-to-fine spatiotemporal fusion approach) (Ran等, 2024) 和灵活时空数据

融合和残差通道注意力网络 RCAN-FSDAF [flexible spatiotemporal data fusion (FSDAF) and residual channel attention network (RCAN)] (Cui等, 2024a) 等网络, 通过进一步设计残差结构 (图4) 和注意力机制 (图5), 显著改进了CNN融合模型, 提升了模型性能。

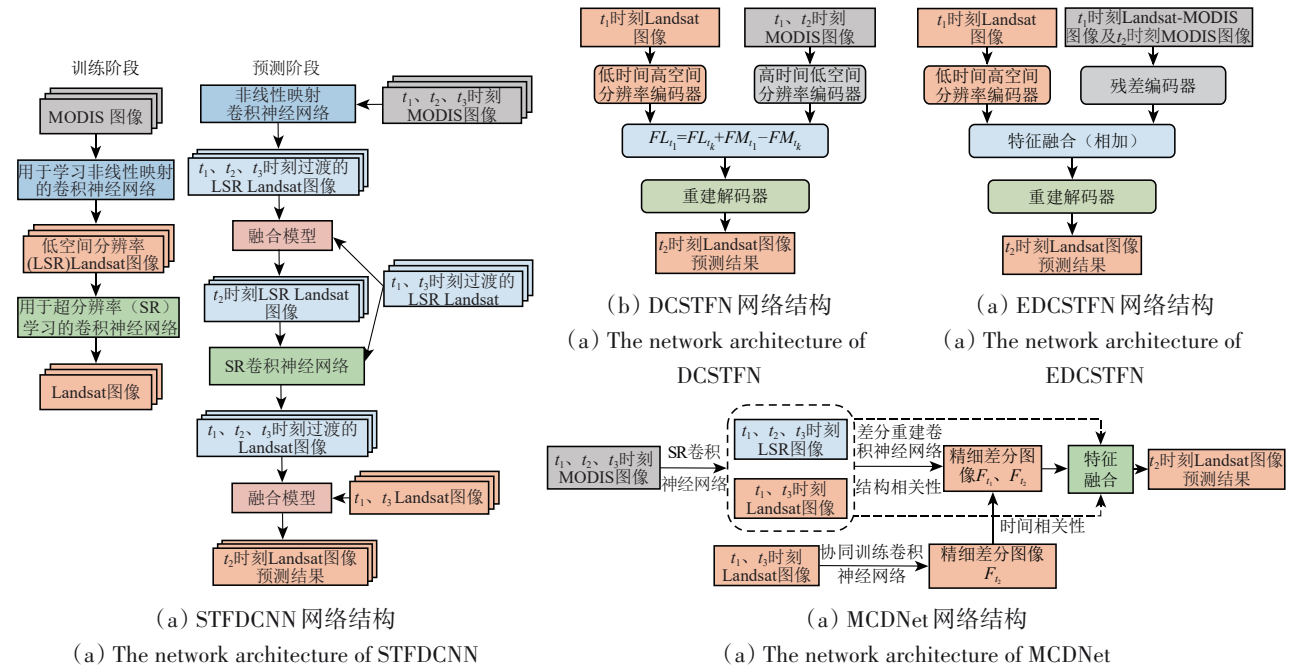


图3 基于CNN的STF网络结构

Fig. 3 CNN-based network architecture for STF

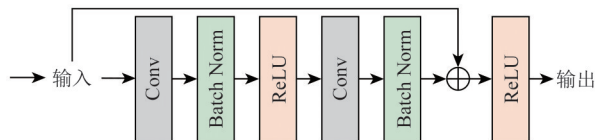


图4 卷积神经网络中典型残差网络结构

Fig. 4 Typical residual network structure in CNN

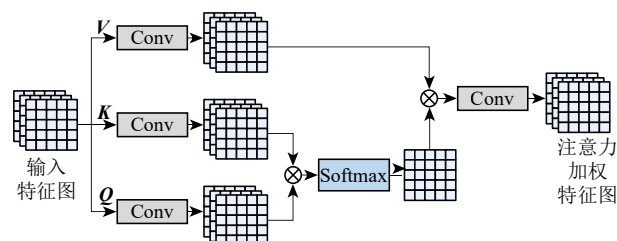


图5 卷积神经网络中注意力模块的示例(Q 、 V 、 K 分别表示查询、键、值3类向量)

Fig. 5 Example of an attention module in CNN (Q , V , and K denote query, key, and value vectors)

但上述基于CNN的方法均从空间域的角度出发, 直接从空间域提取像素信息, 忽略了频率信

息。最新研究基于交叉配对的增强小波时空融合网络 ECPW-STFN (Enhanced Cross-Paired Wavelet based Spatio-Temporal Fusion Networks) (Zhang等, 2024a) 提出了一种基于频域增强的卷积模型, 采用小波变换将图像特征从空间域映射到频率域, 从频率域的角度分离高、低频特征, 并分别进行处理, 取得了较为优异的融合效果。哈尔 (Haar) 小波变换的过程如图6所示, 其中低频分量 A 包含了图像整体的光谱和背景信息, 高频分量 V 、 H 、 D 包含了图像的纹理细节信息。

小波变换在计算机视觉领域受到了广泛关注, 如超分辨率 (Xu等, 2025)、遥感图像变化监测 (Wang等, 2025) 和语义分割 (Yang等, 2024) 等。ECPW-STFN首次将小波变换引入STF, 通过分离图像的高、低频特征进行针对性训练, 显著提升了复杂城市地物变化区域的细节保留能力, 其网络结构如图7所示。该方法采用交叉配对输入的方式, 将变化特征和参考特征分别映射到频率

域, 并设计小波损失来保留空间细节特征, 在减少图像依赖的前提下, 避免融合效果的下降。

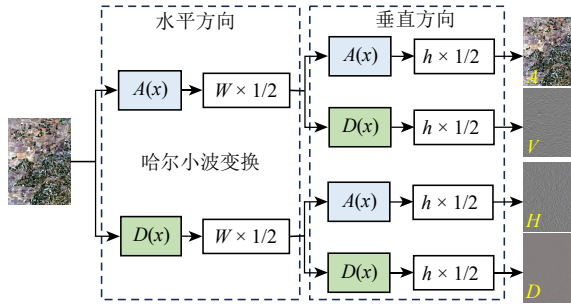


图6 哈尔小波变换频率特征分解的原理(W 为特征图宽度, h 为特征图高度)

Fig. 6 The principle of frequency feature decomposition of Haar wavelet transform (W and h denote the width and height of the feature map)

总体而言, CNN因其显著的特征提取能力和较高的训练效率, 仍是许多优秀的时空融合模型的基础。但其模型结构感受野受限, 在处理全局

信息和长距离依赖关系上存在一定的局限性, 并且在图像融合过程中会带来噪声。

2.2.2 Transformer

Transformer的引入为遥感图像STF提供了新的解决方案, 与CNN相比, 它能够学习全局变换信息, 动态建模像素级别的关系, 解决CNN在处理长距离依赖和全局信息时的不足的问题。Transformer通过自注意力机制 (self-attention) 为每个位置构造查询 (Query, Q)、键 (Key, K)、值 (Value, V) 3类向量, 先计算 Q 与所有 K 的相似性权重, 再用权重加权对应的值来获得该位置的表示, 从而建模图像中全局的长距离依赖关系。常用形式为

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

式中, d_k 表示 K 的维度。

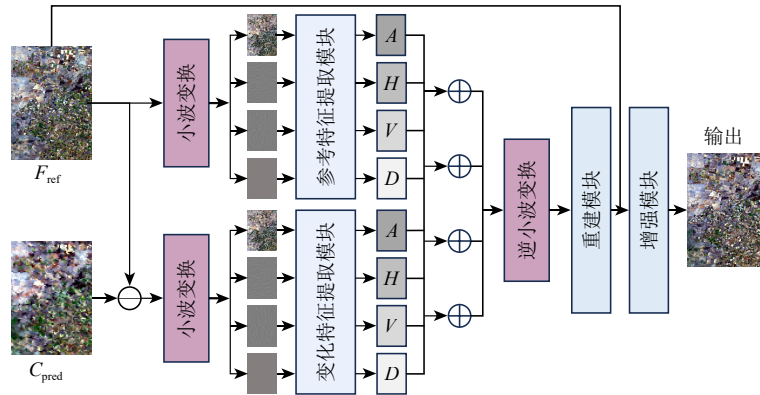


图7 ECPW-STFN网络结构图

Fig. 7 Network architecture of the ECPW-STFN

多流时空融合网络 MSNet (multi-stream spatiotemporal fusion network) (Li 等, 2021a) 首次将Transformer用于遥感图像STF, 并与CNN结合构成双流结构。MSNet通过CNN提取粗图像和细图像的时间信息和空间细节, 并利用Transformer编码器学习粗图像中局部和全局时间变化信息的关系, 最后采用平均加权策略进行融合, 从而有效的缓解了噪声敏感和异质区模糊问题, 见图8 (a)。增强多流时空融合网络 (EMSNet) (Li 等, 2022b) 设计卷积投影替代线性投影, 取消位置编码改进Transformer, 并结合扩张卷积提升融合精度与鲁棒性, 见图8 (b)。基于Swin Transformer的时空融合方法 SwinSTFM (spatiotemporal fusion based swin

transformer) (Chen 等, 2022) 使用两个Swin Transformer分别处理时空信息, 通过移位窗口机制实现跨窗口信息交互, 并引入反馈机制迭代优化网络模型, 增强对复杂非线性关系的建模能力, 见图8 (c)。基于纹理Transformer和CNN的多阶段时空融合网络 MSFusion (multi-stage spatiotemporal fusion based on texture Transformer and convolutional neural network) (Yang 等, 2022) 设计Texture Transformer改进Transformer, 并利用多尺度特征融合, 提高图像的细节信息和整体效果, 见图8 (d)。但Transformer忽略了遥感图像二维结构, 且计算量大, 后续研究基于大核注意力的多阶段时空融合网络 MLKNet (multi-stage spatiotemporal fusion

network based on a large kernel attention) (Jiang 等, 2024) 提出采用大核注意力取代 Transformer, 以保留遥感图像的二维结构并降低计算成本, 见图 8 (e)。基于编-解码 Transformer 结构的双流时空融合技术 STF-Trans (two-stream spatiotemporal fusion technique based on encoder-decoder Transformer architecture) (Benzenati 等, 2024)、STM-STFNet

