Robust Spatiotemporal Fusion of Satellite Images: A Constrained Convex Optimization Approach

Ryosuke Isono, *Graduate Student Member, IEEE*, Kazuki Naganuma, *Graduate Student Member, IEEE*, and Shunsuke Ono, *Senior Member, IEEE*

***Abstract*— This article proposes a novel spatiotemporal (ST) fusion framework for satellite images, named robust optimization-based ST fusion (ROSTF). ST fusion is a promising approach to resolve a tradeoff between the temporal and spatial resolution of satellite images. Although many ST fusion methods have been proposed, most of them are not designed to explicitly account for noise in observed images, despite the inevitable influence of noise caused by the measurement equipment and environment. Our ROSTF addresses this challenge by formulating noise removal and ST fusion as a unified optimization problem. First, we define observation models for satellite images that may be contaminated with random noise, outliers, and/or missing values. Next, we introduce certain assumptions that naturally hold between the observed images and the target high-resolution image. Then, based on these models and assumptions, we formulate the fusion problem as a constrained optimization problem and develop an efficient algorithm based on a preconditioned primal–dual splitting method (P-PDS) for solving the problem. The performance of ROSTF was verified using simulated and real data. The results show that ROSTF performs comparably to several state-of-the-art ST fusion methods in noiseless cases and outperforms them in noisy cases.**

***Index Terms*— Constrained optimization, primal–dual splitting method, spatiotemporal (ST) fusion.**

I. Introduction

The analysis of temporal image series is necessary and important in many remote sensing applications, such as vegetation/crop monitoring and estimation [1], evapotranspiration estimation [2], atmosphere monitoring [3], land-cover/land-use change detection [4], surface dynamic mapping [5], ecosystem monitoring [6], soil water content analysis [7], and detailed analysis of human–nature interactions [8]. These applications require time series of high-spatial-resolution (HR) images to properly model the ground surface. In addition, time series of high-temporal-resolution images are also needed to capture the changes in the ground surface that occur over short periods of time.

However, there is a tradeoff between the temporal and spatial resolution of satellite sensors, and no single sensor can satisfy both requirements simultaneously. For example, the Landsat sensors can acquire images with an HR of 30 m, but they have a revisit period of up to 16 dates. On the other hand, the Moderate Resolution Imaging Spectroradiometer (MODIS) sensors can acquire images for the same scene at least once per date, but the images are at a low spatial resolution (LR) of 500 m [9]. Therefore, the simultaneous acquisition of image series of high spatial and high temporal resolution is a major challenge in the remote sensing community [10]. A simple solution to this challenge is to perform super-resolution on the corresponding single LR image to estimate the unobserved HR image [11], [12]. However, it is too difficult because the spatial resolution gap between the two satellite images is often quite large.

Spatiotemporal (ST) fusion (ST fusion) addresses this challenge by utilizing pairs of HR and LR images taken on reference dates that are temporally close to the target date. Specifically, the unobserved HR image on the target date is estimated by combining detailed spatial structure extracted from the HR images on the reference dates and spectral changes captured from the differences between the LR images on the reference and target dates. In ideal situations, where a large number of reliable reference images are available, achieving accurate ST fusion would be straightforward because the correct spatial structure and spectral changes are readily available. In real-world applications, however, such situations are very rare. Therefore, to achieve the desired ST fusion in real-world applications, the following two requirements are important.

1) *Minimum Reference Dates:* ST fusion methods that require minimum reference dates are preferred. In many remote sensing applications, only one pair of images on a reference date may be available due to cloud contamination, inconsistencies in image acquisition timing, or other factors. In addition, preparing another pair of images can be time-consuming. Therefore, ST fusion methods using a single pair of HR and LR images on a reference date apply to a wider range of cases than those using multiple pairs, although such a situation is obviously challenging [13].

Manuscript received 1 February 2024; revised 19 March 2024; accepted 2 April 2024. Date of publication 8 April 2024; date of current version 15 April 2024. This work was supported in part by Japan Science and Technology Agency (JST) ACT-X under Grant JPMJAX23CJ; in part by JST Precursory Research for Embryonic Science and Technology under Grant JPMJPR21C4; in part by JST AdCORP under Grant JPMJKB2307; in part by Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant 22H03610, Grant 22H00512, and Grant 23H01415; and in part by the Grant-in-Aid for JSPS Fellows under Grant 23KJ0912. *(Corresponding author: Ryosuke Isono.)*

The authors are with the Department of Computer Science, Tokyo Institute of Technology, Yokohama 226-8503, Japan (e-mail: isono.r.ac@m.titech.ac.jp; naganuma.k.aa@m.titech.ac.jp; ono@c.titech.ac.jp).

