Semiblind Compressed Sensing: A Bidirectional-Driven Method for Spatiotemporal Fusion of Remote Sensing Images

Peng Liu, *Member, IEEE*, Lizhe Wang, Jia Chen, and Yongchuan Cui, *Graduate Student Member, IEEE*

***Abstract*—Spatiotemporal remote sensing imaging is one of the most important ways to continuously monitor the Earth. Due to some technical limitations, it is still not easy to obtain images with high-temporal-high-spatial resolution. In this article, we propose a new spatiotemporal remote sensing image fusion method with semiblind deep compressed sensing (SDCS). The reconstruction by SDCS includes two stages: compressed sensing observation and deep post processing. In the stage of CS observation, we design a sensing matrix to connect two spatiotemporal sequences. It can make sure that both the RIP condition of CS and the correspondence of spatiotemporal features are satisfied at the same time, and then CS observation provides a good initial estimates. In the stage of deep postprocessing, it is data-driven, and we designed a deep CNN architecture with multivariate activation function. The second stage not only smoothes out the noise but also reduces the errors from unprecise sampling matrix and compensates for the image differences caused by different imaging conditions. The proposed method is tested on two Landsat and MODIS datasets. Some of state-of-the-art algorithms are comprehensively compared with the proposed SDCS. The experiment results and ablation analysis confirm the better performances of the proposed method when compared with others.**

***Index Terms*—Compressed sensing, data-driven, image fusion, model-driven.**

I. Introduction

TIME-SERIAL remote sensing images have attracted considerable attention of researchers in recent years. Temporal and spatial resolutions are two of the most essential characteristics of remote sensing data. Spatio-temporal remote sensing image fusion (STRSIF) is a crucial technique in the field of remote sensing. It combines the advantages of remote sensing images with different spatial and temporal resolutions and then generates fused images with both high spatial and temporal qualities.

In many remote sensing applications, we often need images that can accurately represent both the detailed spatial features and the densely temporal changes of the observed area. However, due to limitations in sensor technology and costs, a single remote sensing image source usually cannot meet these dual requirements. Some remote sensing satellites, such as CBERS, SPOT5, GF2 [1], and Landsat provide data with high-spatial-low-temporal (HSLT) resolution. Other satellites, such as MODIS, AVHRR, and SPOTVGT, provide high-temporal-low-spatial (HTLS) data because of their short revisit cycles. More and more applications require data with high-spatial-high-temporal (HSHT) resolution, especially for situations such as land use/cover mapping, change detection, monitoring ecosystem dynamics, etc.. STRSIF enables us to merge images captured at different times and with different spatial resolutions, providing more comprehensive and timely information or DEM [2], [3] for various fields such as environmental monitoring, urban planning, agricultural management, disaster prevention, etc. By integrating temporal and spatial information, we can better understand the dynamic processes and changes occurring on the Earth's surface. Therefore, developing the STRSIF [4] technology is highly significant [5], [6].

However, due to the limitations of orbit and sensor, it is challenging to generate remote sensing images with both high temporal and extreme spatial resolution. Follow the summary of [7], for remote sensing spatiotemporal fusion, the difficulties mainly come from three points:

1) in the temporal dimension, the dramatic uncertainty of land feature change is hard to predict;
2) in the spatial dimension, the huge difference between high-resolution images and low resolution; and
3) for different sensors, there are inevitably systematic errors in the imaging process.

In this article, we analyze some essential problems of the observation model of remote sensing image and proposed semiblind compressed sensing (CS) for image fusion. The rest of the article is organized as follows: Section II is related works. We review the observation of imaging in Section III and summarize three fundamental problems in Section IV and introduce CS in Section V. In Section VI, the problem definition and variable definitions are addressed and the proposed method is presented

Received 4 July 2024; revised 12 September 2024; accepted 13 September 2024. Date of publication 26 September 2024; date of current version 24 October 2024. This work was supported in part by the National Science Foundation under Grant 41925007, Grant 41971397, and Grant 41471368. *(Corresponding author: Lizhe Wang.)*

Peng Liu and Yongchuan Cui are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, People's Republic of China, and also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100094, People's Republic of China.

Lizhe Wang and Jia Chen are with the School of Computer Science, China University of Geosciences, Wuhan 430074, China (e-mail: lizhe.wang@foxmail.com).

Digital Object Identifier 10.1109/JSTARS.2024.3463750

in detail and the main flowchart is given. The experiments and results are shown in Section VII. Finally, we conclude the article in Section VIII.

II. Related Works

Current STRSIF algorithms can be roughly divided into several categories [8] such as: weight methods, unmixing methods, Bayesian methods, sparse-learning methods, deep-learning methods, and hybrid methods, etc.

Weight methods [9], [10], [11], [12], [13], [14] estimate the change of information through the linear combination of HTLS pixels according to the weight of time, spectral, and distance, and then adds the change to the HSLT image. The classical method in this category is STARFM [9] algorithm. STARFM assumes an ideal case and designs weights by taking neighborhood HSLT pixels in various factors to obtain an HTHS image. There are some promotion methods: such as STAARCH [10], ESTARFM [11], mESTARFM [12], and ISKRFM [13]. A single weight-based method cannot ideally deal with STRSIF well since it assumes a linear relationship basically. The trend is to combine it with other methods or find the relationship by more sophisticated models such as Fit-FC [14], which introduces a regression model based on STARFM's strategy.

Unmixing methods [15], [16], [17], [18], [19], [20] consider that low spatial resolution pixels are linear mixing of endmember contained in high spatial resolution images. These methods generally define endmembers by preclassifying high spatial resolution images, such as [16] and [21]. Only using pixel unmixing, the fusion equation is difficult to solve. Some scholars have combined STARFM and pixel unmixing methods to carry out spatiotemporal fusion [22], [23]. Based on pixel unmixing, fusion can be reduced to the inverse problem, which has a clear physical meaning, and the abrupt details of land types contained in HTLS images can be recovered to a certain extent. However, this kind of method focuses more on the relationship between spectrum and spatial domain. In the time domain, it is usually assumed that the proportion (abundance) is unchanged for different substances in the mixed pixel with low spatial resolution. This assumption of abundance invariance seriously weakens the ability of the model to coordinate spatiotemporal relationships, so the final fusion results often show obvious blocky characteristics similar to the classification map.

Bayesian methods [24], [25], [26], [27], [28] consider that the STRSIF problem can be regarded as a maximum posterior problem for solving the optimal state with known observations. Therefore, how to define the relationship between the input HTLS and HSLT images and the output HTHS images in the form of probability becomes the key to this type of method. For example, covariance functions are used in [24], low-pass filterings are used in [25], and joint covariances are used in [28]. In general, numerous hyperparameters are often difficult to set in Bayesian methods.

Sparse representation method mainly uses a specific basis function to represent image sparsely and fuse them in the transform space. The early research such as SPSTFM [29] introducing the popular nonanalytic dictionary learning into the study of spatiotemporal image fusion. Later improvements include reducing the number of images and increasing flexibility [29], adding error boundary normalization [30], introducing structural sparse [31], etc. The method based on sparse representation often draws on the training method of double dictionaries in the superresolution reconstruction method [32]. Under certain conditions, their performance is better than the early linear method such as STARFM. If the low-resolution data are considered as the result of downsampling of high-resolution data, then the problem of sampling reconstruction needs to be considered. In this case, methods based on CS can be regarded as the extension of sparse representation methods. Under the CS theory [33], the reconstruction of the sampled signal no longer depends on the bandwidth of the original signal, but depends on the structure of the information in the signal and whether the sensing matrix meets the isometry constraint (RIP) [33] conditions. In the field of remote sensing image fusion, multispectral-panchromatic image fusion [34], [35] have employed this idea. There is also some research which tried to solve the exploration of spatiotemporal image fusion based on CS [36]. However, these CS methods mainly focus on the processing of image spatial resolution relationship, and cannot fully explain the contradiction of spatiotemporal opposition caused by different sampling frequencies among spatiotemporal remote sensing datasets. Therefore, its performances are often limited because the resolution difference is large and the sampling matrix is not totally known.

Deep learning methods draw on the large amounts of data and the progress of deep learning in superresolution reconstruction [37] to establish implicit relationships between low and high resolution images [38]. Deep learning methods have powerful feature extraction ability with the support of Big Data. Through the design of network structure and loss function [39], the adaptability to spatiotemporal fusion can be improved, such as STFDCNN [40], DCSTFN [41], EDCSTFN [42], StfNet [43], and BiaSTF [44]. GAN methods show many advantages in spatiotemporal fusion, such as STFGAN [45], GAN-STFM [46], and CycleGAN-STF [47]. However, the architecture of generator or discriminator, the loss function, and feature extraction, etc., all need to be consider further to adapt for spatiotemporal fusion. Deep-learning based approaches have a lot of potential, but they also face obvious problems: On the surface, spatio-temporal fusion in remote sensing often suffers from large resolution differences, which will lead to unstable results. The deeper reason is that fusion of spatiotemporal remote sensing data faces complex contradiction relationship between multiple sampling sequences. Some existing deep learning frameworks that simply increase the depth of neural networks or simply focus on the way of network connections. These architectures do not clearly reflect the nature of spatiotemporal relationship of multisource remote sensing images, so that they are not conducive to exerting the powerful feature learning ability of deep networks.

Hybrid methods [21], [22], [23], [48], [49], [50] are tend to combine the advantages of several above mentioned methods to achieve the goal of improving STRSIF accuracy. Methods such as FSDAF [48] combined with the idea of unmixing and weighting, which solves the information variation in the temporal domain (by unmixing method) and spatio domain (by

spatial interpolation). Another example is STRUM [23], which directly blends changes in HTLS pixels by Bayesian theory to estimate changes in HSLT endmembers for STRSIF. In the future, the forms hybrid methods may be more diverse so that the advantages of different approaches can be further taken advantage of.

From another point of view, we can divide abovementioned methods into two large categories: model-driven methods and data-driven methods. Model-driven methods such as unmixing methods [16], CS method [36], and BRDF [51] method are all with clear physical meaning. On the contrary, data-driven methods such as deep learning method, they do not take much emphasis on physical meaning but data features. Although, data-driven methods (especially deep learning) already show very promising results, its stability and theory boundary is not as clear as model-driven methods (such as CS observation model). At the same time, in fusion process, the large resolution gap, unprecise sampling matrix, and complex imaging condition etc. always make the conventional CS approach less effective. Therefore, in this article, we proposed to connect a blind CS model with deep learning to form semiblind deep compressed sensing (SDCS). In this new method, it include two stages: CS observation stage and deep reconstruction stage, and then the advantages from both CS observation model and deep feature learning can be well utilized. In summary, the contributions of this work are:

1) Summarize the imaging model and its three fundamental problems, and propose that CS as a special form of imaging model can be combined with deep learning to achieve fusion.
2) In the CS observation stage, design a sensing matrix satisfying both sampling mapping and RIP condition to provide an initial fusion estimation.
3) In the deep learning stage, design a deep architecture with multivariate activation function (MAF) to further improve the fusion effect.

In the following, we will elaborate on the proposed method and verify the results based on comprehensive experiments.

III. Model of Degradation for Remote Sensing

Generally, we can take the process of remote sensing imaging as the process of signal acquisition. In this process, we get the signal of the ground object through the sensor. The common forms of imaging are varied, such as optical, SAR, hyperspectral, and even point clouds, etc. There are also geoprocesses that need to be monitored based on special sensors, which are just discrete signals rather than images. In almost all of them, there are information losses or damage, so imaging is a degrading process. As in Fig. 1, most of degradation processes of remote-sensing images (major in optical systems) can be expressed as

$$Y = \mathcal{H} * X + \epsilon = \mathbf{H}_1\mathbf{H}_2 \cdots \mathbf{H}_k X + \epsilon \tag{1}$$

where Y is the degraded and observed image, X is the original signal, $\mathcal{H}$ is the degradation matrix, and ϵ is the additive noise. The operator $*$ is the convolution. The noise ϵ is often assumed as a zero-mean white Gaussian process with variance σ^2. The total degradation $\mathcal{H} = \mathbf{H}_1\mathbf{H}_2 \cdots \mathbf{H}_k$, where $\mathbf{H}_k$ is the degradation by a single factor, for example, $\mathbf{H}_1$, $\mathbf{H}_2$, and $\mathbf{H}_3$ can be the degradation in temporal domain, spatial domain, spectral domain, etc. Images Y and X can be denoted by vectors, that $Y = [y_1, \ldots, y_N]^T$ and $X = [x_1, \ldots, x_N]^T$. $\mathbf{H}_k X$ is the matrix-vector form, where $\mathbf{H}_k$ is convolution matrix which is approximated by a block-circulant matrix.

The representation of different types of image degradation, such as blurring, thin cloud, missing information, noise, shadow, and downsampling, mainly depends on the assumption on the degradation matrix $\mathbf{H}_k$. For example, when $\mathbf{H}_k$ is the unit matrix we consider the image quality improvement as a denoising problem; when $\mathbf{H}_k$ is the Gaussian like kernel we consider the image quality improvement as a deburring problem; when $\mathbf{H}_k$ is the downsampling matrix we consider it as an image fusion or superresolution problem; when $\mathbf{H}$ represents the atmospheric scattering model we consider the image quality improvement as a thin cloud removal problem; when the $\mathbf{H}_k$ is the process of Ray-casting reflection we consider the image quality improvement as a shadow removal problem.

There have developed many different theories and methods for different degradation model of $\mathbf{H}_k$. $\mathcal{H}$ is the result of the combined action of multiple degradation processes $\mathbf{H}_k$, but many studies only consider one or a few degradation processes.

IV. Three Fundamental Problems in Image Reconstruction

Considering the observation model in (1), a general form of image quality improvement (or reconstruction) problem can be denoted as

$$T(\mathcal{H}, X) = \|Y - \mathcal{H} * X\|_2^2 + \lambda_2\psi(X) + \lambda_1\phi(\mathcal{H}) \tag{2}$$

where $\| - \|_2^2$ is L_2 norm, $\psi(X)$ is the constraint for image, and $\phi(\mathcal{H})$ is the constraint for transfer function.

Different types of degradation usually vary widely and lead to different methods of solution. In this article, we believe that, for all the studies of image quality improvement, there are always three fundamental problems: 1) transfer function, 2) regularization schemes, and 3) noise status. In this section, the three fundamental problems will be comprehensively addressed in detail.

A. Sensing Matrix (Transfer Function)

Among the three fundamental problems, the sensing matrix or degradation matrix $\mathcal{H}$ is the most important. It is used to describe the mapping between the observed image (Y) and the ideal image (X). It represents the most essential characteristics of the different degradation processes. We usually discriminate the type of degradations such as blurring, noising, shading, clouding, etc., based on its observation model.