(Qian 等, 2024) 等进一步证明了 Transformer 类的时空融合方法在克服卷积模型限制方面的有效性。

Transformer 虽擅长长距离依赖与全局信息建模, 但其标准自注意力存在二次复杂度开销。尽管 SwinSTFM 利用移位窗口机制有效降低了复杂度, 但仍在精细二维结构与效率间面临权衡。

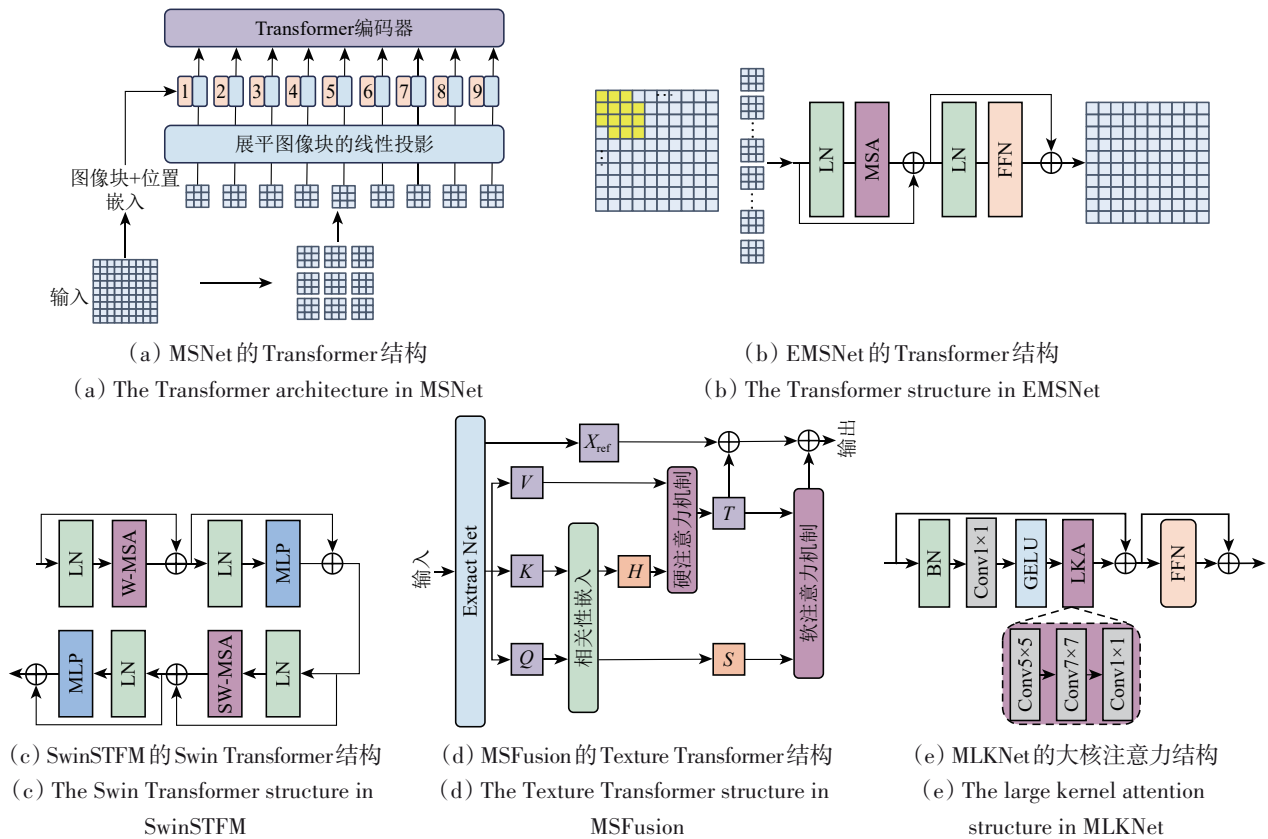


图8 基于Transformer的STF网络结构

Fig. 8 Transformer-based network architecture for STF

2.2.3 GAN

如图 2 (f) 所示, GAN 由生成器 (Generator, G) 与判别器 (Discriminator, D) 组成。生成器试图将随机噪声映射为逼真样本, 判别器学习区分真样本与生成样本。 G 与 D 对抗训练, 通过最小-最大博弈优化对抗损失。经典形式如下:

$$\min_G \max_D \mathbb{E}_{x \sim p_{\text{data}}} [\ln D(x)] + \mathbb{E}_{z \sim p_z} [\ln (1 - D(G(z)))] \quad (2)$$

式中, x 表示真实样本; p_{data} 为真实数据分布, z 表示噪声/隐向量, p_z 为噪声先验分布。

GAN 在图像增强和超分辨率重建方面取得了突出的成果, 并开始应用于遥感图像的时空融合

领域 (Xie 等, 2025)。时空融合生成对抗网络 STFGAN (spatiotemporal fusion using a gan) (Zhang 等, 2021b) 方法继承了基于超分辨率生成对抗网络 SRGAN (super resolution gan) (Ledig 等, 2017) 的超分辨率思想, 将时空融合作为一个超分辨率问题来解决, 采用端到端的两阶段图像融合生成对抗网络 IFGAN (two-stage framework with an end-to-end image fusion gan), 缓解了粗图像与精细图像在空间分辨率上的巨大差距, IFGAN 网络结构见图 9 (a)。多级特征融合生成对抗网络 MLFF-GAN (multilevel feature fusion with gan) (Song 等, 2022) 采用多级特征自适应融合框架, 利用自适应实例归一化与自注意力 AdaIN-AM (adaptive

instance normalization and self-attention module) 双模块耦合, 实现了融合过程中全局分布对齐与局部细节增强的联合优化, AdaIN-AM 网络结构见图 9 (b)。基于低时敏度的高精度时空融合生成对抗网络 HPLTS-GAN (high-precision spatiotemporal fusion generative adversarial network based on low temporal sensitivity) (Lei 等, 2024) 为解决时间数据依赖, 提出了一种自适应空间分布变换 ASDT (adaptive spatial distribution transformation) 模块, 通过结合粗分辨率图像的空间结构来增强细分辨率图像的时空一致性, ASDT 网络结构见图 9 (c)。使用条件生成对抗网络的灵活参考不敏感时空融合方法 GAN-STFM (flexible reference-insensitive spatiotemporal fusion model using conditional generative adversarial network) (Tan 等, 2022) 将参考图像作为 GAN 中的条件输入, 引入条件 GAN 和可切换归一化技术, 突破传统模型对多参考图

像及严格时间一致性的依赖, 如图 9 (d) 所示。此外, 基于多尺度和卷积块注意力块的时空融合生成对抗网络 MCBAM-GAN (the gan spatiotemporal fusion model based on multiscale and convolutional block attention module) (Liu 等, 2023)、自适应多尺度时空融合网络 AMS-STF (adaptive multiscale network for spatiotemporal fusion) (Pan 等, 2023) 等通过多尺度特征提取和注意力机制增强时空特征表达, 并改进生成器与判别器, 提升模型对突变地形的适应能力。

尽管 GAN 能生成清晰、具有自然纹理的影像, 主观视觉质量较高, 在遥感图像 STF 领域取得了优异的效果。但仍存在训练不稳定、易发生模式崩溃问题。且原生对抗损失难以直接保证像素或光谱精度, 需结合像素损失、光谱一致性约束或判别器设计以保持遥感数据的光谱/结构一致性。

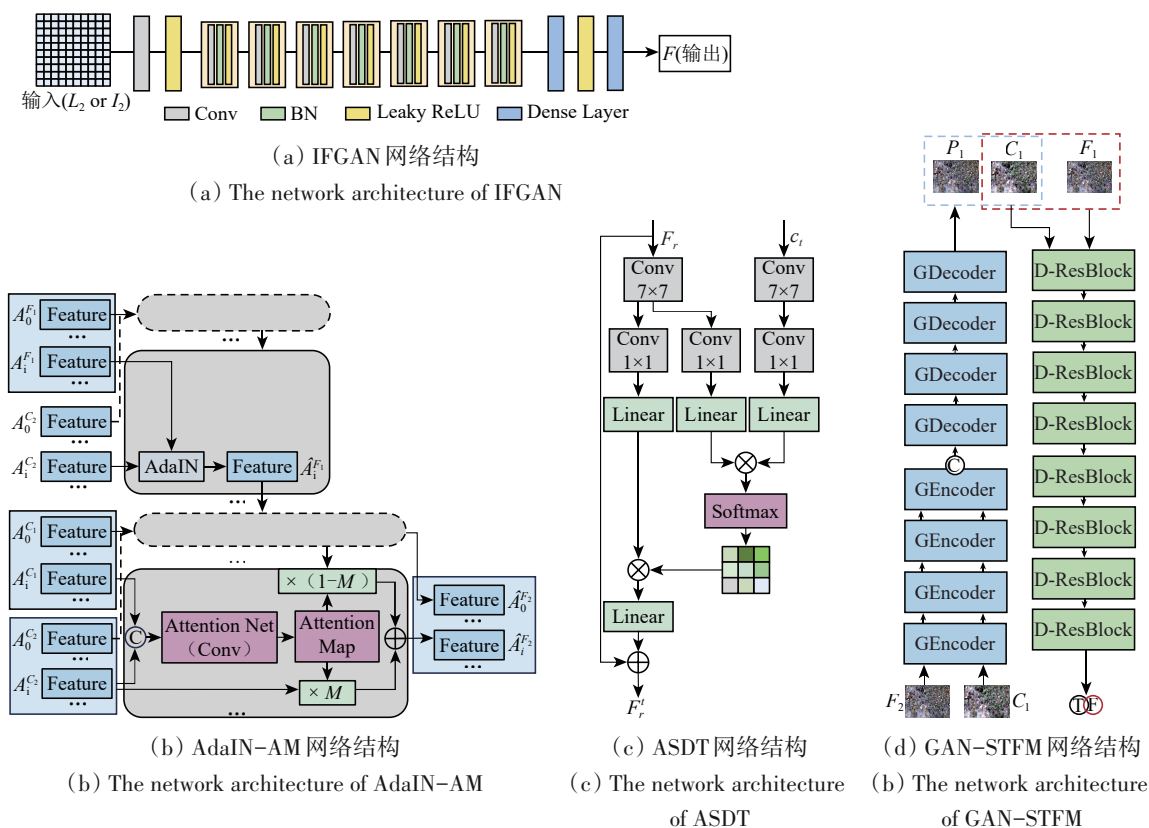


图 9 基于 GAN 的 STF 网络结构

Fig. 9 GAN-based network architecture for STF

2.2.4 RNN

RNN 通过记忆机制, 能够有效建模时间依赖性, 在处理时间、文本和视频等序列数据中展现

出显著优势。强大的时间序列建模能力使其能够有效处理图像中的时间序列信息, 这对于 STF 具有重要价值 (Fung 等, 2019)。Fung 等 (2019) 首次