Digital Object Identifier 10.1109/TGRS.2024.3385917

2) *Robustness to Noise:* Due to the measurement equipment and/or environment, satellite images are often contaminated with various types of noise, such as random noise, outliers, and missing values [14], [15]. An estimated HR image that remains noisy would significantly affect subsequent processing. Therefore, it is imperative to develop noise-robust ST fusion methods.

A. Prior Research

Many ST fusion methods have been proposed over the past decades. They are generally categorized into five groups [16]: unmixing-based, weight-function-based, Bayesian-based, learning-based, and hybrid methods. Unmixing-based methods estimate the pixel values of an HR image by unmixing the pixels of the input LR images based on the linear spectral mixing theory [17], [18]. Weight-function-based methods generate an HR image by combining information of all the input images based on weight functions [9], [19]. Bayesian-based methods use Bayesian estimation theory to fuse the input images in a probabilistic manner [20], [21]. In the Bayesian framework, ST fusion can be considered a maximum a posteriori (MAP) problem, that is, the goal is to obtain an HR image on the target date by maximizing its conditional probability relative to the input HR and LR images. Learning-based methods model the relationship between observed HR and LR image pairs and then predict an unobserved HR image using machine-learning algorithms such as sparse representation learning [22], [23], random forest [24], and neural networks [25], [26], [27], [28], [29], [30], [31]. Hybrid methods integrate two or more techniques from the above four categories [32], [33].

Some of the unmixing-based, weight-function-based, and Bayesian-based methods allow an arbitrary number of reference dates. In other words, they can handle the cases with a single reference date. However, they are sensitive to noise because they estimate an HR image at the pixel level based entirely on reference images, which may be noisy. Thus, if the input images are noisy, the output image will be severely degraded.

On the other hand, in the context of learning-based methods, a robust ST fusion network (RSFN) [28] has been established to account for Gaussian noise. RSFN automatically filters noise and prevents predictions from being corrupted by using convolutional neural networks (CNNs), generative adversarial networks (GANs), and an attention mechanism. Specifically, the RSFN uses the attention mechanism to ignore noisy pixels in two reference HR images and focus on clean pixels. This method only works in situations where noisy pixels in the two reference HR images do not appear at the same location. In real-world measurements, however, such situations are rare because noise generally contaminates the entire image, not just parts of it. Furthermore, as mentioned above, RSFN requires two reference dates.

B. Contributions and Paper Organization

Now, a natural question arises: *How to achieve robust ST fusion with only a single reference date?* In this article, we propose a novel ST fusion framework for satellite images, named *robust optimization-based ST fusion (ROSTF)*, to estimate an HR image on the target date while simultaneously denoising an HR image on a single reference date.

Before formulating our optimization problem, we newly define two observation models (they will be detailed in Section III-A).

1) The first model describes the relationship between an observed noisy image and the oracle noiseless image. The model is designed under the assumption that the observed image is not only contaminated with random noise but also with outliers and missing values. Specifically, random noise is modeled by Gaussian noise while outliers and missing values are modeled by sparse noise.
2) The second model represents the relationship between an HR image and the corresponding LR image, based on a super-resolution model [34].

We also introduce the following two assumptions about satellite images (they will be detailed in Section III-B).

1) The reflectance may change significantly between the reference and target dates, but the land structure (the locations of the edges) does not. This is a very natural assumption in the context of ST fusion.
2) An HR image and the corresponding LR image have similar brightness. This assumption is necessarily true if the HR and LR sensors have similar spectral resolutions, as is the case with sensors like Landsat and MODIS [9].

Based on the observation models and assumptions, we formulate the fusion problem as a constrained convex optimization problem. Subsequently, we develop an efficient algorithm based on the preconditioned primal–dual splitting method (P-PDS) [35] with an operator-norm-based design method of variable-wise diagonal preconditioning, named OVDP [36], which can automatically determine the appropriate stepsizes for solving the optimization problem.

The main contributions of the article are given as follows.