What is $\mathcal{H}$ like and how to define a $\mathcal{H}$ for degradation are often the keys to image reconstruction. First, we have to define a $\mathcal{H}$ for image quality improvements before all things we can do. For example, for thin cloud removal $\mathcal{H}$ is pixelwise form because the degradation for each pixel has nothing to do with its neighborhoods; for a deblurring problem, $\mathcal{H}$ is a kernel function

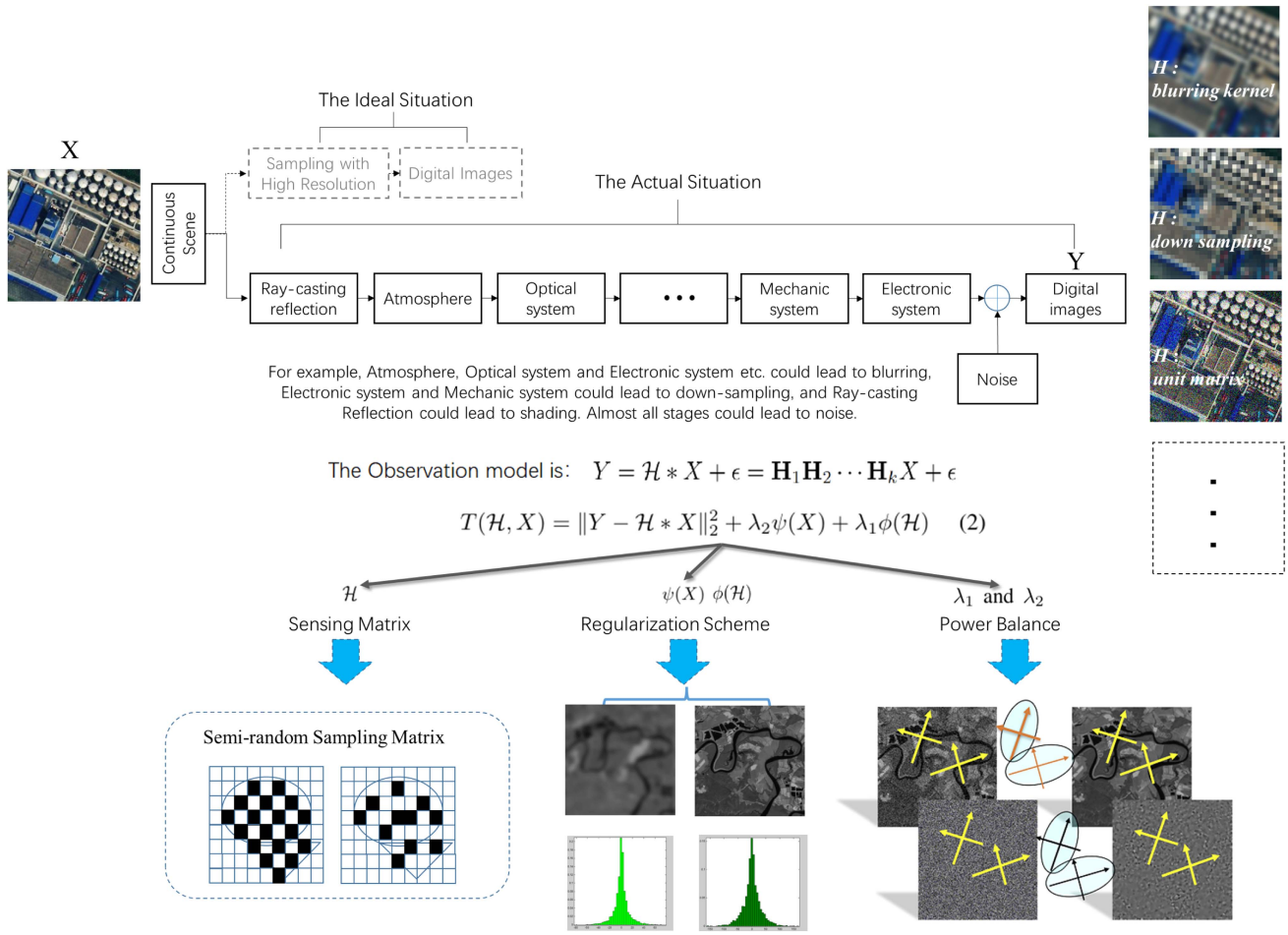

Fig. 1. Observation model and three fundamental problems. Image fusion is a special form of image quality improvement of generalized form.

who has far more small support than an image because the blurring effectiveness usually only extend in a small area; for fusion or superresolution problem, $\mathcal{H}$ may be a downsampling matrix because we focus on the spatial resolution or spectral resolution. However, how to define a good $\mathbf{H}_k$ or $\mathcal{H}$ is still very open problem for some type of degradation. Second, it is how to solve the model with this $\mathcal{H}$. A corresponding solution method must be developed when $\mathcal{H}$ is defined. The most simple one is denoising when $\mathcal{H}$ is a unit matrix. For the deblurring problem and some thin cloud removal problem, $\mathcal{H}$ cannot be inverse directly because of its underdetermined characteristics. Since the rank of $\mathcal{H}$ is often smaller than the dimension of $\mathcal{H}$, it is a singular matrix in most cases. Most image quality improvement problems are ill-posed or seriously ill-posed problems. $\mathcal{H}$ often is unknown making many observation equations very hard to solve in nature.

The most famous conclusion about $\mathcal{H}$ may be sampling theorem. The Nyquist–Shannon sampling theorem provides a sufficient condition for the sampling and reconstruction of a band-limited signal. The popular CS theory in recent years can be regarded as an upgraded version of the sampling theorem. CS believes that data reconstruction is mainly determined by three factors: the structure of the signal, the way of sampling, and the algorithm for solving. In this way, we have opened a new door for data reconstruction research.

In the later chapters of this article, we will discuss in detail how to carry out STRSIF based on CS theory. We can implement image fusion from three aspects:

1) A CS observation model is used to describe the mapping between the observed image (Y) and the ideal image (X);
2) The deterministic constraint is obtained by using the spatial position correspondence between the observed image and the ideal image;
3) The image should be transformed to be sparse, and $\mathcal{H}$ need to meet the requirements of the CS observation model by randomness or other characteristics.

B. Regularization Schemes

Regularization schemes are used to introduce prior knowledge and allow a robust approximation of ill-posed inverse problems, which is the second fundamental problem. Since most degradation models are seriously ill-posed, the remote sensing image quality improvement based on (2) will have to employ regularization to find a stable solution. Both X and $\mathcal{H}$ need regularizations, as well as $\psi(X)$ and $\phi(\mathcal{H})$. The research on

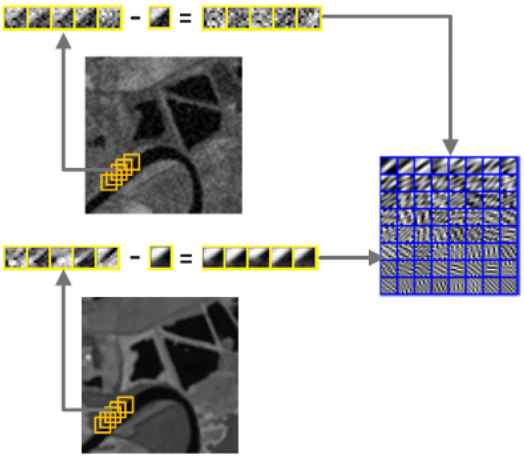

Fig. 2. Training a dictionary by group patches from both the target and reference images [52].

defining $\psi(X)$ is more common and prosperous. In recent years, the study of regularization has developed very fast. Almost all smooth schemes in low-level vision tasks can be used as regularization in remote sensing image quality improvements.

Regularization schemes are prior knowledge on X: It is an assumption of data characteristics when solving the equation. The most common regularization is L_2 norm regularization. Due to the popularity of variational image processing and CS, the regularization of L_1 norm also plays an important role in the field. Except for L_2 and L_1 regularization, many advanced methods for low-level vision tasks are also used as regularization. For example, as in summary of [52], the TV or partial-differential equation method [53] makes use of the geometric features of the image, the wavelet method makes use of the statistical features of the wavelet coefficients [54], the MRF method makes use of the relationship of neighborhood pixels [55], the dictionary learning method [56] mainly makes use of the sparsity of the representation coefficients, the nonlocal means method [57] makes use of the redundancy in the image texture features, and the expected patch log-likelihood method [58] makes use of the statistical features of the image patches. There are also excellent methods, such as block matching and 3-D filtering (BM3D) [59], which employ both the redundancy in texture and the sparsity in the transform domain. These different regularizations are based on different theories and assumptions, and they all exhibit promising performances. All methods are used for image-denoising and finally employed for regularization schemes. They are all assumptions on the characteristics of the data. Of course, the now popular deep learning can also be seen as a general form of regularization, and this regularization based on big data shows more powerful capabilities. For the regularization in image fusion, it is crucial to measure and correlate the characteristics of multisource observation data to form complementary information gain. As the example in Fig. 2, only by constructing a regularization term based on complementary information gain, which provides beneficial information of multiple sources for solving the original data X, can regularization be used to support fusion problems.

Regularization and Law of Geography: We propose that regularization is the concrete embodiment of "The First Law of Geography" in the field of remote sensing data reconstruction. The First Law of Geography, according to Waldo Tobler, is "everything is related to everything else, but near things are more related than distant things" [60]. This first law is the foundation of spatial dependence and spatial autocorrelation and is used specifically for the inverse distance weighting method for spatial interpolation and to support the regionalized variable theory for kriging [61]. The first law of geography is the fundamental assumption used in all spatial analysis [62]. Very similar, regularization is exactly this kind of assumption of data, but it is for remote sensing observation data. In the community of machine learning, the basis or dictionary in sparse representation, the patches in nonlocal means, or the convolution in deep learning, are all concrete methods of regularization to find similarities or relationships in the data by transforming them into a new space. Furthermore, we need to refer to "the second law of geography" as well as Law of Spatial Heterogeneity. In practical applications, the corresponding regularization methods may be derived from the way of the heterogeneity.

We should note that in CS theory, data properties such as sparsity are considered together with the sensing matrix, which is considered to be a very advanced concept in this article. The idea is promising, and not just limited to CS. Of course, this also shows the importance of regularization, because we need the assumption of data characteristics for a good reconstruction.

C. Noise Status (Balance of Power)

The noise status will determine the Balance of Power in the solution of the object function (2), which is the third fundamental problem. In the equation for image quality improvement, the fidelity term mainly provides the power of antidegradation, while the regularization term mainly provides the power for equation stability. The most important characteristic of the two powers is that they are the unity of opposites. First, they are oppositive because antidegradation will enhance the texture details of the image but regularization will smooth out a part of the texture details. Second, they are unitive need each other because antidegradation can provide input information for regularization and regularization can suppress large noise in antidegradation to avoid contaminating the images.

As in (2), λ (λ_1 and λ_2) denote the balance between the two powers of antidegradation and regularization. In most cases, it is taken as a hyperparameter problem. First, since the ground truth image is unknown, we cannot directly judge how much noise and how many texture details there are in the current image. As a result, it is very hard to set the values of λ when solving the equation. Second, the fidelity term is often a linear form, but the regularization term can be either a linear or nonlinear form. If the initial noise in the observation data Y is known and the regularization term is linear, the values of λ can be precisely estimated. It means that we can set the optimal hyperparameter in solving the linear equation. However, most of advanced image quality improvement methods involved highly nonlinear regularization or even nonlinear fidelity. As a result, the strength

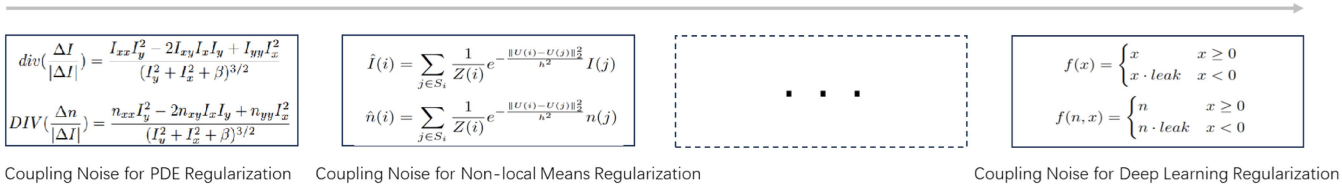

Fig. 3. Examples of coupling noise with synchronization simulation for different advanced nonlinear regularization. For PDF regularization, the synchronization equation is constructed with gradients. For nonlocal means regularization, the synchronization equation is constructed with the kernel of features. For deep learning regularization, the synchronization equation is constructed with activation functions. The synchronization equation model present the relationship between noise state and quality improvement. The noise equation interacts with the data equation and iterates synchronously. The correlation between noise state and quality improvement process is quantified, the statistical characteristics of residual noise can be accurately estimated, and the two processes of noise removal and quality improvement can be coordinated.

of the two powers is both anisotropic and dynamic, especially for nonlinear regularization. This is the fundamental cause of the difficulty in finding the balance between the two powers.

Coupling noise with synchronization simulation: The regularization parameter maintains the balance between the regularization term and the fidelity term. The selection of an appropriate regularization parameter has been the object of studies in the field of inverse problems, and some classical methods address the issue, such as cross-validate, L-curve, discrepancy principle, and Bayesian estimation etc. However, these classical methods are hard to directly use in advanced nonlinear regularization. To correctly estimate the status of the noise remaining in the image, we believe that coupling noise by synchronization simulation is promising [63]. As in Fig. 3, it synchronously iterates a synthesized noise with the observed image in the solving procedure. The similarity in the statistical properties of the real noise and the synthetic noise can be maintained in iteration. We then establish the relationship between the statistical characteristics of synthetic noise and the regularization parameter. In every iteration, the regularization parameter can be calculated by using a derivable formula for the relationship.

D. Deep Learning Versus Imaging Model

For the early studies of data reconstruction, an explicit model with a clear observation model plays more important role. However, with the development of deep learning and automatic feature learning, it is possible to weaken the observation model such as (1). We can find that based on popular deep learning methods in recent years, it seems possible to reconstruct data without these three fundamental problems. However, the three fundamental problems with data reconstruction have not gone away. The deep learning approach, because it is an implicit model, actually solves them in a special way all together. The perception matrix that plays the role of mapping and the regularization that maintains the stability of the equation have changed their forms, and they are implemented in the ways of convolution, connection, activation, pooling, even attention, etc. The process of fusion can be easily realized in the hidden layers of CNN or other deep architectures, where the physical meaning of the observation or data distribution model is dramatically weakened. We call this kind of trend of fusion as implicit model and data-driven, which is denoted as in [4].

In a typical method of implicit model and data-driven, we do not need to consider the observation equation $Y = \mathcal{H} * X + \epsilon$. One of the advantages is that the fusion process can avoid some problems resulting from the complexity of $\mathcal{H}$, especially for heterogeneous remote sensing data, etc. However, the explainability or interpretability of the fusion becomes so weak that we cannot judge how and why the fusion produces better performances in some cases. Furthermore, data-driven methods often show overfitting in some cases. Another problem is that data-driven reconstruction usually needs more training data and more computation sources. We believe that both model-driven and data-driven methods have their advanced merits. Explicit reconstruction with observational models is still very valuable, especially in fusion reconstruction scenarios that require clear physical meaning. There are already studies that try to combine the two schemes in fusion. This article will also develop this bidirectional-driven approach, based on semiblind CS, to achieve spatiotemporal fusion.