在 STF 中运用 RNN, 提出了 Hopfield 神经网络时空数据融合算法 HNN-SPOT (hopfield neural network spatiotemporal data fusion model)。它将高分辨率空间的每个像素视为一个神经元, 通过平衡精细图像的空间细节和粗图像的光谱响应值进行融合。HNN-SPOT 减少了对数据前提条件的需求, 但难以应对较大的时间和空间变化。Yang 等 (2021) 提出了混合深度学习时空融合模型, 它结合超分辨率卷积神经网络 SRCNN (Super-Resolution Convolutional Neural Network) 和长短期记忆 LSTM (Long Short-

Term Memory) 网络直接对图像的时、空依赖关系建模, 如图 10 所示。该方法利用 LSTM 来提取 SRCNN 派生图像中具有时间物候变化特征的时间演变特征, 并通过控制信息在记忆中的存储和更新, 捕捉图像间复杂的时间依赖关系, 提高了模型预测快速或非线性物候变化方面的能力。类似的, 基于时间序列的时空融合网络 TSSTFN (Time-Series-based Spatiotemporal Fusion Network) (Zhan 等, 2024) 也使用 LSTM 来建模长期依赖关系, 捕捉长时序图像中的动态变化模式。

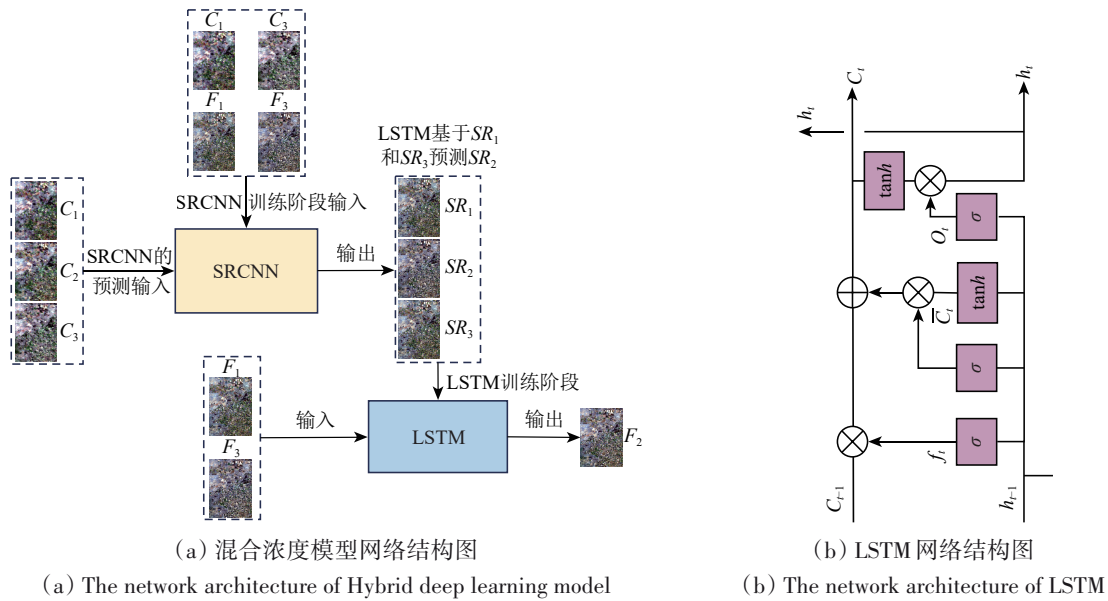


图 10 RNN 网络结构汇总

Fig. 10 Overview of the RNN architecture

2.2.5 扩散模型

在上述基于深度学习的 STF 方法中, 存在空间、时间和光谱不确定性, 以及复杂场景下 GAN 模式崩溃问题。而扩散模型具有显著的灵活性和鲁棒性, 近年来备受关注, 已成功地用于解决各种复杂视觉挑战 (Kawar 等, 2022; Rombach 等, 2022)。最新研究表明, 扩散模型在遥感图像 STF 任务中取得了新进展。如图 11 所示, 扩散模型的扩散过程包括 q (遵循马尔可夫链前提下的扩散过程, 该过程可逐步加噪) 和 p_θ (去噪过程, 该过程可逐步去噪) 两阶段, 通过迭代优化, 从噪声中生成目标图像 (Huang 等, 2024)。过程 q 将高斯噪声不断添加到干净的图像 x_0 中, 直至图像接近纯噪声。过程 p_θ 从图像中去除高斯噪声, 恢复干

净图像。加噪过程和去噪过程见式 (3) 和式 (4)。

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon_t, \epsilon_t \sim N(0, 1) \quad (3)$$

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{x_t - \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon_t(x_t, \hat{F}_2, F_1, C_2, t)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1}} \cdot \epsilon_t(x_t, \hat{F}_2, F_1, C_2, t) \quad (4)$$

式中, F_1 为参考日期的高分辨率图像, C_2 为预测日期的低分辨率图像, $\hat{F}_2$ 为预测结果, x_0 为干净图像 (原始无噪声图像), x_t 为第 t 步的加噪图像, ϵ_t 为高斯噪声, t 为时间步, $\bar{\alpha}_t$ 为累积保留系数, 是调度方法在前 t 步的累计作用结果。

基于条件扩散模型的时空融合模型 DiffSTF (spatiotemporal fusion method based on the conditional diffusion model) (Ma 等, 2024) 首次将条件扩散

模型用于遥感图像 STF，扩散和去噪过程如图 11 所示。该方法仅利用参考日期的细图像和预测日期的粗图像作为输入，同时对尺度和传感器偏差进行建模，提高突变物候变化适应性。另一个方法——时空融合扩散模型 STFDiff (Spatio-Temporal Fusion method with Diffusion models) (Huang 等, 2024)，提出通用条件扩散模型，网络流程如图 12 所示。扩散过程中，干净图像 x_0 逐渐被破坏成高斯噪声。去噪过程中，通过编—解码结构 DS-Unet 来预测噪声。最后，通过从噪声图像中迭代去除噪声，生成空间细节与细图像相似、时间动态与粗图像相似的融合图像。

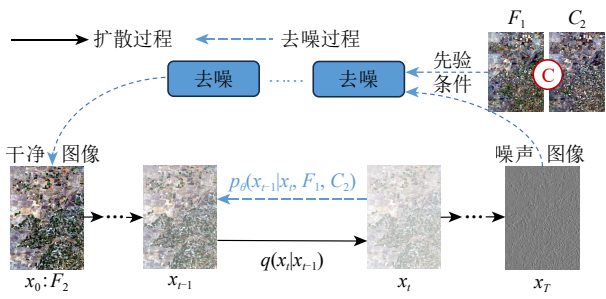


图 11 DiffSTF 网络流程
Fig. 11 Network architecture of the DiffSTF

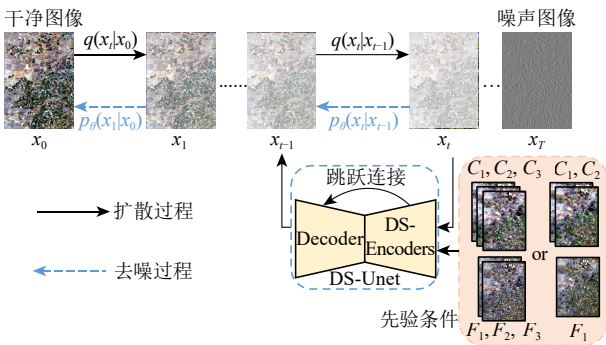


图 12 STFDIFF 网络流程
Fig. 12 Network architecture of the STFDIFF

虽然扩散模型的生成能力更强、训练稳定，且在复杂、多模态分布建模上表现优异，但为保证光谱精度或时序一致性，常需在损失或条件设计中额外加入物理/谱域约束，同时其采样需要大量逐步迭代，导致资源消耗较高。

3 数据集和评价指标

3.1 数据集介绍

本文汇总了 7 个用于遥感图像时空融合的数据集，包括 CIA (Emelyanova 等, 2013)、LGC (Emelyanova 等, 2013)、AHB (Li 等, 2020a)、TJ (Li 等, 2020a)、DX (Li 等, 2020a) 和 WH (Zhang 等, 2024b) 共 6 个公开数据集以及自制的 SW 数据集，相关信息如表 3 所示。其中，CIA 和 LGC 被广泛用于时间序列遥感研究。AHB、TJ 和 DX 为测试 STF 方法在检测农村、山区地区物候变化，城区物候变化及土地覆盖变化任务中的准确性提供基准。WH 数据集为 STF 提供了高分辨率数据基准。本团队构建的 SW 数据集包含农田、城市、村庄、裸土和水体等多种特征，旨在研究 STF 模型在不同卫星图像上适用性，为探索融合方法对不同数据源的适应性提供了思路。最后，本文将上述数据集及论文开源代码的在线获取地址进行了整理与汇总 (表 4)。

3.2 评价指标

借鉴 Zhu 等 (2022) 提出的“光谱—空间双维度”性能评估思想，为兼顾光谱一致性与空间结构保真性，确保性能评估的系统性与科学性，本文系统整理了前沿方法中常用于 STF 的评估指标，分别从空间结构维度和光谱一致性维度进行归纳，如表 5 所示。