1) *Robustness to Random Noise, Outliers, and Missing Values:* As described above, no existing ST fusion methods can handle random noise, outliers, and missing values, although this type of noise contaminates satellite images due to the measurement equipment and environment. Thanks to the formulation that incorporates the first observation model developed in Section III-A, ROSTF is robust to such mixed noise.
2) *Single Reference Date:* Assumption 1) is very simple but effective for ST fusion. By incorporating this assumption as a constraint in the optimization problem, we realize the mechanism to promote the estimated HR image on the target date and the denoised HR image on the reference date to have edges at similar locations. Thanks to such a mechanism, ROSTF performs well even based on a single reference date.
3) *Facilitation of Parameter Adjustment:* The objective function of the optimization problem of ROSTF consists only of image regularization terms to promote spatial piecewise smoothness. The other components, corresponding to data fidelity based on the two observation models and our two assumptions, are imposed as

hard constraints. Such a formulation using constraints instead of adding terms to the objective function has the advantage of simplifying parameter setting [37], [38], [39], [40], [41], [42]: the appropriate parameters in the constraints do not depend on each other and can be determined independently for each constraint.

4) *Automatic Stepsize Adjustment:* To solve our optimization problem for ST fusion, we develop an efficient algorithm based on P-PDS with OVDP. The appropriate stepsizes of the standard PDS [43] (and most other optimization methods) would be different depending on the problem structure, which means that we have to adjust them manually. On the other hand, P-PDS with OVDP can automatically determine the appropriate stepsizes based on the problem structure, and thus our algorithm is free from such a troublesome task.

This paper is organized as follows. We first cover the mathematical preliminaries of our method in Section II and then proceed to the establishment of our method in Section III. In Section IV, we demonstrate the performance of ROSTF and the effectiveness of each key component of ROSTF through comparative experiments and ablation studies, respectively. Experimental results show that ROSTF performs comparably to several state-of-the-art ST fusion methods for noiseless images and outperforms them for noisy images, thanks to the effective work of each key component.

The preliminary version of this work, without considering sparse noise, mathematical details, comprehensive experimental comparison, deeper discussion, or implementation using P-PDS with OVDP, has appeared in conference proceedings [44].

II. Preliminaries

A. Notations

Let $\mathbb{R}$ be the set of all real numbers. Vectors and matrices are denoted by bold lower and upper case letters, respectively. We treat a multispectral image with spatial resolution $N_1 \times N_2$ and spectral resolution (the number of bands) B as a vector $\mathbf{x} = ([\mathbf{x}]_1^\top, \dots, [\mathbf{x}]_B^\top)^\top \in \mathbb{R}^{NB} (N = N_1 N_2)$, where $[\mathbf{x}]_b \in \mathbb{R}^N$ is the vector representing the bth band pixel values. Here, the nth pixel of $[\mathbf{x}]_b$ is denoted by $x_{b,n} \in \mathbb{R}$. Let $\Gamma_0(\mathbb{R}^{NB})$ be the set of all proper lower-semicontinuous convex functions defined on $\mathbb{R}^{NB}$. The ℓ_1-norm, ℓ_2-norm, and $\ell_{1,2}$-norm of $\mathbf{x}$ are defined as $\|\mathbf{x}\|_1 := \sum_b \sum_n |x_{b,n}|$, $\|\mathbf{x}\|_2 := (\sum_b \sum_n |x_{b,n}|^2)^{1/2}$, and $\|\mathbf{x}\|_{1,2} := \sum_n (\sum_b |x_{b,n}|^2)^{1/2}$, respectively. For an image $\mathbf{x} \in \mathbb{R}^{NB}$, let $\mathbf{D}_v$ and $\mathbf{D}_h \in \mathbb{R}^{NB \times NB}$ denote the matrices for computing the differences between vertical/horizontal adjacent pixels, respectively, and let $\mathbf{D} := (\mathbf{D}_v^\top\ \mathbf{D}_h^\top)^\top \in \mathbb{R}^{2NB \times NB}$. The hyperslab centered at ω with a radius α is denoted as

$$S_\alpha^\omega := \left\{ \mathbf{z} \middle| |\omega - \mathbf{1}^\top \mathbf{z}| \leq \alpha \right\}. \tag{1}$$

Also, the ℓ_p-norm ball ($p = 1, 2$) with the center $\mathbf{c}$ and radius ε is denoted as

$$B_{p,\varepsilon}^{\mathbf{c}} := \left\{ \mathbf{z} \middle| \|\mathbf{z} - \mathbf{c}\|_p \leq \varepsilon \right\}. \tag{2}$$