V. Compressed Sensing

In this section, we address the model of CS, which is a special case in observation model and data reconstruction and will be used in the following process of spatiotemporal data fusion. CS [33] is an efficient way to acquire and reconstruct a signal from a series of sample measurements. For an original remote sensing signal $X \in R^A$, similar to (1), the observation $Y \in R^B$ can be represented as

$$Y = \Phi X + \epsilon \tag{3}$$

where Φ is the measurement matrix mapping from R^A to R^B, B is typically much smaller than A, and ϵ is the noise. Matrix Φ represents a dimensionality reduction. Many studies have been made for restoring X from Y. For signal X, such as remote sensing image, it is usually sparse in some domains and can be represented by the basis D:

$$X = D\alpha \tag{4}$$

where α are the coefficients of X with the basis D, and D is an atom set which is also denoted as a dictionary. With above definitions in (3) and (4), observation Y is expressed as

$$Y = \Phi D\alpha + \epsilon. \tag{5}$$

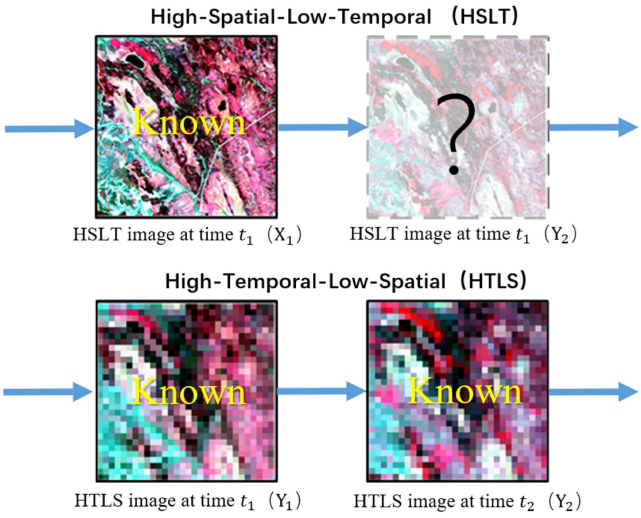

Fig. 4. Problem definition. The HSLT images and HTLS images are at time t_1 and t_2. The unknown image at time t_2 needs to be estimated.

TABLE I
SYMBOL TABLE

Symbol	Definition
X_k, Y_k	Original HTLS image and HSLT image at time k, $k = 1, 2$.
X_k^i, Y_k^i	Arbitrary the ith HTLS image or HSLT image at time k.
$\mathbb{X}_k^i$, $\mathbb{Y}_k^i$	Patch set for the HSLT image X_k^i and HTLS image Y_k^i, and $\mathbb{X}_k = \{\mathfrak{x}_k^{(i,c)}\}_{c=1}^C$ and $\mathbb{Y}_k = \{\mathfrak{y}_k^{(i,c)}\}_{c=1}^C$, where $\mathfrak{x}_k^{(i,c)}$ and $\mathfrak{y}_k^{(i,c)}$ are subclass patches.
$\mathcal{X}$, $\mathcal{Y}$	$\mathcal{Y}$ is the matrix for $[\sqrt{\eta_1}Y_1, \sqrt{\eta_2}Y_2]^T$, and $\mathcal{X}$ is $[\sqrt{\eta_1}X_1, \sqrt{\eta_2}X_2]^T$.
Φ	Sampling matrix in a standard compressed sensing.
D	Over complete dictionary trained from the corresponding X_k. We assume that $D_1 \approx D_2$.
α	Sparse coefficients obtained by decomposing image data with the dictionary D.
M_k	Sampling matrix between HTLS image Y_k and HSLT image X_k.
$\mathcal{F}(\cdot)$	Deep CNN sturcture of second stage estimates
$g_m(\cdot)$	MAF in $\mathcal{F}(\cdot)$

Finally, the objective function of a CS problem is

$$\min_{\alpha} |\alpha|_0 , \text{ subject to } Y = \Phi D\alpha. \tag{6}$$

The minimization of ℓ_0-norm in (6) can be converted to ℓ_1-norm problem when restricted isometry property (RIP) [64], [65], [66] (for Φ and D) is satisfied. By this way, the converted problem of ℓ_1-norm is solvable in polynomial time. Imposing ℓ_2-norm on the data-fitting term, and applying a Lagrangian form, (6) becomes

$$\min_{\alpha} |\alpha|_1 + \lambda \|Y - \Phi D\alpha\|_2^2 . \tag{7}$$

In the next section, CS model is applied to spatiotemporal data fusion and used to present the resolution relationship between multisource images.

VI. Spatiotemporal Data Fusion

A. Problem Definition of Spatiotemporal Data Fusion

In this article, the known and unknown variables in the STRSIF are shown in Fig. 4. For the task of STRSIF, all images need to be accurately matched in a unique geographic location. Both HTLS and HSLT images are calibrated to the same physical quantity. The images involved in STRSIF are: one pair of images with HSLT image X_1 and HTLS image Y_1; and one pair of images with HSLT image X_2 and HTLS image Y_2. The unknown HTHS image X_2 is the target that needs to be estimated. For convenient, the symbol definitions used in this article are listed in Table I.

In this part, we design a two-stage deep-CS model for spatiotemporal image fusion. At the first stage, the observation model of (3) is taken as forward model, and the observation matrix is only partly known. At the second stage, the deep networks as postprocessing of CS are added into our fusion. This postprocessing not only smooths out the noise and artifact in the initial estimation but also compensates for the errors from the unprecise downsampling matrix M. Many studies [67], [68] [69], [70] categorized this method into deep CS. We can also call it bidirectional-driven method as it includes in both forward model-driven and backward data-driven.

B. Spatiotemporal Data Fusion With CS

For the problem of spatiotemporal data fusion, we use the CS to represent the relationship of images with different spatial-resolutions. As in Fig. 4, our goal is to predict an image of with high spatial resolution of Landsat ETM+ at time k with one pairs of Landsat ETM+ and MODIS images (acquired at $k-1$) and one MODIS image (acquired at k). For arbitrary time k, there is

$$Y_k = M_k X_k + \epsilon_k \tag{8}$$

where the MODIS image Y_k is defined as the observation downsampled from X_k, and M_k (same as Φ_k) represents the downsampling operation. They are similar to the observation equation of CS. If α_k is sparse enough ($X_k = D_k\alpha_k$) and the sensing matrix $M_k D_k$ satisfy RIP condition, α_k (as well as X_k) can be solved uniquely in polynomial time.

It seems like a solvable problem since it is not difficult to find D_k that can sparsely represent spatiotemporal data. However, in spatiotemporal data fusion, there are at least three issues that we need to give special considerations. First, spatiotemporal data fusion is based on two temporal sequences, so in the reconstruction we can make use both the sparse characteristics and the spatiotemporal correlation characteristics. Second, the measurement matrix M_k is unknown, which makes the it far more difficult than common CS problem. Third, due to the situation that images are acquired by different sensors in different conditions (such as time, spectrum, resolution, etc.), it is often difficult to ensure RIP. Therefore, extra constraints should be introduced into the fusion model based on (7), which will help us find its solution. Since they are two temporal sequences, we can assume that X_2 is not far from X_1. Then for time $k = 2$,

there is

$$L_\alpha(D,\alpha)=\lambda_1\|X_1-D\alpha\|_F^2+\lambda_2\|Y_2-M_2D\alpha\|_F^2+|\alpha|_1 \tag{9}$$

where $D\alpha=X_2$, and λ_1 and λ_2 are the parameters. We need to notice that we do not use D_1 and D_2 but only D. In time sequence images, we can assume that $D_1\approx D_2$, and they are both sparse enough to represent our target image X_2.

More important, in (9), there are two ways that provide information to our target X_2: one is the observation model of downsampling, the other is its similarity to the neighborhood image X_1. Obviously, it will be more suitable to find a X_2 in (9) than in a single CS model of (7). However, it is still not enough to reconstruct a fusion image X_2, because the measurement matrix M_2 is unknown and it is difficult to ensure RIP condition for CS.

Without a determined measurement matrix M_2, it will lead to a blind CS problem. Since we can know some information about M_2 by neighborhood images, it is a semiblind CS problem. In next section, we will discuss how to estimate a sampling matrix M_2.

C. Design Sampling Matrix for Spatiotemporal Data Fusion

One of the key points for CS reconstruction is the relationship between the sampling matrix (measurement matrix) and dictionary, which belongs to the first fundamental problem mentioned above. To meet this requirement, commonly used measurement matrices include Gaussian matrices, Bernoulli matrices, partial Fourier matrices, partial random Toeplitz or circulant matrix, and partial Hadamard matrix, etc. The coherence of a matrix falling within the lower and upper bounds helps us to reconstruct the image effectively.

However, in spatiotemporal fusion, we have not totally determined sampling matrix. It is semiblind CS since there is only a partly known sampling matrix. STRSIF with semiblind CS puts new requirements on constructing sampling matrix. On the one hand, the sampling matrix should reflect the projection relationship of different resolution in remote sensing images, rather than a completely random matrix. On the other hand, the sampling matrix needs to be incoherent with the sparse transformation matrix (dictionary) as far as possible to meet RIP condition.

As in (5) and (9), when α is sufficiently sparse, it is necessary to ensure that M_2 and D are incoherent. Obviously, matrix M_2 will be incoherent with D when it satisfies high randomness. However, the high randomness of M_2 in STRSIF is not practical. The reason is that M_2 represents a downsampling process from high spatial resolution to low spatial resolution, so it means that M_2 must be able to describe both the feature structure and sampling relationship of spatiotemporal images. Therefore, M_2 should not be a completely random matrix. In this article, we for first time proposed that the fusion problem needs to establish a semirandom observation matrix that satisfies both the spatial correspondence and the random property.

The semirandom observation matrix must exist, but it is not easy to find accurately. We can design the sampling matrix based on more relaxed conditions. Let us review the RIP condition. Let $\Sigma_{\mathbf{k}}$ be the union of all subspaces spanned by all subsets of $\mathbf{k}$ columns of M_2. The matrix M_2D has the restricted isometry property (RIP) adapt M_2 with $\delta_{\mathbf{k}}$ referring to [64], [65], [66] if

$$(1-\delta_{\mathbf{k}})\|\alpha\|_2^2\le\|M_2D\alpha\|_2^2\le(1+\delta_{\mathbf{k}})\|\alpha\|_2^2. \tag{10}$$

We can understand formula (10) as: M_2D will be better if this product matrix is almost orthogonal. In addition, since the time sequences are similar in image features, we have reason to assume $M_1\approx M_2\approx M$. When considering both the downsampling relationship between spatiotemporal data and the fundamental RIP condition in CS, we construct a new object function for measurement matrix as

$$\begin{aligned}L_M(M)&=\eta_1\|Y_1-MX_1\|_F^2\\&\quad+\eta_2\|Y_2-MX_2\|_F^2+\|\mathbf{I}-MDD^TM^T\|_F^2\end{aligned} \tag{11}$$

where $\mathbf{I}$ is a unit matrix. Term $\|\mathbf{I}-MDD^TM^T\|_F^2$ means $MD(DM)^T$ is very close to unit matrix $\mathbf{I}$, as well as MD is a orthogonal matrix. We actually implement the RIP of (10) by constraints of $\|\mathbf{I}-MDD^TM^T\|_F^2$. Since D is known, M is not necessary to be very high random but incoherence to a known D.

For an existing D, to find the solution of M, the object $L_M(M)$ is rewritten as

$$\begin{aligned}&L_M(M)\\&=\left\|\begin{bmatrix}\sqrt{\eta_1}Y_1\\\sqrt{\eta_2}Y_2\end{bmatrix}-M\begin{bmatrix}\sqrt{\eta_1}X_1\\\sqrt{\eta_2}X_2\end{bmatrix}\right\|_F^2+\|\mathbf{I}-MDD^TM^T\|_F^2.\end{aligned} \tag{12}$$

For the simplicity, it is changed as

$$L_M(M)=\|\mathcal{Y}-M\mathcal{X}\|_F^2+\|\mathbf{I}-MDD^TM^T\|_F^2 \tag{13}$$

where $\mathcal{Y}$ is the matrix for $[\sqrt{\eta_1}Y_1,\sqrt{\eta_2}Y_2]^T$, and $\mathcal{X}$ is $[\sqrt{\eta_1}X_1,\sqrt{\eta_2}X_2]^T$. We take the derivative of the object function with respect to M and get

$$\frac{\partial L_M}{\partial M}=2(\mathcal{Y}\mathcal{X}^T-M\mathcal{X}\mathcal{X}^T)+\frac{\partial\|\mathbf{I}-MDD^TM^T\|_F^2}{\partial M}=0 \tag{14}$$

where

$$\frac{\partial MDD^TM^T}{\partial M}=(DD^T+D^TD)M^T. \tag{15}$$

The derivative of the second term is

$$\begin{aligned}&\frac{\partial\|\mathbf{I}-MDD^TM^T\|_F^2}{\partial M}\\&=(\mathbf{I}+\mathbf{I}^T)MDD^T+4MDD^TM^TMDD^T.\end{aligned} \tag{16}$$

Bring (16) into (14), we have

$$2(\mathcal{Y}\mathcal{X}^T-M\mathcal{X}\mathcal{X}^T)+2MDD^T+4MDD^TM^TMDD^T=0. \tag{17}$$

Since $MDD^TM^T\approx\mathbf{I}$, there is

$$2(\mathcal{Y}\mathcal{X}^T-M\mathcal{X}\mathcal{X}^T)+2MDD^T+4MDD^T=0. \tag{18}$$

Rearrange it as

$$M(\mathcal{X}\mathcal{X}^T-3DD^T)=\mathcal{Y}\mathcal{X}^T. \tag{19}$$

The estimated sampling matrix is

$$M = \mathcal{Y}\mathcal{X}^T(\mathcal{X}\mathcal{X}^T - 3DD^T)^{-1}. \tag{20}$$

It is worth to notice that X_2 in $\mathcal{X}$ is not really known, so the sampling matrix M will be updated with each new $\mathcal{X}$ in the iteration.

In the CS stage of spatiotemporal data fusion, we need to first find a dictionary D, and then use (11) and (20) to find a sampling matrix M, and finally predict the target image X_2 by (9). The dictionary D is relative easy to train. In practice, (9) and (11) can also be alternatively solved as (21) shown at the bottom of this page.

When both D and M are solved, we finish the first stage of deep CS model. However, the semiblind CS reconstruction with a partly known M and over large resolution gap is still prone to instability only by regularization of l_1 norm as in (9). We need to introduce more information by deep features of images into the fusion process.

D. Add Deep Network as Postprocessing into Fusion

As mentioned above, a single CS reconstruction only by the first stage is hard to provide a precise and stable fusion for the target image. The constraint items of sparsity for representing the HTHS image can effectively maintain the fusion result to a certain extent. However, a single sparse regularization is not enough. This belongs to the second fundamental problem in Section IV. Due to the large difference in resolution between HTLS image and HSLT image and unprecise sampling matrix M, the fusion results will be degraded, especially for small detail features and changing areas.

In this case, we implement the second stage reconstruction by deep learning, which is also a special case of deep CS as pointed in [71]. The estimation in the second stage is defined as: $\tilde{X}_2 = \mathcal{F}(\hat{X}_2, Y_2, X_1 : \theta)$, where $\mathcal{F}(\cdot)$ which is a deep network architecture, θ are parameters in the network, and the reference image such as X_1 and Y_2 and initial estimates $\hat{X}_2$ (from CS) are inputs. Combining with object function $L_\alpha(D, \alpha)$, it is kind of deep CS by the taxonomy of [71]. It means that we need samples of true X_2 in the training stage, which are from other time sequences, and its loss function can be denoted as

$$Loss(\hat{X}_2) = \|\mathcal{F}(\hat{X}_2, Y_2, X_1 : \theta) - X_2\|_2^2 \tag{22}$$

where $\hat{X}_2 = D\hat{\alpha}$ and $\hat{X}_2$ is the initial estimation from CS, and X_2 is the ground truth image in training pairs. In this article, in addition to L_2 loss in (22), we also use L_1 norm and structure similarity loss at the same time.