表 3 遥感图像 STF 数据集

Table 3 Datasets for STF of remote sensing images

数据集名称	尺寸	时间跨度	数据源	变化类型
CIA	1720 pixels×2040 pixels	约 1 年	Landsat-MODIS	作物物候时间变化
LGC	3200 pixels×2720 pixels	约 1 年	Landsat-MODIS	地物突变 (洪水)
AHB	2480 pixels×2800 pixels	约 5 年	Landsat-MODIS	农区、山区物候变化
TJ	2100 pixels×1970 pixels	约 6 年	Landsat-MODIS	城区物候变化
DX	1640 pixels×1640 pixels	约 6 年	Landsat-MODIS	土地覆盖类型变化
WH	1000 pixels×1000 pixels	约 7 年	GF-Landsat	城市土地覆盖变化
SW	1792 pixels×1280 pixels	约 3 年	GF-MODIS	土地利用和覆盖变化

表4 STF 开放资源
Table 4 Open resources for STF

名称	链接
数据集	CIA
	http://dx.doi.org/10.4225/08/5111AC0BF1229
	LGC
	http://dx.doi.org/10.4225/08/5111AD2B7FEE6
	AHB
	TJ
方法	DX
	WH
	https://pan.baidu.com/s/1ymgud6tnY6XB5CTCXPUfnw
	https://github.com/lixinghua5540/Wuhan-dataset
	DCSTFN
	https://github.com/theonegis/rs-data-fusion
	EDCSTFN
	https://github.com/theonegis/edcstfn
	SwinSTFM
	https://github.com/microsoft/Swin-Transformer
	GAN-STFM
	https://github.com/theonegis/ganstfm
	MLFF-GAN
	https://github.com/songbingze/MLFF-GAN
	STFDiff
	https://github.com/prowDIY/STF
	ECPW-STFN
	https://github.com/lixinghua5540/ECPW-STFN

表5 STF 评估指标
Table 5 STF evaluation indexes

维度	名称	公式
空间结构	SSIM	$SSIM = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$
	CC	$CC = \frac{\sum_{i=1}^W \sum_{j=1}^H (x(i,j) - \mu_x)(y(i,j) - \mu_y)}{\sqrt{\sum_{i=1}^W \sum_{j=1}^H (x(i,j) - \mu_x)^2 (y(i,j) - \mu_y)^2}}$
	Edge	$Edge = x_{i,j} - x_{i+1,j+1} + y_{i,j+1} - y_{i+1,j} $ $LBP = decimal(d_1d_2, \dots, d_8)$
	LBP	$d_i = \begin{cases} 1 & \text{如果 } D_i > D_c \\ 0, & \text{其他} \end{cases}$
	RMSE	$RMSE = \sqrt{\frac{\sum_{i=1}^H \sum_{j=1}^W (y(i,j) - x(i,j))^2}{H \times W}}$
光谱一致性	SAM	$D_{SAM}(x) = \cos^{-1} \frac{\mathbf{d}^T \mathbf{x}}{(\mathbf{d}^T \mathbf{d})^{\frac{1}{2}} (\mathbf{x}^T \mathbf{x})^{\frac{1}{2}}}$
	AD	$AD = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W (y(i,j) - x(i,j))$
	ER-GAS	$ERGAS = 100 \frac{h}{l} \sqrt{\frac{1}{N} \sum_{k=1}^N \left(\frac{RMSE(B_k)}{\mu(k)} \right)^2}$

注： $y(i,j)$ 和 $x(i,j)$ 分别表示在位置 (i,j) 处的参考图像和融合图像的像素值， H 和 W 分别表示图像的高和宽， μ_x 、 σ_x 和 μ_y 、 σ_y 分别表示参考和融合图像的均值和方差， σ_{xy} 表示二者的协方差， μ 表示图像的平均值， $\mathbf{x}$ 和 $\mathbf{d}$ 分别表示融合图像和参考图像的光谱向量， h 和 l 分别表示融合图像和参考图像的空间分辨率， N 表示光谱波段数量， k 表示每个波段的索引， B_k 表示多光谱图像， $\mu(k)$ 表示多光谱图像辐射率的平均值， D_i 表示在 3×3 移动窗口中围绕中心像素的像素值， D_c 为中心像素值。

空间结构维度用于评估融合影像的空间结构、纹理与边缘保真度，包括结构相似性 SSIM

(structural similarity)、边缘强度差异 Edge (Robert's edge)、空间相关系数 CC (spatial Correlation Coefficient) 与局部二值模式 LBP (Local Binary Pattern)。光谱一致性维度用于衡量融合影像在光谱精度、辐射能量及色彩还原方面与参考影像的一致性，包括均方根误差 RMSE (Root Mean Square Error)、平均误差 AD (Average Difference)、光谱角相似度 SAM (Spectral Angle Mapper) 与全局相对光谱损失 ERGAS (Erreur Relative Globale Adimensionnelle de Synthèse)。

SSIM (Wang 等, 2004) 基于人类视觉系统的结构敏感特性，通过亮度、对比度与结构比较来评估图像的结构相似性，反映预测图像的感知质量。CC (Yang 等, 2022) 衡量二者的空间细节一致性，通过高通滤波提取高频特征并计算各波段相关性，以刻画空间结构保真度。Edge (Zhu 等, 2022) 衡量预测与参考影像在边缘结构上的差异，评估空间结构和细节的保持能力。LBP (Zhu 等, 2022) 描述纹理分布的一致性，反映空间纹理特征的相似度。RMSE 计算二者之间的像素偏差，反映全局光谱误差。SAM (Yuhua 等, 1992) 通过光谱向量夹角评估光谱一致性。AD (Zhu 等, 2022) 衡量预测与参考影像的平均光谱偏移趋势，体现系统性偏差。ERGAS (Renza 等, 2013) 衡量两幅图像的整体光谱失真程度，反映全局辐射差异。理想情况下，Edge、LBP、RMSE、SAM、AD 和 ERGAS 的最优值为 0，而 SSIM 和 CC 的最优值为 1。

4 实验分析

本实验选取FSDAF、EDCSTFN、ECPW-STFN、SwinSTFM、MLKNet、MLFF-GAN 和 STFDiff 作为基准模型，采用RMSE、SAM、SSIM、ERGAS 和 CC 作为评估指标评，在 CIA、DX 和 SW 数据集上进行对比分析。其中FSDAF、EDCSTFN、MLKNet、MLFF-GAN 和 STFDiff 为三输入模型，SwinSTFM、ECPW-STFN 为二输入模型。此外，EDCSTFN 与 ECPW-STFN，MLKNet 与 SwinSTFM 分别属于同类型模型，以便有效观察不同网络结构及不同输入组合对融合效果的影响。

4.1 定量分析

表 6 为各方法在 CIA、DX 和 SW 上的定量结果。在 3 个数据集中，传统方法结果较差，但在空

间结构保持方面作用显著，且在城镇复杂的 DX 数据集中取得了不错的效果。扩散模型在各项指标中均得到了最优或次优的结果，但也带来更多计算量，增加了计算负担。在特征复杂的 DX 和 SW 数据集中，生成式 GAN 和扩散模型均取得较优的结果。三输入模型中的 MLKNet 在各方面均优于二输入模型中的 SwinSTFM，这主要得益于更多的输入可以提供更多的时序信息和时空约束，使模型能够提取更多的特征细节和先验知识，从而更好地捕捉目标的时空变化，提升复杂变化场景下的鲁棒性与预测精度。在二输入的模型中，基于频域增强的 ECPW-STFN 在多项指标中取得了次优的效果，验证了从频域的角度进行时空融合的有效性。此外，采用自注意力的 MLFF-GAN 虽取得了较好的结果，但也带来了计算负担。

表 6 STF 定量比较结果
Table 6 Quantitative comparisons on STF results

方法		模型						
		FSDAF	EDCSTFN	ECPW-STFN	MLFF-GAN	MLKNet	SwinSTFM	STFDiff
Param (M)/Flops(G)		—/—	0.19/24.39	4.39/30.55	33.84/39.87	43.77/83.84	37.53/28.08	33.18/209.78
CIA	RMSE ↓	0.0210	0.0242	0.0203	0.0193	<u>0.0186</u>	0.0189	0.0180
	SAM ↓	0.0645	0.0668	0.0615	0.0583	0.0549	<u>0.0567</u>	0.0584
	SSIM ↑	0.9011	0.9029	0.9135	0.9153	0.9213	0.9116	<u>0.9165</u>
	ERGAS ↓	1.1571	1.4565	1.1060	1.0412	1.0195	1.0432	<u>1.0229</u>
	CC ↑	0.8637	0.8048	0.8711	0.8860	<u>0.8894</u>	0.8809	0.8955
DX	RMSE ↓	0.0299	0.0297	0.0297	<u>0.0282</u>	0.0293	0.0285	0.0272
	SAM ↓	0.0673	0.0676	0.0679	0.0610	<u>0.0654</u>	0.0668	0.0657
	SSIM ↑	0.8402	0.8602	<u>0.8541</u>	0.8431	0.8365	0.8369	0.8444
	ERGAS ↓	1.3199	1.4136	1.2744	1.2483	1.2081	1.2720	<u>1.2220</u>
	CC ↑	0.8591	0.8472	0.8628	<u>0.8670</u>	0.8531	0.8568	0.8733
SW	RMSE ↓	0.0690	0.0451	0.0367	<u>0.0311</u>	0.0335	0.0355	0.0280
	SAM ↓	0.1657	0.1194	0.1179	0.0863	0.1089	0.1324	<u>0.0997</u>
	SSIM ↑	0.6985	0.7904	0.7901	0.8282	<u>0.8056</u>	0.7703	0.7951
	ERGAS ↓	5.0529	2.9748	2.7918	<u>1.9779</u>	2.4632	2.7203	1.8869
	CC ↑	0.8847	0.8910	<u>0.9205</u>	0.9177	0.8963	0.8973	0.9346

注：粗体表示最佳结果，下划线表示次优结果；Param (M)指的是参数量(Parameters, 单位 M=Million)，表示模型中可学习参数的数量；Flops (G)指的是浮点数运算量[Floating Point Operations, 单位 G=Giga(10⁹)]，表示模型前向推理需要的计算量；向上箭头代表指标越大越好，向下箭头代表指标越小越好。

4.2 定性分析

为直观展示各方法的优劣，本文对融合结果进行可视化，并对局部区域进行放大处理，以分析和展示融合结果细节信息，如图 13、图 14 和图 15 所示。

图 13 展示了 CIA 数据集的可视化结果，第 1 列为真实的观测图像。CIA 数据集大部分区域被农田与植被覆盖，地物类别清晰，第 2 行展示了不同方法融合的真彩影像，以直观对比空间细节和纹理信息的保持情况。第 3 行为假彩图，反映了近红外