The indicator function $\iota_C : \mathbb{R}^N \to (-\infty, \infty]$ of a nonempty closed convex set C is defined as

$$\iota_C := \begin{cases} 0, & \text{if}\, \mathbf{x} \in C, \\ \infty, & \text{otherwise}. \end{cases} \tag{3}$$

B. Proximal Tools

The optimization problem of ROSTF, to be formulated in Section III-B, involves nonsmooth convex functions. To solve such a problem, we introduce the notion of the *proximity operator* of index $\gamma > 0$ of $f \in \Gamma_0(\mathbb{R}^{NB})$ as follows:

$$\mathrm{prox}_{\gamma f} : \mathbb{R}^{NB} \to \mathbb{R}^{NB} : \mathbf{x} \mapsto \underset{\mathbf{y} \in \mathbb{R}^{NB}}{\mathrm{argmin}}\, f(\mathbf{y}) + \frac{1}{2\gamma} \|\mathbf{x} - \mathbf{y}\|_2^2. \tag{4}$$

The Fenchell–Rockerfellar conjugate function f^* of the function $f \in \Gamma_0(\mathbb{R}^{NB})$ is denoted by

$$f^*(\mathbf{y}) := \sup_{\mathbf{x} \in \mathbb{R}^{NB}} \{\langle \mathbf{x}, \mathbf{y} \rangle - f(\mathbf{x})\}. \tag{5}$$

Thanks to Moreau's identity [45], the proximity operator of f^* is efficiently calculated as

$$\mathrm{prox}_{\gamma f^*}(\mathbf{x}) = \mathbf{x} - \gamma \mathrm{prox}_{\frac{1}{\gamma} f}\left(\frac{1}{\gamma}\mathbf{x}\right). \tag{6}$$

Below, we present the specific proximity operators of the functions used in this article. The proximity operator of the $\ell_{1,2}$-norm is given by

$$\left[\mathrm{prox}_{\gamma \|\cdot\|_{1,2}}(\mathbf{x})\right]_{b,n} = \max\left\{1 - \frac{\gamma}{\sqrt{\sum_{b'=1}^{B} |x_{b',n}|^2}}, 0\right\} x_{b,n}. \tag{7}$$

The proximity operator of the indicator function of the hyperslab in (1) is expressed as follows:

$$\mathrm{prox}_{\gamma \iota_{S_\alpha^\omega}}(\mathbf{x}) = \begin{cases} \mathbf{x} + \dfrac{\eta_1 - \mathbf{1}^\top \mathbf{x}}{NB} \mathbf{1}, & \text{if}\, \mathbf{1}^\top \mathbf{x} < \eta_1, \\ \mathbf{x} + \dfrac{\eta_2 - \mathbf{1}^\top \mathbf{x}}{NB} \mathbf{1}, & \text{if}\, \mathbf{1}^\top \mathbf{x} > \eta_2, \\ \mathbf{x}, & \text{otherwise} \end{cases} \tag{8}$$

where $\eta_1 = \omega - \alpha$ and $\eta_2 = \omega + \alpha$. The proximity operators of the indicator functions of the ℓ_2-norm ball and the ℓ_1-norm ball in (2) are calculated by

$$\mathrm{prox}_{\gamma \iota_{B_{2,\varepsilon}^{\mathbf{c}}}}(\mathbf{x}) = \begin{cases} \mathbf{x}, & \text{if}\, \mathbf{x} \in \iota_{B_{2,\varepsilon}^{\mathbf{c}}}, \\ \mathbf{c} + \dfrac{\varepsilon(\mathbf{x} - \mathbf{c})}{\|\mathbf{x} - \mathbf{c}\|_2}, & \text{otherwise} \end{cases} \tag{9}$$

and a fast ℓ_1-ball projection algorithm [46], respectively.

C. P-PDS With OVDP

The standard PDS [43] is a versatile and efficient proximal algorithm that can solve a wide class of nonsmooth convex optimization problems without using matrix inversion. However, it is troublesome to manually set the appropriate stepsizes of the standard PDS. Therefore, we adopt P-PDS [35] with OVDP [36], a method that automatically determines the appropriate stepsizes according to the problem structure.