Now we have a deep CS with the second stage in the fusion process as in Fig. 5. The deep CS with two stages belongs to both model-driven and data-driven. The information from CS observation is model-driven, but the information from deep learning $\mathcal{F}(\hat{X}_2, Y_2, X_1 : \theta)$ is data-driven. CS reconstruction with sampling matrix M is explicit but $\mathcal{F}(\hat{X}_2, Y_2, X_1 : \theta)$ reconstruction is implicit. They all have their advantages. CS model-driven is explainable and easy to known which features can be reconstructed. Deep feature with data-driven can utilize more information from external data and improve the performance of fusion. We can also call this method as bidirectional-driven model.

Specifically, as in Fig. 6, we design a parallel CNN structure for second stage $\mathcal{F}(\hat{X}_2, Y_2, X_1 : \theta)$, where the target image $\hat{X}_2$ and the reference image X_1 and Y_2 are all transformed into the similar feature space by the convolutional layers. Since we have a CS initial estimates, in the second stage, the fusion reconstruction does not adopt regular branch aggregation structure as many existing research. We proposed to integrate the information in a new way: features from different branches enter into a new MAF to fuse the reconstruction of X_2.

Before presenting the proposed MAF, we first review the active process of feature layers in conventional CNN. The well-known ReLU activation function [72] is defined as

$$f(x) = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{23}$$

where x is the feature value in the arbitrary index from the convolutional layers.

ReLU in (23) or Leaky-ReLU [73] as the activation function are commonly used in many networks as they are simply operation and improve the sparseness of the network. For information confluence from multiple branches, the judgments by a conventional activation function are not always correct. Especially when there are many errors from CS reconstruction, it is hard to decide how many details should be removed or reserved.

We believe that the priors from neighborhood data (Y_2 or X_1) can be introduced in deep reconstruction by a more direct way. In this article, we propose a parallel deep CNN structure (PD-CNN) with a MAF for the fusion in the second stage. The PD-CNN is constructed with three similar deep CNNs as in Fig. 6. The inputs are image $\hat{X}_2$, X_1, and Y_2, and the output is $\tilde{X}_2$. All the three branches have the same convolutional layers. To reduce system errors, $\hat{X}_2$ and X_1 are styled to Y_2 by AdaIN [74] function before entering into MAF. We can think that PD-CNN transforms all input image into the similar feature space. We need to compare the feature values of the three images, and determine which are the important image features and the error of CS reconstruction. Therefore, We consider using the information of the reference image to assist the activation of the features of target image. A new activation function g_m is defined as

$$g_m(x_{cs}, x_{a1}, x_{a2}, y_2) = \begin{cases} x_{a2} + x_{cs} & x_{a2} \geq 0, x_{a1} \geq 0 \\ x_{a2} \cdot l + x_{cs} & x_{a2} \geq 0, x_{a1} < 0 \\ x_{a2} \cdot l & x_{a2} < 0, x_{a1} \geq 0 \\ x_{a2} \cdot l + y_2 & x_{a2} < 0, x_{a1} < 0 \end{cases} \tag{24}$$

$$\begin{cases} L_\alpha(D, \alpha) = \lambda_1\|X_1 - D\alpha\|_F^2 + \lambda_2\|Y_2 - MD\alpha\|_F^2 + |\alpha|_1 \\ L_M(M) = \eta_1\|Y_1 - MX_1\|_F^2 + \eta_2\|Y_2 - MX_2\|_F^2 + \|\mathbf{I} - MDD^TM^T\|_F^2. \end{cases} \tag{21}$$

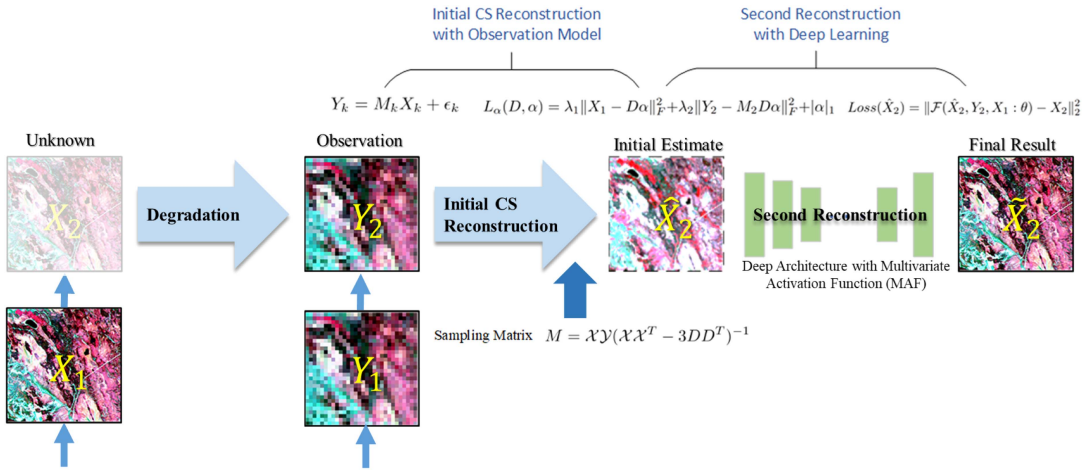

Fig. 5. SDCS reconstruction with two stages. At the first stage, we design a sampling matrix and provide an initial estimation; at the second stage, we construct deep architecture to make the final estimation by compensating for the spectral and spatial features.

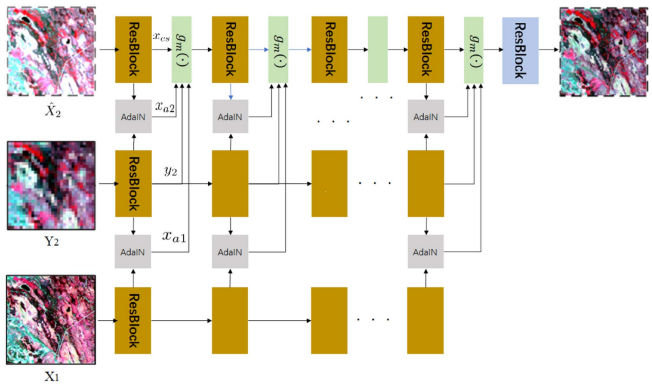

Fig. 6. Second stage reconstruction by PD-CNN architecture. $g_m(\cdot)$ is the MAF. ResBlock is the block of residual networks. AdaIN is the adaptive instance normalization.

where l is a leak parameter that represents the slope, y_2 represent the feature values from Y_2, respectively, x_{a1} is the feature values of X_1 out from AdaIN function, and x_{cs} is the feature values of $\hat{X}_2$ out from AdaIn function. x_{a2} is with the similar definition as x_{a1}. For the proposed PD-CNN, there are two types of activation functions. The conventional ReLU activation function as in (23) is used in the network of ResBlock. The proposed activation function of g_m in (24) is used in the interval between different ResBlocks.

In the proposed activation function $g_m(x_{cs}, x_{a1}, x_{a2}, y_2)$, there are four inputs and four cases, which provide more information for activation. For the first case, when $x_{a2} \geq 0$ and $x_{a1} \geq 0$, it means the features from t_1 and t_2 are both remarkable, and then features directly from CS initial reconstruction and their styled feature all need to be reserved. For the second case, when $x_{a2} \geq 0$ and $x_{a1} < 0$, features directly from CS initial reconstruction should be multiplied by leak parameter. For the third case, when $x_{a2} < 0$ and $x_{a1} \geq 0$, it means that features are changed in time dimension, and the initial estimate x_{cs} is eliminated and only the styled feature is reserved. For the fourth case, when $x_{a2} < 0$ and $x_{a1} < 0$, the features from t_1 and t_2 are both negative so that in addition to multiplying leak parameter we use y_2 to compensate for the activation. Overall, MAF g_m is piecewise function, so that it is more suitable for the task of information fusion in the multibranch networks for postprocessing.

Now, we have designed PD-CNN with MAF for the second stage of deep CS reconstruction ($\tilde{X}_2$). The combination of two stage reconstruction is with both model-driven and data-driven reconstruction. The complete semiblind CS fusion model is shown in Fig. 5.

E. Data Preprocessing and Complete Fusion Model

1) Data Preprocessing: Data preprocessing is a necessary way to enhance the sparsity of data and to find a reasonable local M for semiblind CS. Especially for the first stage, it is unrealistic to use a single sampling matrix M to represent all feature relationships between HTLS images and HSLT images, because Landsat and MODIS data inevitably suffer from system errors, interferences in imaging and abrupt changes in land cover, etc. In this article, to ensure the stability and robust in semiblind CS reconstruction, we use group sparse representation in dictionary learning and group CS in matrix M training.

Similar to [75], for an arbitrary ith HTLS image and HSLT image at time k are segmented into patch set $\mathbb{X}_k^i$ and $\mathbb{Y}_k^i$. To balance feature diversity and sparsity, each patch dataset $\mathbb{X}_k^i$, is clustered into C classes according to the Euclidean distance, as well as $\mathbb{X}_k = \{\mathfrak{x}_k^{(i,c)}\}_{c=1}^C$. For the HTLS image, we have a similar patch set $\mathbb{Y}_k^i = \{\mathfrak{y}_k^{(i,c)}\}_{c=1}^C$. However, each cluster in $\mathbb{Y}_k^i$ is created by allocating each patch to its nearest cluster center corresponding in $\mathbb{X}_k^i$. For each subclass, we search for a corresponding dictionary $D_k^{(i,c)}$. The sampling matrix $M_k^{(i,c)}$ is trained using the cluster pair $\langle \mathfrak{x}_k^{(i,c)}, \mathfrak{y}_k^{(i,c)} \rangle$. More detailed description on patch groups refers to [75]. In this way, the localization dictionary more sparsely represents the data and the localization sampling matrix is adaptive, so that the condition of reconstruction for CS is more easily satisfied.

After preprocessing, for the arbitrary ith pair subclass $\langle \mathfrak{x}_k^{(i,c)}, \mathfrak{y}_k^{(i,c)} \rangle$ in image $\{X_k^i, Y_k^i\}$, the training and prediction are all similar. Therefore, for convenient, we eliminate the index (i, c) and only use $\{X_k, Y_k\}_{k=1}^2$ as a group in derivation formula of performing fusion.

2) Complete Fusion Model: Now we summarize the complete fusion model. As illustrated in Fig. 5, it includes two stages: the CS initial estimation and the postreconstruction by deep learning.

At the first stage, an initial estimation is obtained by CS reconstruction. There are three unknown variables D, α, and M in CS reconstruction. Dictionary D can be found by solving (9) by group sparse representation. Each patch group has an individually trained dictionary D and individual sampling matrix M. Sampling matrix M is a solution of (20). M is only partly known and disturbed by condition of image acquisition. The solution of (20) will make sure that both the RIP condition of CS and the correspondence of spatiotemporal features are satisfied at the same time. Matrix M is also adaptive and correspond to local D. Therefore, the local downsampling matrix within a group is easier to be estimated and the intrinsic mapping relationship of spatiotemporal data is better established. At the end of the first stage, we have a new training dataset as $\{\{X_1^i, X_2^i\}, \hat{X}_2^i, \{Y_1^i, Y_2^i\}\}_{i=1}^N$, where $\hat{X}_2^i$ is the initial estimation by CS.

At the second stage, after the initial fusion $\hat{X}_2^i$ is already generated, we need to use the PD-CNN structure with MAF to reconstruct the image based on the initial estimates $\hat{X}_2^i$. PD-CNN structure has three branch, and it is added as a postprocessing of the fusion after CS reconstruction. The new dataset are organized as Fig. 6 and input into of PD-CNN for deep reconstruction. One of key problem is to decide what features should be activated and what information should be introduced into the final fusion image. The newly designed MAF function can help to utilize the reference information of neighborhood spatiotemporal image. The other key problem is to reduce system errors in this stage, so that we use AdaIN block for the two branches. The second stage not only smoothes out the noise but also reduces the errors from unprecise sampling matrix and compensates for the influences by different imaging conditions.

Now, we summarized the training steps of SDCS as Algorithm 1. The predicting steps is similar to Algorithm 1 but only performs prediction in the second stage.

Algorithm 1: Training for STRSIF with SDCS.

Input: Training dataset $S_{tr} = \{\{X_1^i, X_2^i\}, \{Y_1^i, Y_2^i\}\}_{i=1}^N$

The First Stage:

1: Preprocess training data S_{tr}: generate patches and construct patch group as Section VI-E1

2: Training the dictionary D by (9)

3: Construct the sampling matrix M by (20)

4: Make an initial estimate with (9) and get $\hat{X}_2^i$

The Second Stage:

1: Construct new training data

$$\hat{S}_{tr} = \{\{X_1^i, X_2^i\}, \hat{X}_2^i, \{Y_1^i, Y_2^i\}\}_{i=1}^N$$

2: Use training data $\hat{S}_{tr}$ to training the second-stage network PD-CNN in Fig. 6.

Output: D, M, and PD-CNN

VII. Experiments and Results

A. Study Areas and Data Sets

In the experiment, four open source datasets for the ColeamBbally Irrigation Area (CIA),[1] the Gwydir Downstream Catchment Area (LGC),[2] the Ar Horqin Banner (AHB) and Tianjin [3] were used to test the proposed model.

CIA is an image dataset of Australia's summer crop growing areas. It is 43 km high in the north–south direction and 51 km wide in the east–west direction. The dataset consists of Landsat-7 ETM+ and MODIS image pairs to form the MODTRAN4 product. The image size of Landsat-7 ETM+ is 720 × 2040. For each pair of images, the MODIS image is interpolated to the same size as the Landsat-7 image. In both CIA and LGC, the band number of Landsat-7 ETM+ and MODIS images is 6. There are 17 pairs of cloud-free images in the CIA dataset. In CIA as in Fig. 7, image phenology changes significantly with the change of time.

LGC is about drainage areas in northern New South Wales. It is 80 km high from north to south and 68 km wide from east to west. The image pairs are Landsat-5 TM image and MODIS MOD09GA product image. Landsat-5 TM satellite image size is 3200 × 2720. Again, MODIS data are upsampled to the same size as Landsat images. There are 14 pairs of images in the LGC dataset. The scenarios provided by LGC are mainly watershed areas. With the change of time, the land cover changes significantly.

AHB, located in Ar Horqin Banner of Inner Mongolia province, China, is a dataset that spans 51 km in the north–south direction and 48 km in the east–west direction. The dataset comprises Landsat-8 OLI and MODIS image pairs to form the AHB dataset. The image size of Landsat-8 OLI is 2480 × 2800 with six bands. For each pair, the MODIS image is resampled to match the spatial resolution of the Landsat-8 image. The AHB dataset features 27 pairs of cloud-free images, capturing significant phenological changes in rural areas over time.

Tianjin is in the northern municipality of China, which covers an area of 68 km in the north–south direction and 66 km in the east–west direction. This dataset is composed of Landsat-8 OLI and MODIS MOD02HKM image pairs, known as the Tianjin dataset. The Landsat-8 OLI images measure 2100 × 1970 with six spectral bands. Similar to the AHB dataset, the MODIS images are interpolated to the pixel dimensions of the Landsat-8 images. The Tianjin dataset includes 27 pairs of cloud-free images, illustrating substantial phenological changes in urban areas throughout the different seasons.