波段的信息保留情况和植被的光谱特征。从视觉角度看，传统方法也取得了很好的结果。例如黄色区域中，EDCSTFN、SwinSTFM 和 ECPW-STFN 融合结果更加模糊，但 ECPW-STFN 优于 EDCSTFN；绿色区域中，FSDAF、MLFF-GAN、MLKNet 和 STFDiff 融合结果相近，纹理更为清晰，更接近真实影像，但 FSDAF 对近红外波段信息保留不足，与真实图像存在偏差，MLKNet、STFDiff 更加接近

观测图像。这进一步说明，三输入的方法可以提取图像更多的特征细节，在视觉效果上优于二输入方法，同时也验证了其对频域建模的有效性。而 ECPW-STFN 和 SwinSTFM 相比，存在边界模糊问题，这是由于自注意力机制可以对全局特征进行建模，而 ECPW-STFN 方法，未采用全局建模方式。

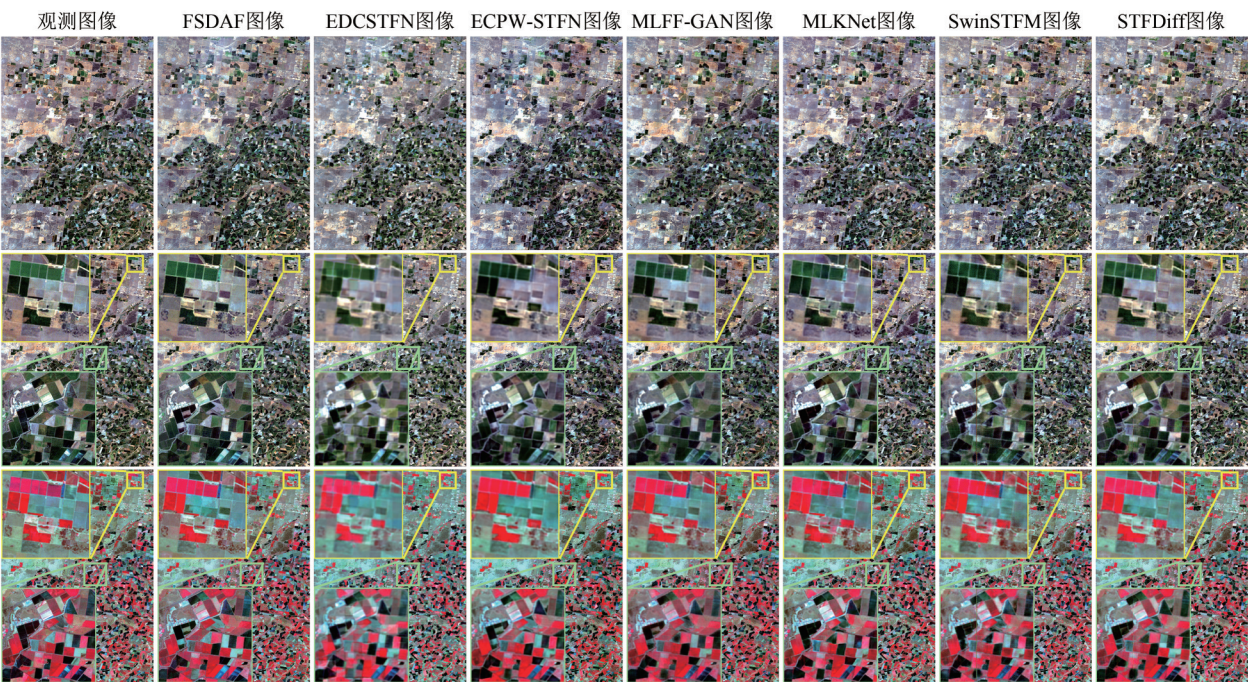


图 13 CIA 数据集可视化结果
Fig. 13 Visualization comparison results on CIA dataset

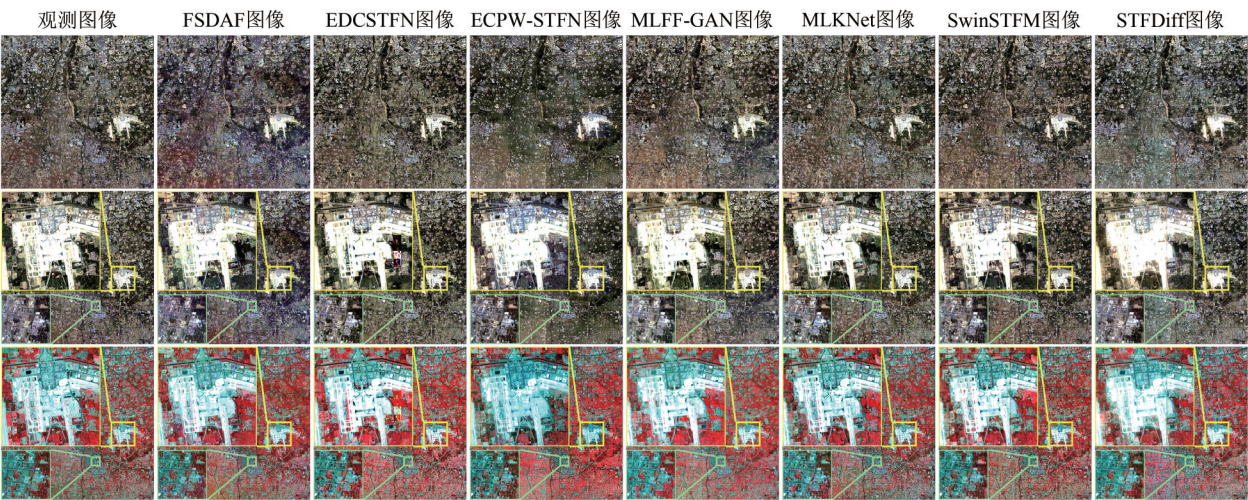


图 14 DX 数据集可视化结果
Fig. 14 Visualization comparison results on DX dataset

图 14、图 15 展示了各方法在 DX、SW 数据集上的融合结果。DX 数据集以城区建筑为主，包含

跨时 5 年的机场建设，如图 14 中黄色框线所示，土地覆盖变化明显。SW 数据集主要以耕地、植被

为主,变化类型以土地利用和土地覆盖为主。从视觉效果来看,传统方法和扩散模型均取得了较好的结果,但从假彩图的绿色框线可以看出,深度学习方法在细节还原上更接近真实影像。其中,基于注意力机制的融合结果更加接近真实观测图像,这说明对全局特征的建模能力与融合质量密

切相关。值得注意的是,DX数据集的城镇特征更加复杂,部分特征细节难以提取,导致其定量指标低于CIA数据集。并且,SW数据集的特征类型复杂多样,且分辨率差异较大,各方法在该数据集上的表现均不理想。

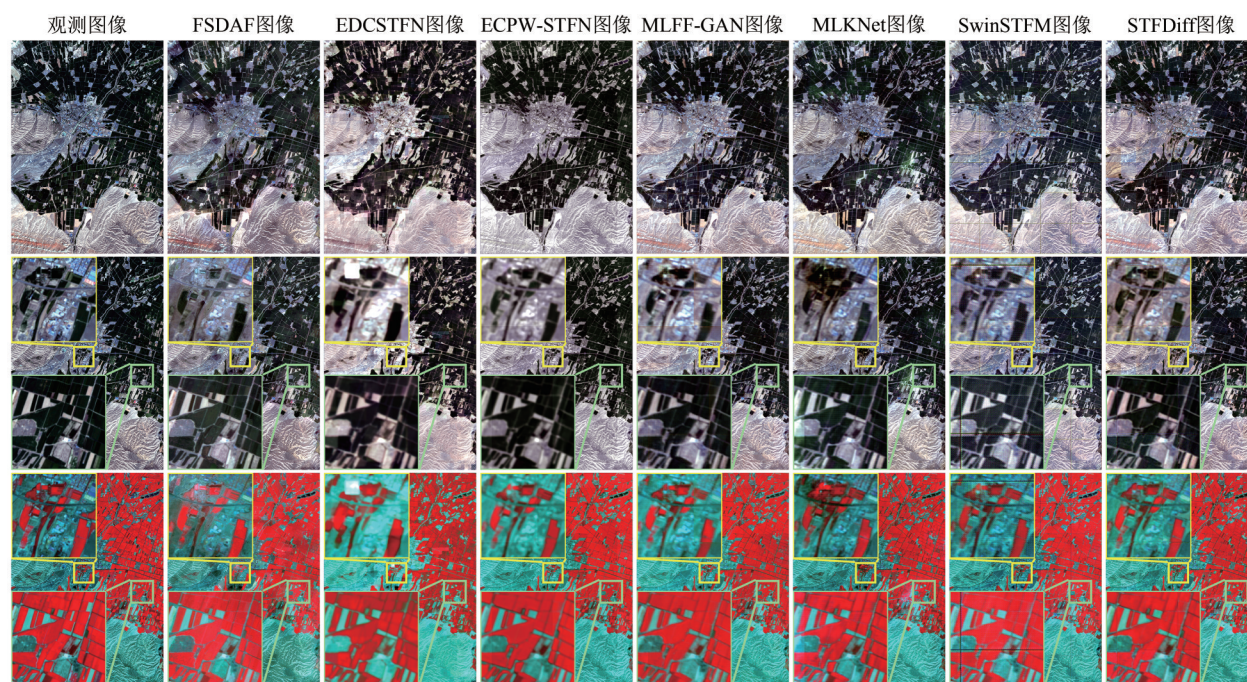


图 15 SW数据集可视化结果

Fig. 15 Visualization comparison results on SW dataset

5 结 语

本文首先介绍了5种传统STF方法,并归纳了各自的原理和优缺点。其次,对基于深度学习的前沿STF方法进行了合理的分类梳理,重点介绍了各类方法的原理和特点,总结了其在时空融合任务中的应用进展,并对现有研究中可公开的研究资源进行了整理汇总。最后,在3个数据集上对深度学习方法进行了对比实验,分析了各方法的优势和不足,并针对STF模型的数据方面及未来发展做出如下展望。

(1) 基于频域增强的STF方法: ECPW-STFN证明了将空间域像素特征转换到频率域进行STF的有效性。该方法通过对高频、低频特征的分别处理,在避免像素损失的同时减少了计算负担。然而,尽管当前小波变换方法在遥感图像变化监测、语义分割和超分辨率等领域应用广泛,但其在遥感图像STF领域的应用极少。未来研究可以考虑将