Let $\mathbf{y}_i \in \mathbb{R}^{K_i}$ $(i = 1, \dots, N)$ and $\mathbf{z}_j \in \mathbb{R}^{L_j}$ $(j = 1, \dots, M)$. Consider convex optimization problems of the following form:

$$\min_{\substack{\mathbf{y}_1,\dots,\mathbf{y}_N\\ \mathbf{z}_1,\dots,\mathbf{z}_M}} \sum_{i=1}^{N} g_i(\mathbf{y}_i) + \sum_{j=1}^{M} h_j(\mathbf{z}_j),$$
$$\text{s.t.} \begin{cases} \mathbf{z}_1 = \sum_{i=1}^{N} \mathbf{G}_{1,i}\mathbf{y}_i, \\ \vdots \\ \mathbf{z}_M = \sum_{i=1}^{N} \mathbf{G}_{M,i}\mathbf{y}_i \end{cases} \tag{10}$$

where $g_i \in \Gamma_0(\mathbb{R}^{K_i})$ $(i = 1, \dots, N), h_j \in \Gamma_0(\mathbb{R}^{L_j})$ $(j = 1, \dots, M)$, and $\mathbf{G}_{j,i} : \mathbb{R}^{K_i} \to \mathbb{R}^{L_j}$ $(i = 1, \dots, N,\ j = 1, \dots, M)$ are linear operators. P-PDS with OVDP solves (10) by the following iterative procedures:

$$\begin{cases} \bar{\mathbf{y}}_1^{(n)} & \leftarrow \mathbf{y}_1^{(n)} - \gamma_{1,1} \sum_{j=1}^{M} \mathbf{G}_{j,1}^{\top}\mathbf{z}_j^{(n)}, \\ \mathbf{y}_1^{(n+1)} & \leftarrow \mathrm{prox}_{\gamma_{1,1} g_1}\left(\bar{\mathbf{y}}_1^{(n)}\right), \\ & \vdots \\ \bar{\mathbf{y}}_N^{(n)} & \leftarrow \mathbf{y}_N^{(n)} - \gamma_{1,N} \sum_{j=1}^{M} \mathbf{G}_{j,N}^{\top}\mathbf{z}_j^{(n)}, \\ \mathbf{y}_N^{(n+1)} & \leftarrow \mathrm{prox}_{\gamma_{1,N} g_N}\left(\bar{\mathbf{y}}_N^{(n)}\right), \\ \bar{\mathbf{z}}_1^{(n)} & \leftarrow \mathbf{z}_1^{(n)} + \gamma_{2,1} \sum_{i=1}^{N} \mathbf{G}_{1,i}\left(2\mathbf{y}_i^{(n+1)} - \mathbf{y}_i^{(n)}\right), \\ \mathbf{z}_1^{(n+1)} & \leftarrow \mathrm{prox}_{\gamma_{2,1} h_1^*}\left(\bar{\mathbf{z}}_1^{(n)}\right), \\ & \vdots \\ \bar{\mathbf{z}}_M^{(n)} & \leftarrow \mathbf{z}_M^{(n)} + \gamma_{2,M} \sum_{i=1}^{N} \mathbf{G}_{M,i}\left(2\mathbf{y}_i^{(n+1)} - \mathbf{y}_i^{(n)}\right), \\ \mathbf{z}_M^{(n+1)} & \leftarrow \mathrm{prox}_{\gamma_{2,M} h_M^*}\left(\bar{\mathbf{z}}_M^{(n)}\right) \end{cases}$$

where $\gamma_{1,i}$ $(i = 1, \dots, N)$ and $\gamma_{2,j}$ $(j = 1, \dots, M)$ are stepsize parameters. The stepsize parameters can be determined as follows [36]:

$$\gamma_{1,i} = \frac{1}{\sum_{j=1}^{M} \left\|\mathbf{G}_{j,i}\right\|_{\mathrm{op}}^{2}}, \qquad \gamma_{2,j} = \frac{1}{N} \tag{11}$$

where $\|\cdot\|_{\mathrm{op}}$ represents the operator norm defined by

$$\|\mathbf{G}\|_{\mathrm{op}} := \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{G}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}. \tag{12}$$

III. Proposed Method

From now on, we will focus on cases with a single reference date. Specifically, we consider a situation where both HR and LR sensors observe the same scene on the single reference date, but on the target date, only the LR sensor observes that scene and not the HR sensor. Let the HR image on the reference date, the LR image on the reference date, and the LR image on the target date be $\mathbf{h}_r \in \mathbb{R}^{N_h B}$, $\mathbf{l}_r \in \mathbb{R}^{N_l B}$, and $\mathbf{l}_t \in \mathbb{R}^{N_l B}$, respectively. Our method, ROSTF, estimates the desired noiseless HR image on the target date, denoted by $\widehat{\mathbf{h}}_t \in \mathbb{R}^{N_h B}$, based on these three observed images, while simultaneously denoising $\mathbf{h}_r$. A general diagram of ROSTF is shown in Fig. 1.