[1] https://dx.doi.org/10.4225/08/5111AC0BF1229CIA
[2] https://dx.doi.org/10.4225/08/5111AD2B7FEE6LGC
[3] https://doi.org/10.1007/s11432-019-2785-y

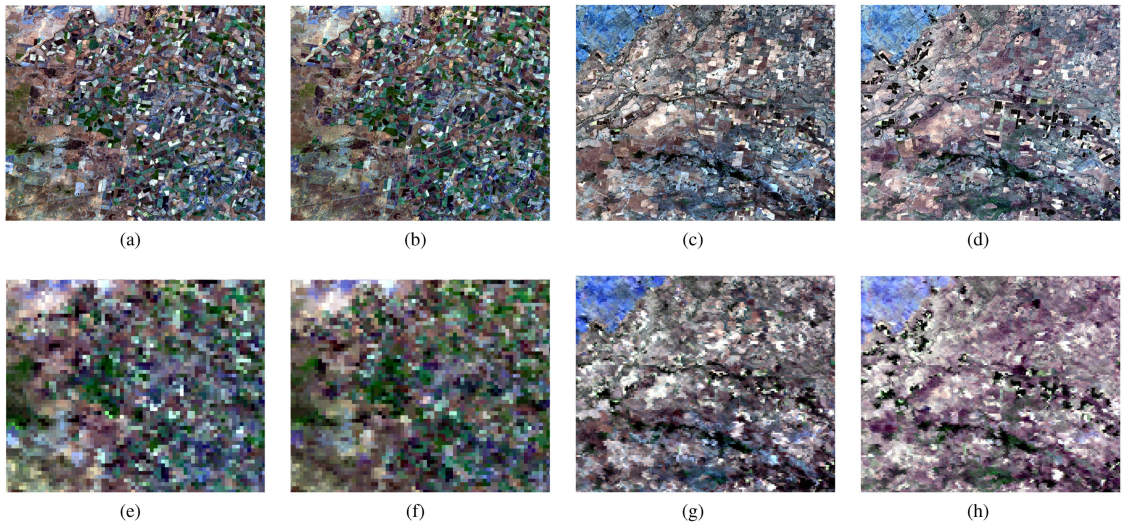

Fig. 7. Examples from CIA (a), (b),(e),(f) and LGC (c), (d), (g), (h) dataset, from which we can observe that there are significant phenological changes. (a) CIA, $t_1 = 2002108$. (b) CIA, $t_2 = 2002117$. (c) LGC, $t_1 = 20050302$. (d) LGC, $t_2 = 20050403$. (e) CIA, $t_1 = 2002108$. (f) CIA, $t_2 = 2002117$. (g) LGC, $t_1 = 20050302$. (h) LGC, $t_2 = 20050403$.

To illustrate the effectiveness of the experiment, experiment setting, quantitative comparison, qualitative comparison, and ablation are introduced in following sections.

B. Experiment Setting

To validate the proposed STRSIF method of semiblind deep compressive sensing (SDCS), we compared SDCS with seven state-of-the-art algorithms (MLFF-GAN [7], FSDAF [23], EDCSTFN [42], GAN-STFM [46], ECPW-STFN [50], SRSF-GAN [76], and STFDiff [77]) to prove its effectiveness. Both quantitative and qualitative comparisons are carried out comprehensively. We also show the function of different parts in SDCS by ablating different parts of its two stages of deep CS.

In the experimental setup, all datasets are partitioned into training and testing subsets. In the CIA dataset, the training subset consists of 11 groups, each containing the first ten pairs of images in chronological order, while the testing subset includes the next five groups in time sequence. Each group contains two pairs of images that are adjacent in time. Similarly, in the LGC dataset, the training subset consists of nine groups, while the testing subset includes the remaining four groups. Similarly, the first 20 groups of pairs of images from the AHB and Tianjin datasets will be employed for training, while five groups will be selected for testing. In the training phase, image pairs from the later period are allowed to be used as known images, while low-resolution images from the previous period are allowed to be used as MODIS images of input at prediction time.

To evaluate the fusion results, we compare the results to the ground truth. Several evaluation indicators are used to evaluate the experimental data. They are the root mean square error (RMSE), the mean absolute loss (MAE), the structural similarity (SSIM), and the spectral angle mapper (SAM). Among them, RMSE, and SSIM mainly depends on spatial details, but SAM mainly reflect on spectral loss.

C. Quantitative Comparison

For quantitative comparison, eight algorithms SDCS, STFDiff [77], ECPW-STFN [50], SRSF-GAN [76], MLFF-GAN [7], GAN-STFM [46], EDCSTFN [42] and FSDAF [23] participated in the experiments. STFDiff [77], ECPW-STFN [50], SRSF-GAN [76], EDCSTFN [42], GAN-STFM [46], and MLFF-GAN [7] are deep-learning based methods. EDCSTFN [42] solely utilizes the architecture of a convolutional neural network. GAN-STFM [46] and MLFF-GAN [7] mainly adopt the architecture of generative adversarial networks. For FSDAF, EDCSTFN, STFDiff [77], ECPW-STFN [50], SRSF-GAN [76], and MLFF-GAN [7], the input data include a low-resolution image at the prediction time and a pair of recent low- and high-resolution images. In contrast, for GAN-STFM [46], the input data include a low-resolution image at the prediction time and a recent high-resolution image. FSDAF [23] is a popular nondeep learning method based on weight and unmixing, which exhibits good performances. In addition, we noticed that the GAN-STFM [46] algorithm only requires two input images.

Table II shows the quantitative index evaluation of the eight algorithms on CIA, LGC, AHB, and Tianjin datasets, respectively.

For CIA dataset, in most cases, the proposed SDCS shows better performances. However, deep learning algorithm MLFF-GAN [7] also show some good performances, and in some cases its quantitative index overpasses some of other methods. This may be attributed to its multilevel feature fusion strategy. For LGC dataset, in most cases, the proposed SDCS shows better performances than others. Surprisingly, conventional algorithm FSDAF [23] performs well on the SAM metric, achieving the highest mean rank, which indicates that FSDAF [23] effectively preserves spectral information. SDCS ranks second in terms of mean SAM, closely trailing FSDAF [23], further demonstrating the robustness of the proposed method in mitigating spectral distortion. For FSDAF [23], we speculate that the pixels in the LGC

TABLE II
RESULTS OF CONTRAST EXPERIMENTS ON CIA, LGC, AHB AND TIANJIN DATASETS

	Model	04-01				04-10				04-17				04-26				Mean			
		RMSE	SAM	SSIM	MAE	RMSE	SAM	SSIM	MAE	RMSE	SAM	SSIM	MAE	RMSE	SAM	SSIM	MAE	RMSE	SAM	SSIM	MAE
CIA	SDCS	**0.0186**	0.0719	**0.9093**	**0.0162**	**0.0163**	**0.0681**	**0.9120**	**0.0139**	**0.0166**	**0.0704**	**0.9024**	**0.0141**	**0.0165**	0.0632	**0.9015**	**0.0139**	**0.0170**	**0.0684**	**0.9063**	**0.0145**
	MLFF-GAN	0.0201	**0.0702**	0.9018	0.0176	0.0196	0.0720	0.8879	0.0169	0.0213	0.0796	0.8476	0.0181	0.0210	0.0694	0.8562	0.0178	0.0205	0.0728	0.8734	0.0176
	STFDiff	0.0218	0.0788	0.8969	0.0192	0.0229	0.0766	0.8873	0.0201	0.0216	0.0813	0.8485	0.0182	0.0205	0.0702	0.8593	0.0172	0.0217	0.0767	0.8730	0.0187
	SRSF-GAN	0.0213	0.0781	0.9049	0.0188	0.0201	0.0759	0.8944	0.0174	0.0218	0.0805	0.8483	0.0186	0.0213	**0.0631**	0.8647	0.0181	0.0211	0.0744	0.8781	0.0182
	ECPW-STFN	0.0318	0.1569	0.7474	0.0293	0.0375	0.1505	0.7286	0.0351	0.0329	0.1412	0.7007	0.0298	0.0281	0.1381	0.7133	0.0250	0.0326	0.1467	0.7225	0.0298
	EDCSTFN	0.0244	0.0906	0.8917	0.0218	0.0274	0.0946	0.8729	0.0248	0.0290	0.1063	0.8280	0.0259	0.0234	0.0764	0.8519	0.0202	0.0261	0.0920	0.8611	0.0232
	GAN-STFM	0.0254	0.0988	0.8758	0.0225	0.0252	0.0940	0.8673	0.0221	0.0265	0.1014	0.8290	0.0231	0.0247	0.0909	0.8361	0.0211	0.0254	0.0963	0.8520	0.0222
	FSDAF	0.0225	0.0811	0.8859	0.0196	0.0241	0.0743	0.8634	0.0209	0.0226	0.0869	0.8282	0.0187	0.0216	0.0715	0.8382	0.0179	0.0227	0.0784	0.8539	0.0193
	Model	01-29				02-24				03-02				04-03				Mean			
		RMSE	SAM	SSIM	MAE	RMSE	SAM	SSIM	MAE	RMSE	SAM	SSIM	MAE	RMSE	SAM	SSIM	MAE	RMSE	SAM	SSIM	MAE
LGC	SDCS	**0.0190**	0.0765	**0.9114**	**0.0166**	**0.0159**	**0.0609**	**0.9368**	**0.0139**	**0.0142**	0.0556	**0.9499**	0.0125	**0.0154**	0.0548	**0.9431**	**0.0136**	**0.0161**	0.0619	**0.9353**	**0.0142**
	MLFF-GAN	0.0198	0.0762	0.8993	0.0172	0.0161	0.0636	0.9280	0.0140	0.0147	0.0580	0.9421	0.0129	0.0162	0.0548	0.9380	0.0143	0.0167	0.0632	0.9268	0.0146
	STFDiff	0.0235	0.0858	0.8939	0.0212	0.0160	0.0629	0.9332	0.0140	0.0161	0.0587	0.9454	0.0144	0.0173	**0.0542**	0.9382	0.0155	0.0182	0.0654	0.9277	0.0163
	SRSF-GAN	0.0217	0.0822	0.9058	0.0193	0.0178	0.0735	0.9268	0.0158	0.0179	0.0628	0.9390	0.0162	0.0167	0.0564	0.9402	0.0150	0.0185	0.0687	0.9280	0.0165
	ECPW-STFN	0.0411	0.1889	0.7346	0.0390	0.0411	0.1992	0.7498	0.0395	0.0434	0.1925	0.7609	0.0421	0.0437	0.2046	0.7539	0.0422	0.0423	0.1963	0.7498	0.0407
	EDCSTFN	0.0313	0.1102	0.8818	0.0289	0.0306	0.1108	0.8985	0.0287	0.0317	0.0972	0.9082	0.0301	0.0316	0.1095	0.9008	0.0297	0.0313	0.1070	0.8973	0.0294
	GAN-STFM	0.0299	0.1206	0.8662	0.0272	0.0308	0.1231	0.8783	0.0286	0.0263	0.0983	0.9073	0.0242	0.0253	0.0975	0.9051	0.0232	0.0281	0.1099	0.8892	0.0258
	FSDAF	0.0215	**0.0740**	0.8885	0.0186	0.0184	0.0614	0.9120	0.0159	0.0144	**0.0498**	0.9367	**0.0123**	0.0167	0.0579	0.9257	0.0145	0.0177	**0.0608**	0.9157	0.0153
	Model	05-12				10-03				10-19				12-06				Mean			
		RMSE	SAM	SSIM	MAE	RMSE	SAM	SSIM	MAE	RMSE	SAM	SSIM	MAE	RMSE	SAM	SSIM	MAE	RMSE	SAM	SSIM	MAE
AHB	SDCS	**0.0264**	0.0945	**0.8740**	**0.0240**	**0.0293**	0.1129	**0.8510**	**0.0266**	**0.0178**	0.0884	0.9048	**0.0157**	0.0325	0.1955	0.9071	0.0317	**0.0265**	0.1228	**0.8842**	**0.0245**
	MLFF-GAN	0.0381	0.1322	0.8551	0.0359	0.0592	**0.0923**	0.8027	0.0569	0.0206	**0.0674**	**0.9162**	0.0190	0.0313	0.2212	0.8918	0.0305	0.0373	0.1283	0.8664	0.0356
	STFDiff	0.0446	0.1073	0.8605	0.0425	0.0491	0.1332	0.8091	0.0466	0.0339	0.1331	0.8767	0.0323	**0.0211**	**0.1941**	**0.9137**	**0.0204**	0.0372	0.1419	0.8650	0.0355
	SRSF-GAN	0.0419	**0.0745**	0.8557	0.0398	0.0485	0.1016	0.8228	0.0462	0.0215	0.0889	0.9157	0.0198	0.0377	0.2158	0.8942	0.0372	0.0374	**0.1202**	0.8721	0.0358
	ECPW-STFN	0.0479	0.1719	0.8540	0.0461	0.0402	0.2444	0.7673	0.0378	0.0299	0.0879	0.8789	0.0285	0.0520	0.2257	0.8716	0.0516	0.0425	0.1825	0.8430	0.0410
	EDCSTFN	0.0461	0.1336	0.8506	0.0440	0.0355	0.2212	0.7950	0.0328	0.0205	0.0929	0.9026	0.0185	0.0509	0.2134	0.8659	0.0504	0.0382	0.1653	0.8535	0.0364
	GAN-STFM	0.0435	0.1544	0.8394	0.0413	0.0481	0.1190	0.8136	0.0457	0.0210	0.1155	0.8938	0.0191	0.0366	0.1988	0.8931	0.0358	0.0373	0.1469	0.8600	0.0354
	FSDAF	0.0575	0.0982	0.8129	0.0558	0.0510	0.2546	0.7456	0.0483	0.0550	0.2115	0.7613	0.0537	0.0546	0.2207	0.8387	0.0537	0.0545	0.1963	0.7896	0.0529
	Model	18-10-1				18-12-4				19-1-21				19-5-29				Mean			
		RMSE	SAM	SSIM	MAE	RMSE	SAM	SSIM	MAE	RMSE	SAM	SSIM	MAE	RMSE	SAM	SSIM	MAE	RMSE	SAM	SSIM	MAE
Tianjin	SDCS	0.0363	**0.1656**	**0.7259**	**0.0309**	**0.0333**	**0.1003**	**0.7870**	**0.0281**	0.0367	0.1029	0.8041	0.0311	**0.0416**	**0.1367**	**0.7077**	**0.0346**	**0.0370**	**0.1264**	**0.7562**	**0.0312**
	MLFF-GAN	0.0417	0.1973	0.6763	0.0363	0.0393	0.1284	0.7413	0.0341	**0.0340**	**0.0963**	0.8010	**0.0283**	0.0423	**0.1367**	0.6858	0.0351	0.0393	0.1397	0.7261	0.0334
	STFDiff	0.0372	0.1832	0.7230	0.0313	0.0361	0.1248	0.7682	0.0304	**0.0340**	**0.0963**	**0.8279**	0.0284	0.0471	0.1615	0.6800	0.0395	0.0386	0.1414	0.7498	0.0324
	SRSF-GAN	**0.0362**	0.1735	0.7255	**0.0309**	0.0339	0.1128	0.7802	0.0286	0.0361	0.1063	0.8008	0.0304	0.0475	0.1688	0.6790	0.0403	0.0384	0.1404	0.7464	0.0325
	ECPW-STFN	0.0694	0.3293	0.6628	0.0641	0.0424	0.1713	0.7519	0.0368	0.0480	0.1777	0.7935	0.0427	0.0814	0.3570	0.6304	0.0740	0.0603	0.2588	0.7097	0.0544
	EDCSTFN	0.0719	0.2875	0.6306	0.0665	0.0426	0.1592	0.7329	0.0365	0.0489	0.1653	0.7530	0.0425	0.0903	0.3710	0.5897	0.0827	0.0634	0.2458	0.6766	0.0570
	GAN-STFM	0.0578	0.2511	0.6102	0.0512	0.0469	0.1847	0.6883	0.0406	0.0462	0.1485	0.7123	0.0389	0.0671	0.2567	0.5818	0.0586	0.0545	0.2102	0.6481	0.0473
	FSDAF	0.0393	0.1956	0.6871	0.0332	0.0381	0.1293	0.7423	0.0324	0.0358	0.1222	0.7957	0.0301	0.0641	0.2707	0.5087	0.0563	0.0443	0.1795	0.6835	0.0380

Bold indicates the best, underlining indicates the second-best.