空间域和频率域进行交互,结合二者的优势,提升预测效果。

(2) Mamba网络的引入与设计: 基于Transformer的遥感图像STF方法凭借其对长时序的处理能力,在诸多先进方法中取得了优异结果,但同时也导致网络更复杂,计算量增加。而Mamba的兴起极大的促进了遥感领域的发展。其凭借线性时间序列建模和选择状态空间,在语义分割、变化监测等领域得到了广泛应用。但截至目前,Mamba尚未涉及遥感图像STF领域,这有望成为下一发展阶段的重要研究方向,应受到广泛关注。

(3) 大尺度分辨率差异图像时空融合: 用于多源遥感图像时空融合的源图像之间分辨率存在差异,常用的Landsat和MODIS影像分辨率相差16倍,而GF-1和MODIS的差异更达到30倍。克服多源图像之间大分辨率差异,实现精确的时空融合,仍然是一个挑战。对于能够对差异进行建

模的正确特征的提取问题,虽然已有一些解决方法(Jiang等, 2024; Erdem和Avdan, 2023),但仍存在上采样方法和位置选择等开放问题。此外,通过有机结合超分辨率来设计网络,实现大分辨率差异图像时空融合,是一个重要思路。

(4) 无监督与半监督学习方法:当前的时空融合方法多采用有监督的深度学习方法,基于无监督和半监督的网络模型较为少见。尽管遥感数据量在不断增加,但受到云雾、光照等因素的影响,能够用于时空融合任务的数据数量仍然较少,其中带有标记的数据数量更是有限,这使得已有的数据不能被充分利用。当前无监督和半监督学习在遥感图像空谱融合研究中有较多应用,未来研究中使用无监督或半监督学习方法进行时空融合具有广泛前景和重要意义。

参考文献(References)

- Belgiu M and Stein A. 2019. Spatiotemporal image fusion in remote sensing. *Remote Sensing*, 11(7): 818 [DOI: 10.3390/rs11070818]
- Benzenati T, Kallel A and Kessentini Y. 2024. STF-Trans: a two-stream spatiotemporal fusion transformer for very high resolution satellites images. *Neurocomputing*, 563: 126868 [DOI: 10.1016/j.neucom.2023.126868]
- Chen B, Huang B and Xu B. 2015. Comparison of spatiotemporal fusion models: a review. *Remote Sensing*, 7(2): 1798-1835 [DOI: 10.3390/rs70201798]
- Chen G S, Lu H L, Zou W T, Li L H, Emam M, Chen X B, Jing W P, Wang J and Li C. 2023. Spatiotemporal fusion for spectral remote sensing: a statistical analysis and review. *Journal of King Saud University-Computer and Information Sciences*, 35(3): 259-273 [DOI: 10.1016/j.jksuci.2023.02.021]
- Chen G Y, Jiao P, Hu Q, Xiao L J and Ye Z J. 2022. SwinSTFM: remote sensing spatiotemporal fusion using swin transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 5410618 [DOI: 10.1109/Tgrs.2022.3182809]
- Cui D Y, Wang S D, Zhao C W and Zhang H B. 2024a. A novel remote sensing spatiotemporal data fusion framework based on the combination of deep-learning downscaling and traditional fusion algorithm. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17: 7957-7970 [DOI: 10.1109/jstars.2024.3382136]
- Cui J W, Li J, Gu X F, Zhang W H, Wang D, Sun X L, Zhan Y L, Yang J, Liu Y and Yang X F. 2024b. Comprehensive analysis of temporal-spatial fusion from 1991 to 2023 using bibliometric tools. *Atmosphere*, 15(5): 598 [DOI: 10.3390/atmos15050598]
- Dong W Q and Meng J H. 2018. Review of spatiotemporal fusion model of remote sensing data. *Remote Sensing for Natural Resources*, 30(2): 1-11 (董文全, 蒙继华. 2018. 遥感数据时空融合研究进展及展望. *国土资源遥感*, 30(2): 1-11) [DOI: 10.6046/gtzyyg.2018.02.01]
- Emelyanova I V, McVicar T R, Van Niel T G, Li L T and Van Dijk A I J M. 2013. Assessing the accuracy of blending Landsat-MODIS surface reflectances in two landscapes with contrasting spatial and temporal dynamics: a framework for algorithm selection. *Remote Sensing of Environment*, 133: 193-209 [DOI: 10.1016/j.rse.2013.02.007]
- Erdem F and Avdan U. 2023. STFRDN: a residual dense network for remote sensing image spatiotemporal fusion. *International Journal of Remote Sensing*, 44(10): 3259-3277 [DOI: 10.1080/01431161.2023.2221800]
- Fu R D, Hu H Y, Wu N, Liu Z N and Jin W. 2024. Spatiotemporal fusion convolutional neural network: tropical cyclone intensity estimation from multisource remote sensing images. *Journal of Applied Remote Sensing*, 18(1): 018501 [DOI: 10.1117/1.JRS.18.018501]
- Fung C H, Wong M S and Chan P W. 2019. Spatio-temporal data fusion for satellite images using hopfield neural network. *Remote Sensing*, 11(18): 2077 [DOI: 10.3390/rs11182077]
- Gao F, Masek J, Schwaller M and Hall F. 2006. On the blending of the Landsat and MODIS surface reflectance: predicting daily Landsat surface reflectance. *IEEE Transactions on Geoscience and Remote Sensing*, 44(8): 2207-2218 [DOI: 10.1109/TGRS.2006.872081]
- Gevaert C M and García-Haro F J. 2015. A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion. *Remote Sensing of Environment*, 156: 34-44 [DOI: 10.1016/j.rse.2014.09.012]
- Ghamisi P, Gloaguen R, Atkinson P M, Benediktsson J A, Rasti B, Yokoya N, Wang Q M, Hofle B, Bruzzone L, Bovolo F, Chi M M and Anders K. 2019. Multisource and multitemporal data fusion in remote sensing: a comprehensive review of the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 7(1): 6-39 [DOI: 10.1109/MGRS.2018.2890023]
- Guo D Z, Shi W Z, Hao M and Zhu X L. 2020. FSDAF 2.0: improving the performance of retrieving land cover changes and preserving spatial details. *Remote Sensing of Environment*, 248: 111973 [DOI: 10.1016/j.rse.2020.111973]
- Guo H C, Ye D Q, Xu H Z Y and Bruzzone L. 2024. OBSUM: an object-based spatial unmixing model for spatiotemporal fusion of remote sensing images. *Remote Sensing of Environment*, 304: 114046 [DOI: 10.1016/j.rse.2024.114046]
- Huang B and Jiang X L. 2021. An enhanced unmixing model for spatiotemporal image fusion. *National Remote Sensing Bulletin*, 25(1): 241-250 (黄波, 姜晓璐. 2021. 增强型空间像元分解时空遥感影像融合算法. *遥感学报*, 25(1): 241-250) [DOI: 10.11834/jrs.20210459]
- Huang B and Song H H. 2012. Spatiotemporal reflectance fusion via sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 50(10): 3707-3716 [DOI: 10.1109/TGRS.2012.2186638]