TABLE III
ABLATION EXPERIMENT ON CIA DATASET

Date	Model				MAE	RMSE	SAM	SSIM
	CS	CNN	Ada	MAF				
04-01	√	×	×	×	0.0950	0.0969	0.7031	0.7093
	√	√	×	×	0.0163	0.0188	0.0778	0.9004
	√	√	√	×	0.0171	0.0195	0.0807	0.8984
	√	√	√	√	**0.0162**	**0.0186**	**0.0719**	**0.9093**
04-10	√	×	×	×	0.0907	0.0927	0.7319	0.6982
	√	√	×	×	0.0151	0.0177	0.0775	0.8979
	√	√	√	×	0.0155	0.0180	0.0788	0.8991
	√	√	√	√	**0.0139**	**0.0163**	**0.0681**	**0.9120**
04-17	√	×	×	×	0.0932	0.0951	0.7892	0.6764
	√	√	×	×	0.0153	0.0180	0.0772	0.8891
	√	√	√	×	0.0159	0.0185	0.0787	0.8906
	√	√	√	√	**0.0141**	**0.0166**	**0.0704**	**0.9024**
04-26	√	×	×	×	0.0972	0.0991	0.7670	0.6835
	√	√	×	×	0.0170	0.0197	0.0709	0.8885
	√	√	√	×	0.0167	0.0193	0.0710	0.8911
	√	√	√	√	**0.0139**	**0.0165**	**0.0632**	**0.9015**
Mean	√	×	×	×	0.0940	0.0959	0.7478	0.6919
	√	√	×	×	0.0159	0.0185	0.0759	0.8940
	√	√	√	×	0.0163	0.0188	0.0773	0.8948
	√	√	√	√	**0.0145**	**0.0170**	**0.0684**	**0.9063**

Bold indicates the best.

dataset are easy to be classified into their right categories, so that FSDAF [23] method can find their appropriate endmember in spectral unmixing and then show more precise fusion results.

For AHB and Tianjin datasets, the SDCS model demonstrates superior performance across most metrics when compared to other models. SRSF-GAN [76], MLFF-GAN, STFDiff [77] also achieve good performances. On the AHB dataset, the SAM metric is slightly inferior to that of SRSF-GAN [76]. Both SRSF-GAN [76] and MLFF-GAN [7] achieve relatively good results, indicating that an appropriate generator, supplemented by adversarial loss, holds potential for spatiotemporal image fusion. However, the training of GANs is often plagued by instability and mode collapse issues. STFDiff [77] incorporates the noise injection and denoising strategy from diffusion models for spatiotemporal fusion, yet its sampling steps are time-consuming and need a substantial amount of data to train an effective denoising model. In contrast to these models, the SDCS method proposed in this article employs a hybrid architecture that integrates both model-driven and data-driven approaches, eliminating the need for adversarial structures, and achieves better results than GAN-based and diffusion-based methods. This further underscores the effectiveness of the proposed approach. Noticeably, ECPW-STFN [50] does not performs as well as other deep learning methods in all datasets in most cases. We speculate that this may be due to the wavelet transform not complementing the network architecture mentioned in their article effectively. In contrast, the CS method combined with PD-CNN proposed in this article, yields very promising results. For deep learning methods, a network architecture needs to be more suitable for spatiotemporal fusion to have better performance.

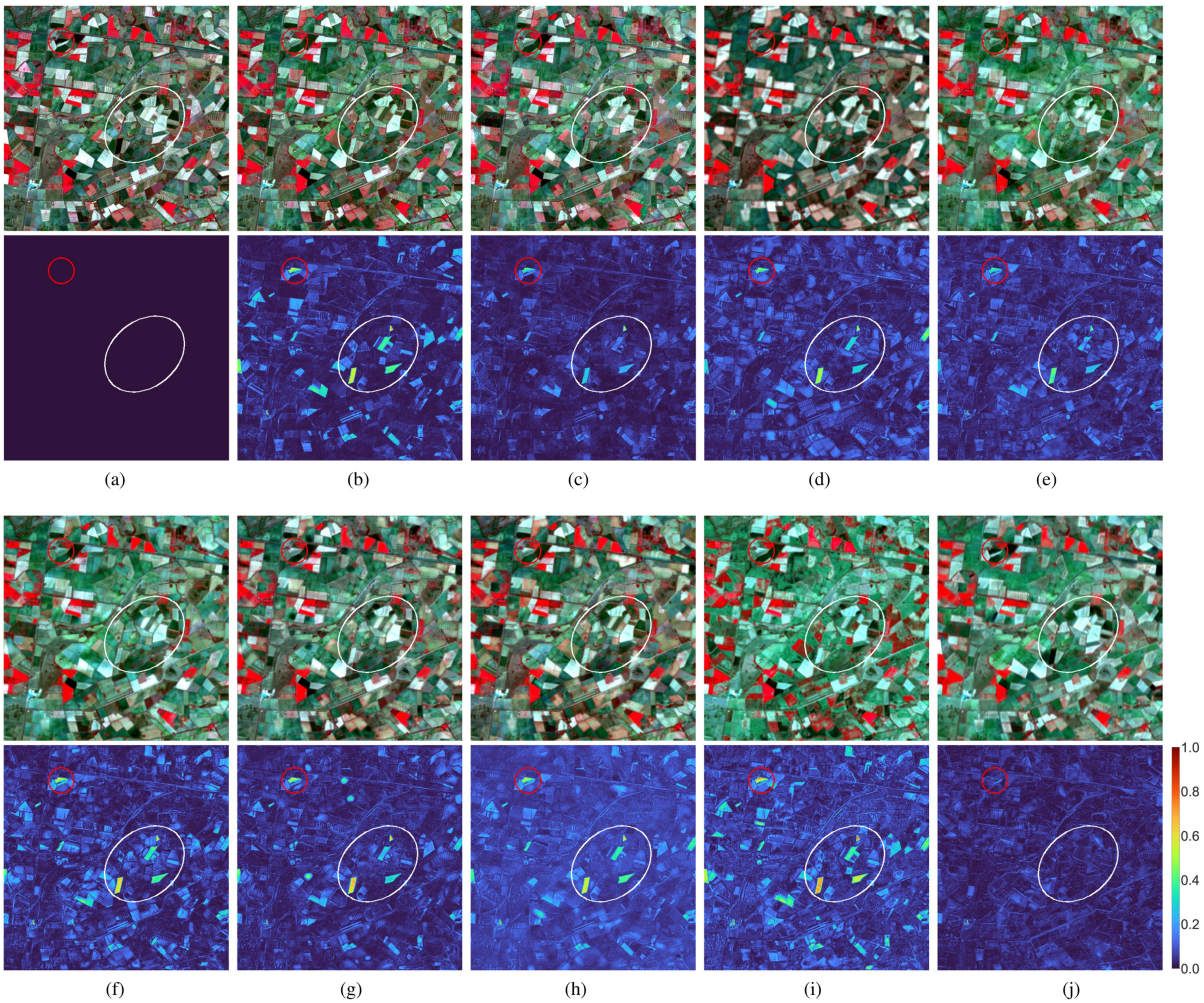

Fig. 8. Visual results of 8 algorithms on CIA dataset. In each subgraph, the lower graph represents the Mean Absolute Error (MAE) map. (a) Real. (b) HR t_1. (c) FSDAF. (d) EDCSTFN. (e) MLFF-GAN. (f) SRSF-GAN. (g) STFDiff. (h) ECPW-STFN. (i) GAN-STFM. (j) SDCS.

Spatiotemporal fusion as an ill-posed problem is difficult to achieve good performance only using model-driven schemes and handcraft features. For SDCS method, the CS reconstruction provide a good initial estimation for fusion, at the same time the second stage with plenty of parameters can well reduce the prediction error of the CS model. Relatively speaking, the overall performances of the bidirectional-driven method SDCS are better than the both deep learning and nondeep learning methods.

Table VI presents a comparison of the parameter count, computational load (multiply–accumulate operations, MACs), and inference time (seconds per image) across various deep learning methods. MACs and inference time require the network to process a 256 × 256 image 6 spectral bands. ECPW-STFN [50] and EDCSTFN [42] have relatively fewer parameters, which may be a direct cause of their suboptimal results. In comparison to other high-performing networks such as MLFF-GAN [7] and SRSF-GAN [76], SDCS has fewer parameters and a reduced computational load. Our proposed method achieves relatively superior performance with a lower parameter count and minimal computational requirements. Regarding inference time, the first phase of SDCS does not utilize GPU acceleration, resulting in a significant time expenditure. Except for the STFDiff [77] model, the inference time of all deep learning models is not much different. The sampling process of STFDiff [77] inherently leads to a notable lag in inference speed compared to the other models.

D. Qualitative Comparison

Fig. 8 (CIA dataset), Fig. 9 (LGC dataset), and Fig. 10 (Tianjin dataset) are visual comparisons for different methods. Aided by CS initial estimation and the design of MAF, the proposed SDCS in most cases reconstructs fusion images by less errors. Top Figs. 8(a), 9(a), and 10(a) are real images at the predicting time. Top Figs. 8(b), 9(b), and 10(b) are images at the reference time. Top Figs. 8(c)–(j), 9(c)–(d), and 10(c)–(j) show the local images of results by false color (band 1, 2, and 3) from different algorithms. At bottom of each subfigure shows the difference map between real image at the reference time and real image at the target time.

For the abruptly changed regions, most methods show obvious errors, although all of them can reconstruct a large part of features by fusion. In the white circle areas, SDCS also predicts some wrong changes, but the wrong prediction is

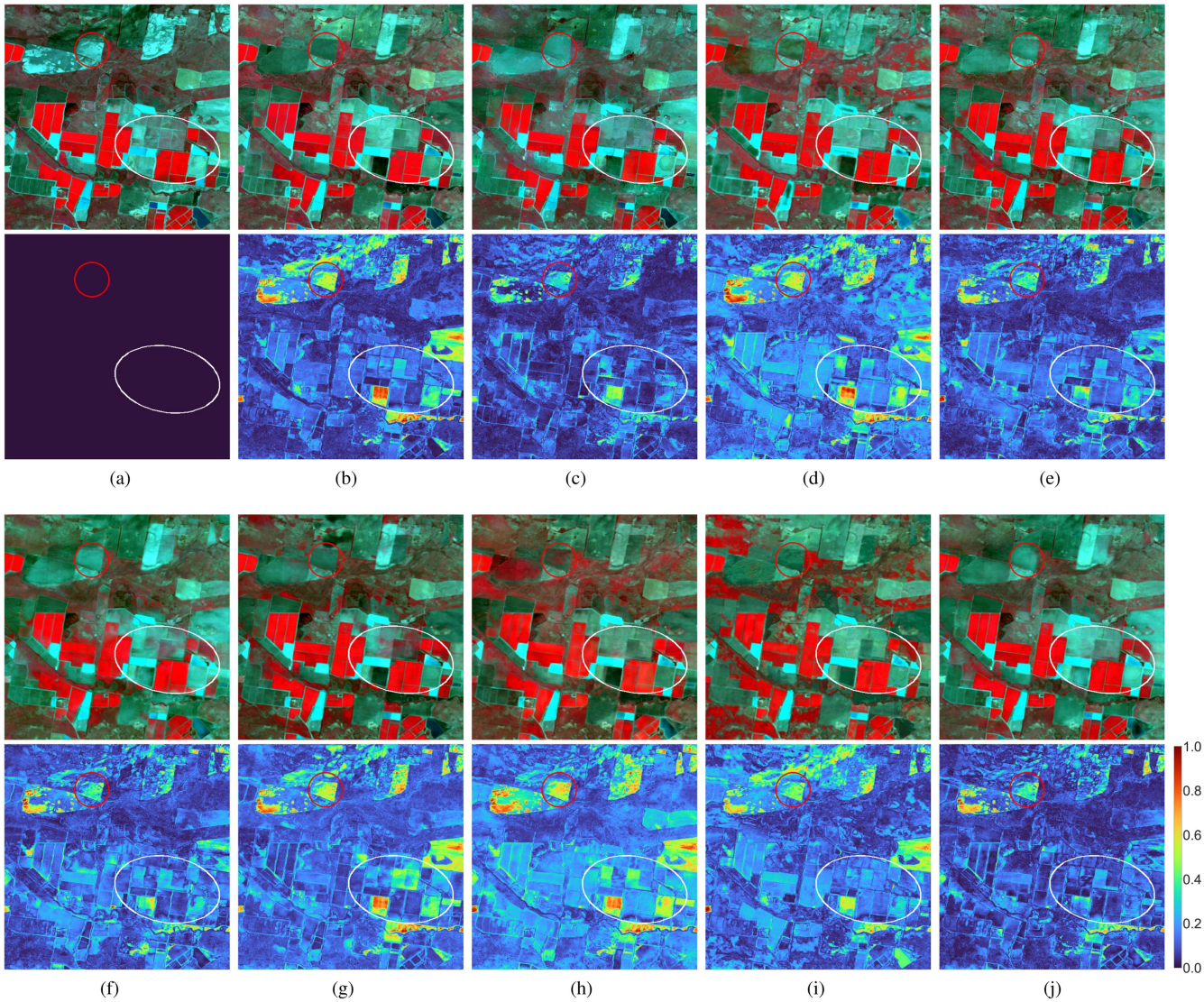

Fig. 9. Visual results of 8 algorithms on LGC dataset. In each subgraph, the lower graph represents the mean absolute error (MAE) map. (a) Real. (b) HR t_1. (c) FSDAF. (d) EDCSTFN. (e) MLFF-GAN. (f) SRSF-GAN. (g) STFDiff. (h) ECPW-STFN. (i) GAN-STFM. (j) SDCS.

greatly weakened due to the constraints of two stage reconstruction. After CS reconstruction providing initial prediction, the MAF can further determine the intensity when injecting the features of the high-resolution image at reference time or the features of the low-resolution image at the prediction time. For GAN-STFM [46], because only two images are taken as input, it is easy to retain the error caused by different sensors so that as in Fig. 8(i) the white circle areas with abrupt differences from the previous image will have greater prediction errors. On the LGC, AHB and Tianjin dataset, the white circle areas also show the similar performances as the CIA dataset. In most cases, SDCS shows fewer errors where there is abrupt changes between the prediction time and the reference time.

For the smoothly or slightly changed regions, most of the methods show good fusion results. In the red circle areas, for example, in bottom Fig. 8, the red circle areas in the images are close to dark blue, so it means that the change is very small. In red circle areas, FSDAF [23] algorithm shows very less errors in the bottom Fig. 9(c). But for other areas, the FSDAF [23] classification process for end-members does not adaptively change with dynamic features, so the deviation in these areas will be very obvious and this reduces its overall performances. We can observe that, there are also some small errors in these areas by the deep learning methods in bottom 8(c)–(j), 9(c)–(j), and 10(c)–(j), because deep learning method may rely too much on the learned experiences. The prediction of MLFF-GAN [7] and SRSF-GAN [76] in Fig. 9(a) and (f) errors are not so obvious. SDCS in Figs. 8(j), 9(j), and 10(j) performs much better when compared with other deep learning methods in slightly changed regions.