- Huang B and Zhao Y Q. 2017. Research status and prospect of spatio-temporal fusion of multi-source satellite remote sensing imagery. *Acta Geodaetica et Cartographica Sinica*, 46(10): 1492-1499 (黄波, 赵涌泉. 2017. 多源卫星遥感影像时空融合研究的现状及展望. *测绘学报*, 46(10): 1492-1499) [DOI: 10.11947/j.AGCS.2017.20170376]
- Huang H, He W, Zhang H Y, Xia Y and Zhang L P. 2024. STFDiff: remote sensing image spatiotemporal fusion with diffusion models. *Information Fusion*, 111: 102505 [DOI: 10.1016/j.inffus.2024.102505]
- Jiang H, Qian Y R, Yang G Q and Liu H. 2024. MLKNet: multi-stage for remote sensing image spatiotemporal fusion network based on a large kernel attention. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17: 1257-1268 [DOI: 10.1109/Jstars.2023.3338978]
- Jing Y H, Shen H F, Li X H, Wu J G and Qiu Z H. 2024. Spatial down-scaling of remote sensing parameters from perspective of data fusion. *Geomatics and Information Science of Wuhan University*, 49(2): 175-189 (景映红, 沈焕锋, 李星华, 吴金橄, 邱中航. 2024. 数据融合视角下的遥感参量空间降尺度. *武汉大学学报(信息科学版)*, 49(2): 175-189) [DOI: 10.13203/j.whugis20220549]
- Justice C O, Vermote E, Townshend J R G, Defries R, Roy D P, Hall D K, Salomonson V V, Privette J L, Riggs G, Strahler A, Lucht W, Myneni R B, Knyazikhin Y, Running S W, Nemani R R, Zhengming Wan N, Huete A R, Van Leeuwen W, Wolfe R E, Giglio L, Muller J, Lewis P and Barnsley M J. 1998. The moderate resolution imaging spectroradiometer (MODIS): land remote sensing for global change research. *IEEE Transactions on Geoscience and Remote Sensing*, 36(4): 1228-1249 [DOI: 10.1109/36.701075]
- Kawar B, Elad M, Ermon S and Song J M. 2022. Denoising diffusion restoration models//*Proceedings of the 36th International Conference on Neural Information Processing Systems*. New Orleans: Curran Associates Inc.: 1714
- Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z H and Shi W Z. 2017. Photo-realistic single image super-resolution using a generative adversarial network//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu: IEEE: 105-114 [DOI: 10.1109/CVPR.2017.19]
- Lei D J, Zhu Q W, Li Y J, Tan J Y, Wang S L, Zhou T T and Zhang L P. 2024. HPLTS-GAN: a high-precision remote sensing spatiotemporal fusion method based on low temporal sensitivity. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 5407416 [DOI: 10.1109/TGRS.2024.3416152]
- Li A H, Bo Y C, Zhu Y X, Guo P, Bi J and He Y Q. 2013. Blending multi-resolution satellite sea surface temperature (SST) products using Bayesian maximum entropy method. *Remote Sensing of Environment*, 135: 52-63 [DOI: 10.1016/j.rse.2013.03.021]
- Li J, Li Y F, He L, Chen J and Plaza A. 2020a. Spatio-temporal fusion for remote sensing data: an overview and new benchmark. *Science China Information Sciences*, 63(4): 140301 [DOI: 10.1007/s11432-019-2785-y]
- Li J X, Hong D F, Gao L R, Yao J, Zheng K, Zhang B and Chanussot J. 2022a. Deep learning in multimodal remote sensing data fusion: a comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, 112: 102926 [DOI: 10.1016/j.jag.2022.102926]
- Li R, Wang M M, Zhang Z J, Hu T and Liu X G. 2022. A review of spatiotemporal fusion methods for remotely sensed land surface temperature. *National Remote Sensing Bulletin*, 26(12): 2433-2450 (李冉, 王猛猛, 张正加, 胡添, 刘修国. 2022. 遥感地表温度产品时空融合方法研究综述. *遥感学报*, 26(12): 2433-2450) [DOI: 10.11834/jrs.20210294]
- Li W S, Cao D W, Peng Y D and Yang C. 2021a. MSNet: a multi-stream fusion network for remote sensing spatiotemporal fusion based on transformer and convolution. *Remote Sensing*, 13(18): 3724 [DOI: 10.3390/rs13183724]
- Li W S, Cao D W and Xiang M H. 2022b. Enhanced multi-stream remote sensing spatiotemporal fusion network based on transformer and dilated convolution. *Remote Sensing*, 14(18): 4544 [DOI: 10.3390/rs14184544]
- Li W S, Yang C, Peng Y D and Zhang X Y. 2021b. A multi-cooperative deep convolutional neural network for spatiotemporal satellite image fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 10174-10188 [DOI: 10.1109/JSTARS.2021.3113163]
- Li X D, Foody G M, Boyd D S, Ge Y, Zhang Y H, Du Y and Ling F. 2020b. SFSDAF: an enhanced FSDAF that incorporates sub-pixel class fraction change information for spatio-temporal image fusion. *Remote Sensing of Environment*, 237: 111537 [DOI: 10.1016/j.rse.2019.111537]
- Li Y F, Li J and He L. 2022. Convolutional neural network based single image pair method for spatiotemporal fusion. *National Remote Sensing Bulletin*, 26(8): 1614-1623 (李云飞, 李军, 贺霖. 2022. 单样本对卷积神经网络遥感图像时空融合. *遥感学报*, 26(8): 1614-1623) [DOI: 10.11834/jrs.20219348]
- Lian Z L, Zhan Y L, Zhang W H, Wang Z J, Liu W B and Huang X H. 2025. Recent advances in deep learning-based spatiotemporal fusion methods for remote sensing images. *Sensors*, 25(4): 1093 [DOI: 10.3390/s25041093]
- Liao L M, Song J L, Wang J D, Xiao Z Q and Wang J. 2016. Bayesian method for building frequent landsat-like NDVI datasets by integrating MODIS and landsat NDVI. *Remote Sensing*, 8(6): 452 [DOI: 10.3390/rs8060452]
- Liu H, Yang G Q, Deng F L, Qian Y R and Fan Y Y. 2023. MCBAM-GAN: the gan spatiotemporal fusion model based on multiscale and CBAM for remote sensing images. *Remote Sensing*, 15(6): 1583 [DOI: 10.3390/rs15061583]
- Liu M, Yang W, Zhu X L, Chen J, Chen X H, Yang L Q and Helmer E H. 2019. An improved flexible spatiotemporal data fusion (IFS-DAF) method for producing high spatiotemporal resolution normalized difference vegetation index time series. *Remote Sensing of Environment*, 227: 74-89 [DOI: 10.1016/j.rse.2019.03.012]
- Liu S J, Liu J, Tan X Y, Chen X H and Chen J. 2024. A hybrid spatio-

- temporal fusion method for high spatial resolution imagery: fusion of gaofen-1 and sentinel-2 over agricultural landscapes. *Journal of Remote Sensing*, 4: 0159 [DOI: 10.34133/remotesensing.0159]
- Liu S J, Zhou J X, Qiu Y A, Chen J, Zhu X L and Chen H. 2022. The FIRST model: spatiotemporal fusion incorrpoting spectral auto-correlation. *Remote Sensing of Environment*, 279: 113111 [DOI: 10.1016/j.rse.2022.113111]
- Ma Y B, Wang Q and Wei J B. 2024. Spatiotemporal fusion via conditional diffusion model. *IEEE Geoscience and Remote Sensing Letters*, 21: 5002405 [DOI: 10.1109/LGRS.2024.3378715]
- Pan X Y, Deng M Y, Ao Z R and Xin Q C. 2023. An adaptive multi-scale generative adversarial network for the spatiotemporal fusion of landsat and MODIS data. *Remote Sensing*, 15(21): 5128 [DOI: 10.3390/rs15215128]
- Peng Y D, Li W S, Luo X B, Du J, Zhang X Y, Gan Y and Gao X B. 2022. Spatiotemporal reflectance fusion via tensor sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 5608318 [DOI: 10.1109/TGRS.2021.3091157]
- Qian Z H, Yue L W, Xie X, Yuan Q Q and Shen H F. 2024. A dual-perspective spatiotemporal fusion model for remote sensing images by discriminative learning of the spatial and temporal mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17: 12505-12520 [DOI: 10.1109/JSTARS.2024.3426944]
- Ran Q, Wang Q H, Zheng K and Li J X. 2024. Multiscale attention spatiotemporal fusion model based on pyramidal network constraints. *IEEE Geoscience and Remote Sensing Letters*, 21: 5004305 [DOI: 10.1109/Lgrs.2024.3432938]
- Renza D, Martinez E and Arquero A. 2013. A new approach to change detection in multispectral images by means of ERGAS index. *IEEE Geoscience and Remote Sensing Letters*, 10(1): 76-80 [DOI: 10.1109/LGRS.2012.2193372]
- Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans: IEEE: 10684-10695 [DOI: 10.1109/CVPR52688.2022.01042]
- Roy D P, Wulder M A, Loveland T R, Woodcock C E, Allen R G, Anderson M C, Helder D, Irons J R, Johnson D M, Kennedy R, Scambos T A, Schaaf C B, Schott J R, Sheng Y, Vermote E F, Belward A S, Bindschadler R, Cohen W B, Gao F, Hipple J D, Hostert P, Huntington J, Justice C O, Kilic A, Kovalsky V, Lee Z P, Lymburner L, Masek J G, Mccorkel J, Shuai Y, Trezza R, Vogelmann J, Wynne R H and Zhu Z. 2014. Landsat-8: science and product vision for terrestrial global change research. *Remote Sensing of Environment*, 145: 154-172 [DOI: 10.1016/j.rse.2014.02.001]
- Shen H F, Meng X C and Zhang L P. 2016. An integrated framework for the spatio-temporal-spectral fusion of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12): 7135-7148 [DOI: 10.1109/TGRS.2016.2596290]
- Shi W Z, Guo D Z and Zhang H. 2022. A reliable and adaptive spatiotemporal data fusion method for blending multi-spatiotemporal-resolution satellite images. *Remote Sensing of Environment*, 268: 112770 [DOI: 10.1016/j.rse.2021.112770]
- Song B Z, Liu P, Li J, Wang L Z, Zhang L, He G J, Chen L J and Liu J B. 2022. MLFF-GAN: a multilevel feature fusion with GAN for spatiotemporal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 4410816 [DOI: 10.1109/TGRS.2022.3169916]
- Song H H and Huang B. 2013. Spatiotemporal satellite image fusion through one-pair image learning. *IEEE Transactions on Geoscience and Remote Sensing*, 51(4): 1883-1896 [DOI: 10.1109/TGRS.2012.2213095]
- Song H H, Liu Q S, Wang G J, Hang R L and Huang B. 2018. Spatiotemporal satellite image fusion using deep convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(3): 821-829 [DOI: 10.1109/JSTARS.2018.2797894]
- Swain R, Paul A and Behera M D. 2024. Spatio-temporal fusion methods for spectral remote sensing: a comprehensive technical review and comparative analysis. *Tropical Ecology*, 65(3): 356-375 [DOI: 10.1007/s42965-023-00318-5]
- Tan Z Y, Di L P, Zhang M D, Guo L Y and Gao M L. 2019. An enhanced deep convolutional model for spatiotemporal image fusion. *Remote Sensing*, 11(24): 2898 [DOI: 10.3390/rs11242898]
- Tan Z Y, Gao M L, Li X H and Jiang L C. 2022. A flexible reference-insensitive spatiotemporal fusion model for remote sensing images using conditional generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 5601413 [DOI: 10.1109/TGRS.2021.3050551]
- Tan Z Y, Yue P, Di L P and Tang J M. 2018. Deriving high spatiotemporal remote sensing images using deep convolutional network. *Remote Sensing*, 10(7): 1066 [DOI: 10.3390/rs10071066]
- Wang Q M, Peng K D, Tang Y J, Tong X H and Atkinson P M. 2021. Blocks-removed spatial unmixing for downscaling MODIS images. *Remote Sensing of Environment*, 256: 112325 [DOI: 10.1016/j.rse.2021.112325]
- Wang S, Cheng D P, Yuan G J and Li J J. 2025. RDSF-Net: residual wavelet mamba-based differential completion and spatio-frequency extraction remote sensing change detection network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18: 11573-11587 [DOI: 10.1109/JSTARS.2025.3559708]
- Wang Z, Bovik A C, Sheikh H R and Simoncelli E P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600-612 [DOI: 10.1109/TIP.2003.819861]
- Wang Z B, Ma Y K and Zhang Y N. 2023. Review of pixel-level remote sensing image fusion based on deep learning. *Information Fusion*, 90: 36-58 [DOI: 10.1016/j.inffus.2022.09.008]
- Wang Z H and Qian Y T. 2024. Super-resolution reconstruction of spatiotemporal fusion for dual-stream remote sensing images based on swin transformer. *Computer Engineering*, 50(9): 33-45 (王志