For the some transition areas which are not so abruptly changed but more serious than a slight changed, the performances of different methods are quite different. For example, in the areas between the red circle and white circle, SDCS often shows fewer prediction errors, because the fusion for local and global features benefits from the design of two stage SDCS. In Fig. 8, FSDAF [23] in Fig. 8(c) is close to SDCS in Fig. 8(j).

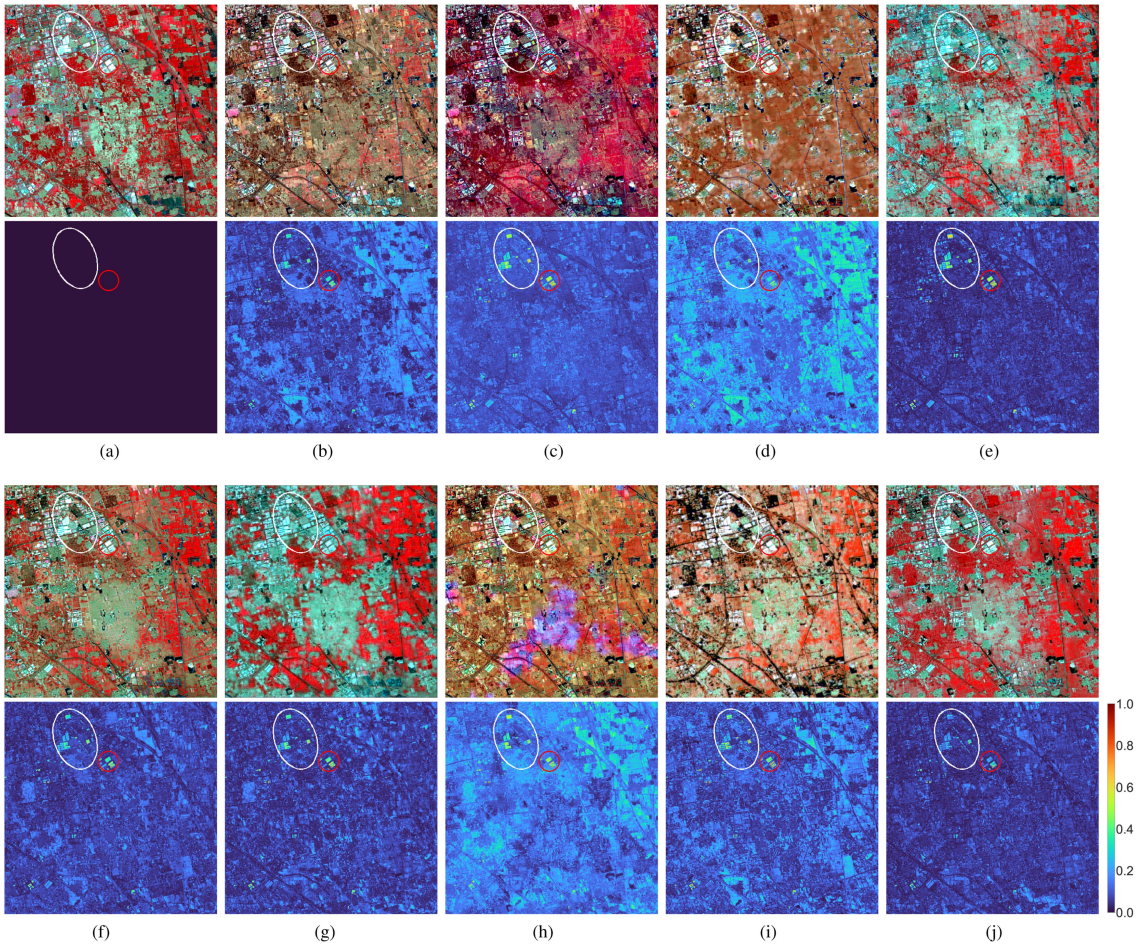

Fig. 10. Visual results of 8 algorithms on Tianjin dataset. In each subgraph, the lower graph represents the MAE map. (a) Real. (b) HR t_1. (c) FSDAF. (d) EDCSTFN. (e) MLFF-GAN. (f) SRSF-GAN. (g) STFDiff. (h) ECPW-STFN. (i) GAN-STFM. (j) SDCS.

In Figs. 9 and 10, SRSF-GAN [76] and MLFF-GAN [7] are close to SDCS. We believe that the CS initial estimation and MAF postprocessing all take into positive effects in the fusion of these areas.

Overall, SDCS performs better in visual on all datasets when compared with other methods.

E. Ablation Experiments

Ablation experiments encompass two aspects: First, the ablation of network components, including the effectiveness of AdaIn and MAF. Second, the ablation of methods of the initial phase, i.e., comparing the initial results generated by other methods with the CS-based approach proposed in this article.

1) Ablation of Network Components: There were four ablation experiments: 1) one stage only with CS reconstruction Fig. 11(g); 2) two stage by CS+CNN but without AdaIn and MAF [Fig. 11(h)]; 3) two stage by CS+CNN+AdaIn but with out MAF [Fig. 11(i)]; 3) complete model in which CS, CNN, AdaIN, and MAF are all added [Fig. 11(j)]. The results for the four stages are listed in Tables III and IV. In the CS stage, there will be a relatively significant error, as this stage focuses more on the reconstruction of spatial information, while the spectral

TABLE IV
ABLATION EXPERIMENT ON LGC DATASET

Date	Model				MAE	RMSE	SAM	SSIM
	CS	CNN	Ada	MAF				
01-29	√	×	×	×	0.0818	0.0840	0.6520	0.7152
	√	√	×	×	0.0275	0.0299	0.1280	0.8795
	√	√	√	×	0.0292	0.0317	0.1398	0.8641
	√	√	√	√	**0.0166**	**0.0190**	**0.0765**	**0.9114**
02-24	√	×	×	×	0.0835	0.0854	0.6623	0.7299
	√	√	×	×	0.0243	0.0263	0.1344	0.9047
	√	√	√	×	0.0245	0.0267	0.1288	0.8900
	√	√	√	√	**0.0139**	**0.0159**	**0.0609**	**0.9368**
03-02	√	×	×	×	0.0827	0.0844	0.6626	0.7496
	√	√	×	×	0.0219	0.0237	0.0980	0.9228
	√	√	√	×	0.0239	0.0258	0.1159	0.9049
	√	√	√	√	**0.0125**	**0.0142**	**0.0556**	**0.9499**
04-03	√	×	×	×	0.0913	0.0929	0.6870	0.7362
	√	√	×	×	0.0214	0.0234	0.0970	0.9137
	√	√	√	×	0.0250	0.0272	0.1105	0.8944
	√	√	√	√	**0.0136**	**0.0154**	**0.0548**	**0.9431**
Mean	√	×	×	×	0.0848	0.0866	0.6660	0.7327
	√	√	×	×	0.0238	0.0258	0.1143	0.9052
	√	√	√	×	0.0256	0.0279	0.1238	0.8884
	√	√	√	√	**0.0142**	**0.0161**	**0.0619**	**0.9353**

Bold indicates the best.

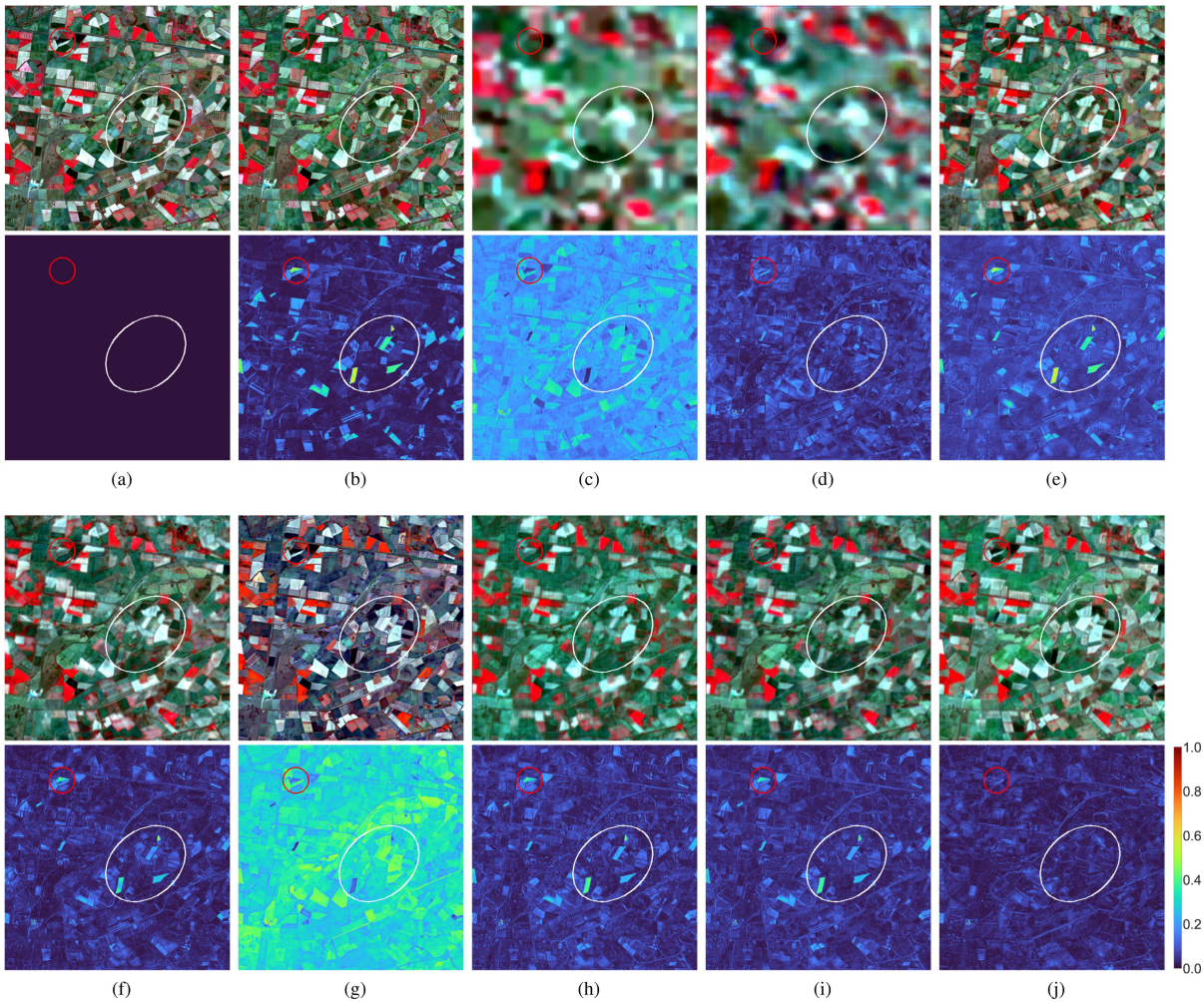

Fig. 11. Results of ablation experiments on CIA dataset. (a) Real image. (b) HR t_1 (c) Stage1-LapSRN. (d) Stage1-LapSRN+Stage2. (e) Stage1-ECPW-STFN (f) Stage1-ECPW-STFN+Stage2 (g) Stage1-CS. (h) CS+CNN but without AdaIN and MAF. (i) CS+CNN+AdIN but without MAF. (j) Our complete model.

transfer is delegated to AdaIn and MAF. Optimal results are achieved when both AdaIn and MAF are fully activated.

Fig. 11(g)–(j) are the visual comparison of four stage ablation experiments for the CIA dataset. we can clearly observe that when CS, CNN, AdaIN, and MAF are added into the model in turns and the errors will be reduced. The visual comparison of the ablation experiment is basically consistent with the quantitative analysis of the previous ablation experiment in Table III.

2) Ablation of the CS Stage: Table V and Fig. 11(c)–(j) illustrates the impact of different first-stage methods on the final outcomes. Among them, LapSRN [78] is a commonly used super-resolution method embedded in OpenCV [79], and ECPW-STFN [50] is one of the deep learning approaches. The experimental results indicate that, although the error in the CS first stage is relatively large [see Fig. 11(g)], by integrating the proposed second-stage method, more superior results [see Fig. 11(j)] can be achieved compared to both LapSRN [78] and ECPW-STFN [50]. LapSRN [78] is solely a super-resolution method, and due to its poor initial spatial resolution [Fig. 11(c)], the final outcomes [see Fig. 11(d)] are also not optimal. Combined with Table II, it can be observed that the proposed second-stage network can significantly enhance its performance when combined with the input from ECPW-STFN [50], which fully demonstrated the robustness of our second-stage method.

TABLE V
ABLATION EXPERIMENTS OF DIFFERENT STAGE I METHODS ON CIA AND LGC DATASETS

Stage 1	CIA dataset					LGC dataset				
	Date	MAE	RMSE	SAM	SSIM	Date	MAE	RMSE	SAM	SSIM
LapSRN		0.0285	0.0318	0.1268	0.8024		0.0248	0.0281	0.1133	0.8205
ECPW-STFN	04-01	0.0193	0.0220	0.0812	0.8830	01-29	0.0194	0.0218	0.0875	0.8998
CS (ours)		**0.0162**	**0.0186**	**0.0719**	**0.9093**		**0.0166**	**0.0190**	**0.0765**	**0.9114**
LapSRN		0.0276	0.0311	0.1258	0.8001		0.0210	0.0239	0.1015	0.8482
ECPW-STFN	04-10	0.0189	0.0217	0.0813	0.8767	02-24	0.0156	0.0177	0.0779	0.9251
CS (ours)		**0.0139**	**0.0163**	**0.0681**	**0.9120**		**0.0139**	**0.0159**	**0.0609**	**0.9368**
LapSRN		0.0265	0.0299	0.1276	0.7900		0.0202	0.0230	0.0924	0.8584
ECPW-STFN	04-17	0.0193	0.0225	0.0900	0.8435	03-02	0.0148	0.0166	0.0675	0.9382
CS (ours)		**0.0141**	**0.0166**	**0.0704**	**0.9024**		**0.0125**	**0.0142**	**0.0556**	**0.9499**
LapSRN		0.0266	0.0302	0.1168	0.7872		0.0193	0.0221	0.0821	0.8581
ECPW-STFN	04-26	0.0185	0.0217	0.0742	0.8497	04-03	0.0156	0.0175	0.0607	0.9343
CS (ours)		**0.0139**	**0.0165**	**0.0632**	**0.9015**		**0.0136**	**0.0154**	**0.0548**	**0.9431**
LapSRN		0.0273	0.0308	0.1243	0.7949		0.0213	0.0243	0.0973	0.8463
ECPW-STFN	Mean	0.0190	0.0220	0.0817	0.8632	Mean	0.0163	0.0184	0.0734	0.9244
CS (ours)		**0.0145**	**0.0170**	**0.0684**	**0.9063**		**0.0142**	**0.0161**	**0.0619**	**0.9353**

The initial estimates will be used as inputs of Stage II. Bold indicates the best.

TABLE VI
PARAMETERS, MACS, AND INFERENCE TIME OF SEVEN DEEP LEARNING METHODS

Model	Parameters	MACs	Time (s/img)
MLFF-GAN	8.7×10^6	1.8×10^{10}	0.0800
STFDiff	7.5×10^6	4.3×10^{10}	2.5208
SRSF-GAN	4.8×10^6	1.6×10^{11}	0.1091
ECPW-STFN	6.9×10^5	3.1×10^{10}	0.0435
EDCSTFN	2.8×10^5	1.8×10^{10}	0.0568
GAN-STFM	4.18×10^6	3.8×10^{10}	0.0627
SDCS (ours)	4.17×10^6	1.2×10^{10}	Stage 1: 138.68 Stage 2: 0.0952

VIII. CONCLUSION

In this article, we proposed a new SDCS method for spatiotemporal fusion. To deal with the large resolution gap, unprecise sampling matrix, and complex imaging condition, etc., the SDCS method is bidirectional and consists of two stages: the CS observation stage and deep reconstruction stage. The advantages from both CS observation model and deep feature learning can be well utilized. The contributions of this work mainly are: in the CS observation stage, we design a sensing matrix satisfying both sampling mapping and RIP condition to provide an initial fusion estimation; in the deep reconstruction stage, we design a deep architecture with MAF to further improve the fusion effect. In the experiments, both quantity and quality comparisons were conducted for the proposed SDCS and other methods (FSDAF, MLFF-GAN, EDCSTFN, etc.). SDCS shows better performances in most cases on both abrupt change areas and transition areas. At the same time, the effectiveness of CS, AdaIN, and MAF in SDCS were evaluated in the ablation experiment. Overall, both of the contrast and ablation experiments confirmed the advantages of SDCS when compared with the other five methods on CIA, LGC, AHB, and Tianjin datasets.

REFERENCES

[1] Y. Zhang, P. Liu, L. Chen, M. Xu, X. Guo, and L. Zhao, "A new multi-source remote sensing image sample dataset with high resolution for flood area extraction: GF-FloodNet," *Int. J. Digit. Earth*, vol. 16, no. 1, pp. 2522–2554, 2023.

[2] L. Yue, B. Gao, and X. Zheng, "Generative DEM void filling with terrain feature-guided transfer learning assisted by remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 6011105.

[3] L. Yue, H. Shen, L. Zhang, X. Zheng, F. Zhang, and Q. Yuan, "High-quality seamless DEM generation blending SRTM-1, ASTER GDEM v2 and ICESat/GLAS observations," *ISPRS J. Photogramm. Remote Sens.*, vol. 123, pp. 20–34, 2017.

[4] P. Liu, J. Li, L. Wang, and G. He, "Remote sensing data fusion with generative adversarial networks: State-of-the-art methods and future research directions," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 295–328, Jun. 2022.

[5] L. Wang, B. Zuo, Y. Le, Y. Chen, and J. Li, "Penetrating remote sensing: Next-generation remote sensing for transparent earth," *Innov.*, vol. 4, no. 6, 2023, Art. no. 100519.

[6] S. Wang, W. Han, X. Zhang, J. Li, and L. Wang, "Geospatial remote sensing interpretation: From perception to cognition," *Innov. Geosci.*, vol. 2, no. 1, pp. 100056–1, 2024.

[7] B. Song et al., "MLFF-GAN: A multilevel feature fusion with gan for spatiotemporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4410816.

[8] X. Zhu, F. Cai, J. Tian, and T. Williams, "Spatiotemporal fusion of multisource remote sensing data: Literature survey, taxonomy, principles, applications, and future directions," *Remote Sens.*, vol. 10, no. 4, 2018, Art. no. 527.

[9] F. Gao, J. Masek, M. Schwaller, and F. Hall, "On the blending of the landsat and MODIS surface reflectance: Predicting daily Landsat surface reflectance," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 8, pp. 2207–2218, Aug. 2006.

[10] T. Hilker et al., "A new data fusion model for high spatial-and temporal-resolution mapping of forest disturbance based on Landsat and MODIS," *Remote Sens. Environ.*, vol. 113, no. 8, pp. 1613–1627, 2009.

[11] X. Zhu, J. Chen, F. Gao, X. Chen, and J. G. Masek, "An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions," *Remote Sens. Environ.*, vol. 114, no. 11, pp. 2610–2623, 2010.

[12] D. Fu, B. Chen, J. Wang, X. Zhu, and T. Hilker, "An improved image fusion approach based on enhanced spatial and temporal the adaptive reflectance fusion model," *Remote Sens.*, vol. 5, no. 12, pp. 6346–6360, 2013.

[13] B. Wu, B. Huang, K. Cao, and G. Zhuo, "Improving spatiotemporal reflectance fusion using image inpainting and steering kernel regression techniques," *Int. J. Remote Sens.*, vol. 38, no. 3, pp. 706–727, 2017.

[14] Q. Wang and P. M. Atkinson, "Spatio-temporal fusion for daily Sentinel-2 images," *Remote Sens. Environ.*, vol. 204, pp. 31–42, 2018.

[15] B. Zhukov, D. Oertel, F. Lanzl, and G. Reinhäckel, "Unmixing-based multisensor multiresolution image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 3, pp. 1212–1226, May 1999.

[16] R. Zurita-Milla, J. G. Clevers, and M. E. Schaepman, "Unmixing-based landsat TM and MERIS FR data fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 3, pp. 453–457, Jul. 2008.

[17] M. Wu, Z. Niu, C. Wang, C. Wu, and L. Wang, "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.*, vol. 6, no. 1, 2012, Art. no. 063507.

[18] W. Zhang et al., "An enhanced spatial and temporal data fusion model for fusing Landsat and MODIS surface reflectance to generate high temporal Landsat-like data," *Remote Sens.*, vol. 5, no. 10, pp. 5346–5368, 2013.

[19] M. Wu, W. Huang, Z. Niu, and C. Wang, "Generating daily synthetic Landsat imagery by combining Landsat and MODIS data," *Sensors*, vol. 15, no. 9, pp. 24002–24025, 2015.

[20] M. Lu, J. Chen, H. Tang, Y. Rao, P. Yang, and W. Wu, "Land cover change detection by integrating object-based data blending model of Landsat and MODIS," *Remote Sens. Environ.*, vol. 184, pp. 374–386, 2016.

[21] Y. Xu, B. Huang, Y. Xu, K. Cao, C. Guo, and D. Meng, "Spatial and temporal image fusion via regularized spatial unmixing," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 6, pp. 1362–1366, Jun. 2015.

[22] D. Xie et al., "An improved starfm with help of an unmixing-based method to generate high spatial and temporal resolution remote sensing data in complex heterogeneous regions," *Sensors*, vol. 16, no. 2, 2016, Art. no. 207.

[23] X. Zhu, E. H. Helmer, F. Gao, D. Liu, J. Chen, and M. A. Lefsky, "A flexible spatiotemporal method for fusing satellite images with different resolutions," *Remote Sens. Environ.*, vol. 172, pp. 165–177, 2016.

[24] A. Li, Y. Bo, Y. Zhu, P. Guo, J. Bi, and Y. He, "Blending multi-resolution satellite sea surface temperature (SST) products using bayesian maximum entropy method," *Remote Sens. Environ.*, vol. 135, pp. 52–63, 2013.

[25] B. Huang, H. Zhang, H. Song, J. Wang, and C. Song, "Unified fusion of remote-sensing imagery: Generating simultaneously high-resolution synthetic spatial–temporal–spectral earth observations," *Remote Sens. Lett.*, vol. 4, no. 6, pp. 561–569, 2013.

[26] H. Shen, X. Meng, and L. Zhang, "An integrated framework for the spatio–temporal–spectral fusion of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7135–7148, Dec. 2016.

[27] L. Liao, J. Song, J. Wang, Z. Xiao, and J. Wang, "Bayesian method for building frequent Landsat-like NDVI datasets by integrating MODIS and Landsat NDVI," *Remote Sens.*, vol. 8, no. 6, 2016, Art. no. 452.

[28] J. Xue, Y. Leung, and T. Fung, "A Bayesian data fusion approach to spatiotemporal fusion of remotely sensed images," *Remote Sens.*, vol. 9, no. 12, 2017, Art. no. 1310.

[29] H. Song and B. Huang, "Spatiotemporal satellite image fusion through one-pair image learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 4, pp. 1883–1896, Apr. 2013.

[30] B. Wu, B. Huang, and L. Zhang, "An error-bound-regularized sparse coding for spatiotemporal reflectance fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6791–6803, Dec. 2015.

[31] J. Wei, L. Wang, P. Liu, and W. Song, "Spatiotemporal fusion of remote sensing images with structural sparsity and semi-coupled dictionary learning," *Remote Sens.*, vol. 9, no. 1, 2016, Art. no. 21.

[32] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[33] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[34] S. Li and B. Yang, "A new pan-sharpening method using a compressed sensing technique," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 2, pp. 738–746, Feb. 2011.

[35] C. Jiang, H. Zhang, H. Shen, and L. Zhang, "A practical compressed sensing-based pan-sharpening method," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 4, pp. 629–633, Jul. 2012.

[36] J. Wei, L. Wang, P. Liu, X. Chen, W. Li, and A. Y. Zomaya, "Spatiotemporal fusion of MODIS and Landsat-7 reflectance images via compressed sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7126–7139, Dec. 2017.

[37] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Transactions on Multimedia*, vol. 21, no. 12, Dec. 2019, pp. 3106–3121, doi: 10.1109/TMM.2019.2919431.

[38] P. Liu, L. Wang, R. Ranjan, G. He, and L. Zhao, "A survey on active deep learning: From model driven to data driven," *ACM Comput. Surv.*, vol. 54, no. 10 s, Sep. 2022, Art. no. 221.

[39] Q. Liu, X. Meng, X. Li, and F. Shao, "Detail injection-based spatiotemporal fusion for remote sensing images with land cover changes," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5401514.

[40] H. Song, Q. Liu, G. Wang, R. Hang, and B. Huang, "Spatiotemporal satellite image fusion using deep convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 821–829, Mar. 2018.

[41] Z. Tan, P. Yue, L. Di, and J. Tang, "Deriving high spatiotemporal remote sensing images using deep convolutional network," *Remote Sens.*, vol. 10, no. 7, 2018, Art. no. 1066.

[42] Z. Tan, L. Di, M. Zhang, L. Guo, and M. Gao, "An enhanced deep convolutional model for spatiotemporal image fusion," *Remote Sens.*, vol. 11, no. 24, 2019, Art. no. 2898.

[43] X. Liu, C. Deng, J. Chanussot, D. Hong, and B. Zhao, "StfNet: A two-stream convolutional neural network for spatiotemporal image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6552–6564, Sep. 2019.

[44] Y. Li, J. Li, L. He, J. Chen, and A. Plaza, "A new sensor bias-driven spatio-temporal fusion model based on convolutional neural networks," *Sci. China Inf. Sci.*, vol. 63, no. 4, 2020, Art. no. 140302.

[45] H. Zhang, Y. Song, C. Han, and L. Zhang, "Remote sensing image spatiotemporal fusion using a generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4273–4286, May 2021.

[46] Z. Tan, M. Gao, X. Li, and L. Jiang, "A flexible reference-insensitive spatiotemporal fusion model for remote sensing images using conditional generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5601413.

[47] J. Chen, L. Wang, R. Feng, P. Liu, W. Han, and X. Chen, "CycleGAN-STF: Spatiotemporal fusion via CycleGAN-based image generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5851–5865, Jul. 2021.

[48] C. M. Gevaert and F. J. García-Haro, "A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion," *Remote Sens. Environ.*, vol. 156, pp. 34–44, 2015.

[49] X. Li, F. Ling, G. M. Foody, Y. Ge, Y. Zhang, and Y. Du, "Generating a series of fine spatial and temporal resolution land cover maps by fusing coarse spatial resolution remotely sensed images and fine spatial resolution land cover maps," *Remote Sens. Environ.*, vol. 196, pp. 293–311, 2017.

[50] X. Zhang, S. Li, Z. Tan, and X. Li, "Enhanced wavelet based spatiotemporal fusion networks using cross-paired remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 211, pp. 281–297, 2024.

[51] D. P. Roy et al., "A general method to normalize Landsat reflectance data to nadir BRDF adjusted reflectance," *Remote Sens. Environ.*, vol. 176, pp. 255–271, 2016.

[52] P. Liu, M. Wang, L. Wang, and W. Han, "Remote-sensing image denoising with multi-sourced information," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 2, pp. 660–674, Feb. 2019.

[53] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D, Nonlinear Phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.

[54] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Trans. Image Process.*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.

[55] A. Barbu, "Learning real-time MRF inference for image denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2009, pp. 1574–1581.

[56] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[57] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2005, vol. 2, pp. 60–65.

[58] D. Zoran and Y. Weiss, "From learning models of natural image patches to whole image restoration," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 479–486.

[59] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[60] W. R. Tobler, "A computer movie simulating urban growth in the Detroit region," *Econ. Geography*, vol. 46, no. sup1, pp. 234–240, 1970.

[61] K. Kemp, *Encyclopedia of Geographic Information Science*. Los Angeles, Washington DC, USA: Sage, 2008.

[62] H. J. Miller, "Tobler's first law and spatial analysis," *Ann. Assoc. Amer. Geographers*, vol. 94, no. 2, pp. 284–289, 2004.

[63] P. Liu, D. Liu, and Z. Liu, "Selection of regularization parameter based on synchronous noise in total variation image restoration," in *Proc. 3rd Int. Conf. Digit. Image Process.*. 2011, vol. 8009, pp. 445–451.

[64] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[65] D. L. Donoho, "For most large underdetermined systems of equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution," *Commun. Pure Appl. Math.*, vol. 59, no. 7, pp. 907–934, Jul. 2006.

[66] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *C. R. Math.*, vol. 346, no. 9-10, pp. 589–592, May 2008.

[67] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, Mar. 2021.

[68] K. Kulkarni, S. Lohit, P. Turaga, R. Kerviche, and A. Ashok, "Reconnet: Non-iterative reconstruction of images from compressively sensed measurements," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 449–458.

[69] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4509–4522, Sep. 2017.

[70] X. Yuan, Y. Liu, J. Suo, and Q. Dai, "Plug-and-play algorithms for large-scale snapshot compressive imaging," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1447–1457.

[71] G. Ongie, A. Jalal, C. A. Metzler, R. G. Baraniuk, A. G. Dimakis, and R. Willett, "Deep learning techniques for inverse problems in imaging," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 39–56, May 2020.

[72] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[73] A. L. Maas et al., "Rectifier nonlinearities improve neural network acoustic models," in *Proc. Int. Conf. Mach. Learn.*, Atlanta, GA, USA, 2013, vol. 30, no. 1, Art. no. 3.

[74] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 1501–1510.

[75] L. Li, P. Liu, J. Wu, L. Wang, and G. He, "Spatiotemporal remote-sensing image fusion with patch-group compressed sensing," *IEEE Access*, vol. 8, pp. 209199–209211, 2020.

[76] Q. Zhao, L. Ji, Y. Su, Y. Zhao, and J. Shi, "SRSF-GAN: A super-resolution based spatial fusion with GAN for satellite images with different spatial and temporal resolutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5408019.

[77] H. Huang, W. He, H. Zhang, Y. Xia, and L. Zhang, "STFDiff: Remote sensing image spatiotemporal fusion with diffusion models," *Inf Fusion.*, vol. 111, 2024, Art. no. 102505.

[78] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5835–5843.

[79] G. Bradski and A. Adrian, "Learning OpenCV: Computer vision with the OpenCV library," *O'Reilly Media, Inc.*, 2008.