- 浩, 钱云涛. 基于 Swin Transformer 的双流遥感图像时空融合超分辨率重建. 计算机工程, 50(9): 33-45 [DOI: 10.19678/j.issn.1000-3428.0068296]
- Wu J G, Cheng Q, Li H F, Li S, Guan X B and Shen H F. 2020. Spatio-temporal fusion with only two remote sensing images as input. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13: 6206-6219 [DOI: 10.1109/JSTARS.2020.3028116]
- Wu M Q, Huang W J, Niu Z and Wang C Y. 2015. Generating daily synthetic landsat imagery by combining landsat and MODIS data. *Sensors*, 15(9): 24002-24025 [DOI: 10.3390/s150924002]
- Wu P H, Yin Z X, Zeng C, Duan S B, Gottsche F M, Ma X S, Li X H, Yang H and Shen H F. 2021. Spatially continuous and high-resolution land surface temperature product generation: a review of reconstruction and spatiotemporal fusion techniques. *IEEE Geoscience and Remote Sensing Magazine*, 9(3): 112-137 [DOI: 10.1109/MGRS.2021.3050782]
- Xie Y P, Hu J, He K, Cao L, Yang K J and Chen L. 2025. The GAN spatiotemporal fusion model based on multiscale convolution and attention mechanism for remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18: 2283-2296 [DOI: 10.1109/Jstars.2024.3507815]
- Xu C, Du X P, Fan X T, Jian H D, Yan Z Z, Zhu J J and Wang R. 2024. FastVSDF: an efficient spatiotemporal data fusion method for seamless data cube. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 5402022 [DOI: 10.1109/TGRS.2024.3353758]
- Xu Y F, Zhou Y, Ma H B, Yang H T, Wang H Y, Zhang S L and Li X W. 2025. Wavelet-based dual discriminator GAN for image super-resolution. *Knowledge-Based Systems*, 317: 113383 [DOI: 10.1016/j.knsys.2025.113383]
- Yang G Q, Liu H, Zhong X W, Chen L and Qian Y R. 2022. Temporal and spatial fusion of remote sensing images: a review. *Computer Engineering and Applications*, 58(10): 27-40 (杨广奇, 刘慧, 钟锡武, 陈龙, 钱育蓉. 2022. 遥感图像时空融合综述. 计算机工程与应用, 58(10): 27-40) [DOI: 10.3778/j.issn.1002-8331.2111-0131]
- Yang G Q, Qian Y R, Liu H, Tang B C, Qi R R, Lu Y and Geng J. 2022. MSFusion: multistage for remote sensing image spatiotemporal fusion based on texture transformer and convolutional neural network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15: 4653-4666 [DOI: 10.1109/Jstars.2022.3179415]
- Yang Y S, Yuan G J and Li J J. 2024. SFFNet: a wavelet-based spatial and frequency domain fusion network for remote sensing segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 3000617 [DOI: 10.1109/Tgrs.2024.3427370]
- Yang Z J, Diao C Y and Li B. 2021. A robust hybrid deep learning model for spatiotemporal image fusion. *Remote Sensing*, 13(24): 5005 [DOI: 10.3390/rs13245005]
- Yuhas R H, Goetz A F H, Boardman J W. 1992. Discrimination among semi-arid landscape endmembers using the Spectral Angle Mapper (SAM) algorithm//Proceedings of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop. [S.l.]: JPL.
- Zhan W F, Luo F, Luo H, Li J L, Wu Y C, Yin Z X, Wu Y L and Wu P H. 2024. Time-series-based spatiotemporal fusion network for improving crop type mapping. *Remote Sensing*, 16(2): 235 [DOI: 10.3390/rs16020235]
- Zhang H, Xu H, Tian X, Jiang J J and Ma J Y. 2021a. Image fusion meets deep learning: a survey and perspective. *Information Fusion*, 76: 323-336 [DOI: 10.1016/j.inffus.2021.06.008]
- Zhang H Y, Song Y Y, Han C and Zhang L P. 2021b. Remote sensing image spatiotemporal fusion using a generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5): 4273-4286 [DOI: 10.1109/Tgrs.2020.3010530]
- Zhang L X, Zeng G P and Xuan Z C. 2022. A survey of fusion methods for multi-source image. *Computer Engineering and Science*, 44(2): 321-334 (张丽霞, 曾广平, 宣兆成. 2022. 多源图像融合方法的研究综述. 计算机工程与科学, 44(2): 321-334) [DOI: 10.3969/j.issn.1007-130X.2022.02.018]
- Zhang X J, Li S, Tan Z Y and Li X H. 2024a. Enhanced wavelet based spatiotemporal fusion networks using cross-paired remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 211: 281-297 [DOI: 10.1016/j.isprsjprs.2024.04.016]
- Zhang X J, Xie L L, Li S, Lei F, Cao L and Li X H. 2024b. Wuhan dataset: a high-resolution dataset of spatiotemporal fusion for remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 21: 2504305 [DOI: 10.1109/Lgrs.2024.3432285]
- Zhu X L, Cai F Y, Tian J Q and Williams T K A. 2018. Spatiotemporal fusion of multisource remote sensing data: literature survey, taxonomy, principles, applications, and future directions. *Remote Sensing*, 10(4): 527 [DOI: 10.3390/rs10040527]
- Zhu X L, Chen J, Gao F, Chen X H and Masek J G. 2010. An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions. *Remote Sensing of Environment*, 114(11): 2610-2623 [DOI: 10.1016/j.rse.2010.05.032]
- Zhu X L, Helmer E H, Gao F, Liu D S, Chen J and Lefsky M A. 2016. A flexible spatiotemporal method for fusing satellite images with different resolutions. *Remote Sensing of Environment*, 172: 165-177 [DOI: 10.1016/j.rse.2015.11.016]
- Zhu X L, Zhan W F, Zhou J X, Chen X H, Liang Z C, Xu S and Chen J. 2022. A novel framework to assess all-round performances of spatiotemporal fusion models. *Remote Sensing of Environment*, 274: 113002 [DOI: 10.1016/j.rse.2022.113002]
- Zhukov B, Oertel D, Lanzl F and Reinhäkel G. 1999. Unmixing-based multisensor multiresolution image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 37(3): 1212-1226 [DOI: 10.1109/36.763276]

Review of deep learning–based spatiotemporal fusion of multisource remote sensing images

KONG Weiquan^{1,2}, QIAN Yurong^{1,2,3}, BAI Lu^{1,2}, FAN Yingying^{1,2}, LUO Sen³, GONG Weijun⁴

1. School of Computer Science and Technology, Xinjiang University, Urumqi 830046, China;

2. Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing, Urumqi 830046, China;

3. Xinjiang Engineering Research Center of Big Data and Intelligent Software, School of Software, Xinjiang University, Urumqi 830091, China;

4. School of Information Technology and Communication, Hexi University, Zhangye 734000, China

Abstract: With the continuous advancement of sensor technology and earth observation capabilities, remote sensing data have become a fundamental information resource supporting scientific research and socioeconomic development, and the demand for high-spatiotemporal-resolution remote sensing imagery has become increasingly urgent. However, constrained by the inherent trade-off between scanning swath width and pixel size, a single satellite sensor can hardly obtain remote imagery with high spatial and temporal resolutions simultaneously. This limitation creates a dilemma in earth observation tasks—such as vegetation phenology monitoring, crop growth monitoring, disaster assessment, urban planning, water quality monitoring, and land cover change detection—where data often offer “high timeliness but lack of detail” or “high precision but lack of continuity.” Consequently, this constraint significantly limits the application potential of remote sensing imagery.

Spatiotemporal fusion of multisource remote sensing imagery offers a viable solution by integrating data from different sensors. By blending high-spatial/low-temporal resolution images with low-spatial/high-temporal ones, this approach mitigates the uncertainty and incompleteness inherent in a single data source. The resulting fused imagery achieves high spatial and temporal resolutions simultaneously. This approach effectively alleviates the trade-off between spatial and temporal resolutions in a single satellite sensor. It also meets the practical demand for remote sensing data with high spatial and temporal resolutions.

However, owing to the gap between physical models and real-world conditions, unreasonable prior assumptions, and reliance on handcrafted features, traditional fusion methods still suffer from limited performance when handling highly nonlinear relationships or large-scale multisource data. Some methods also perform poorly in heterogeneous regions and struggle to effectively fuse multisource remote sensing data across spatiotemporal scales. The emergence of deep learning has introduced advanced spatiotemporal fusion techniques, such as convolutional neural networks and diffusion models. These approaches offer strong representation learning capability, high adaptability, and end-to-end modeling advantages. They can effectively extract multisource spatiotemporal features and provide a new paradigm for spatiotemporal fusion.

To keep pace with recent advances in this field, this study provides an in-depth analysis of deep learning-based multisource remote sensing image spatiotemporal fusion methods in accordance with their underlying principles and characteristics. These methods are categorized into five major classifications, namely, convolutional neural networks, transformers, generative adversarial networks, recurrent neural networks, and diffusion models. Subsequently, commonly used open-source datasets, deep learning algorithms, and performance evaluation metrics for remote sensing image spatiotemporal fusion are summarized and introduced. Curated links to datasets and model codes are also provided to support future research in this field.

In this study, state-of-the-art methods, including traditional and deep learning-based approaches, are selected for quantitative and qualitative evaluations across the CIA, DX, and SW datasets. The spatiotemporal fusion algorithms are validated, and a comparative analysis of their respective strengths and limitations is conducted on the basis of the fusion results. Finally, the development of various types of spatiotemporal fusion algorithms and the results of experimental analyses are synthesized to summarize the main challenges faced in the field of spatiotemporal fusion of remotely sensed images at the present stage and the potential future development directions.

Key words: spatiotemporal fusion, multi-source remote sensing images, deep learning, algorithmic principles, technique classification, datasets

Supported by National Natural Science Foundation of China (No. 62266043, 62566021); Excellent Youth Foundation of Xinjiang Uygur Autonomous Region (No. 2023D01E01); Outstanding Young Talent Foundation of Xinjiang Uygur Autonomous Region (No. 2023TSYCCX0043); Tianshan Innovation Team Program of Xinjiang Uygur Autonomous Region (No. 2023D14012); Key Research and Development Program of the Xinjiang Uygur Autonomous Region (No. 2023B01029-1, 2023B01029-2); Graduate Research and Innovation Program of Xinjiang Uygur Autonomous Region (No. XJ2025G089, XJ2025G036); Natural Science Foundation of Xinjiang Uygur Autonomous Region (No. 2022D01B123); Natural Science Foundation Project of the Basic Research Program under the Gansu Provincial Science and Technology Plan (No. 25JRRG037)