A. Observation Models

Let $\widehat{\mathbf{h}} \in \mathbb{R}^{N_h B}$ and $\widehat{\mathbf{l}} \in \mathbb{R}^{N_l B}$ be a noiseless HR image and a noiseless LR image, respectively, taken on the same date. We introduce observation models for a noisy HR image $\mathbf{h} \in \mathbb{R}^{N_h B}$ and a noisy LR image $\mathbf{l} \in \mathbb{R}^{N_l B}$. Specifically, we consider that the observed satellite images $\mathbf{h}$ and $\mathbf{l}$ are possibly contaminated with random noise, outliers, and missing values. Random noise added to $\widehat{\mathbf{h}}$ and $\widehat{\mathbf{l}}$ is modeled by Gaussian noise $\mathbf{n}_h$ and $\mathbf{n}_l$ with standard deviation σ_h and σ_l, respectively, while outliers and missing values affecting $\widehat{\mathbf{h}}$ and $\widehat{\mathbf{l}}$ are modeled by sparsely distributed noise $\mathbf{s}_h$ and $\mathbf{s}_l$ with the superimposition ratio r_h and r_l, respectively. By modeling the noise in this manner, the observation models for $\mathbf{h}$ and $\mathbf{l}$ are described as

$$\begin{aligned} \mathbf{h} &= \widehat{\mathbf{h}} + \mathbf{n}_h + \mathbf{s}_h, \\ \mathbf{l} &= \widehat{\mathbf{l}} + \mathbf{n}_l + \mathbf{s}_l. \end{aligned} \tag{13}$$

Here, $\sigma_h > \sigma_l$ and $r_h > r_l$ generally hold since HR images often contain more severe noise than LR images. This is because the amount of light received per pixel decreases as the number of pixels increases [47].

On the other hand, $\widehat{\mathbf{l}}$ can be approximated by the image obtained by blurring and downsampling $\widehat{\mathbf{h}}$, known as a typical super-resolution model [34], as follows:

$$\widehat{\mathbf{l}} = \mathbf{S}\mathbf{B}\widehat{\mathbf{h}} + \mathbf{m} \tag{14}$$

where $\mathbf{B} \in \mathbb{R}^{N_h B \times N_h B}$ is the spatial spread transform matrix introduced in [48], $\mathbf{S} \in \mathbb{R}^{N_l B \times N_h B}$ is the downsampling matrix, and $\mathbf{m} \in \mathbb{R}^{N_l B}$ is the modeling error. This model has been widely used in the ST fusion literature [49].

B. Problem Formulation

We introduce the following two assumptions about the noiseless HR and LR images on the reference and target dates, that is, $\widehat{\mathbf{h}}_r$, $\widehat{\mathbf{l}}_r$, $\widehat{\mathbf{h}}_t$, and $\widehat{\mathbf{l}}_t$.

1) The reflectance may change significantly between the reference and target dates, but the land structure does not. This is a very natural assumption in ST fusion and is implicitly accepted in previous studies. If the land structure has not changed significantly, the edges of $\widehat{\mathbf{h}}_r$ and $\widehat{\mathbf{h}}_t$ appear at almost the same locations, implying that the difference between $\mathbf{D}\widehat{\mathbf{h}}_r$ and $\mathbf{D}\widehat{\mathbf{h}}_t$ tends to be small. We measure the similarity of these edges using the ℓ_p $(p = 1$ or $2)$ norm as $\|\mathbf{D}\widehat{\mathbf{h}}_r - \mathbf{D}\widehat{\mathbf{h}}_t\|_p$.
2) The HR and LR images taken on the same date have similar average brightness per band. For example, the difference in average brightness of the bth band of $\widehat{\mathbf{h}}_t$ and $\widehat{\mathbf{l}}_t$, expressed as
$$\left| \frac{1}{N_l}\mathbf{1}^{\top}\left[\widehat{\mathbf{l}}_t\right]_b - \frac{1}{N_h}\mathbf{1}^{\top}\left[\widehat{\mathbf{h}}_t\right]_b \right| \tag{15}$$
is expected to be small. This is necessarily true if the HR and LR sensors have similar spectral resolutions, as is the case for Landsat and MODIS [9].

Based on these assumptions and the observation models in (13) and (14), we formulate the fusion problem as the

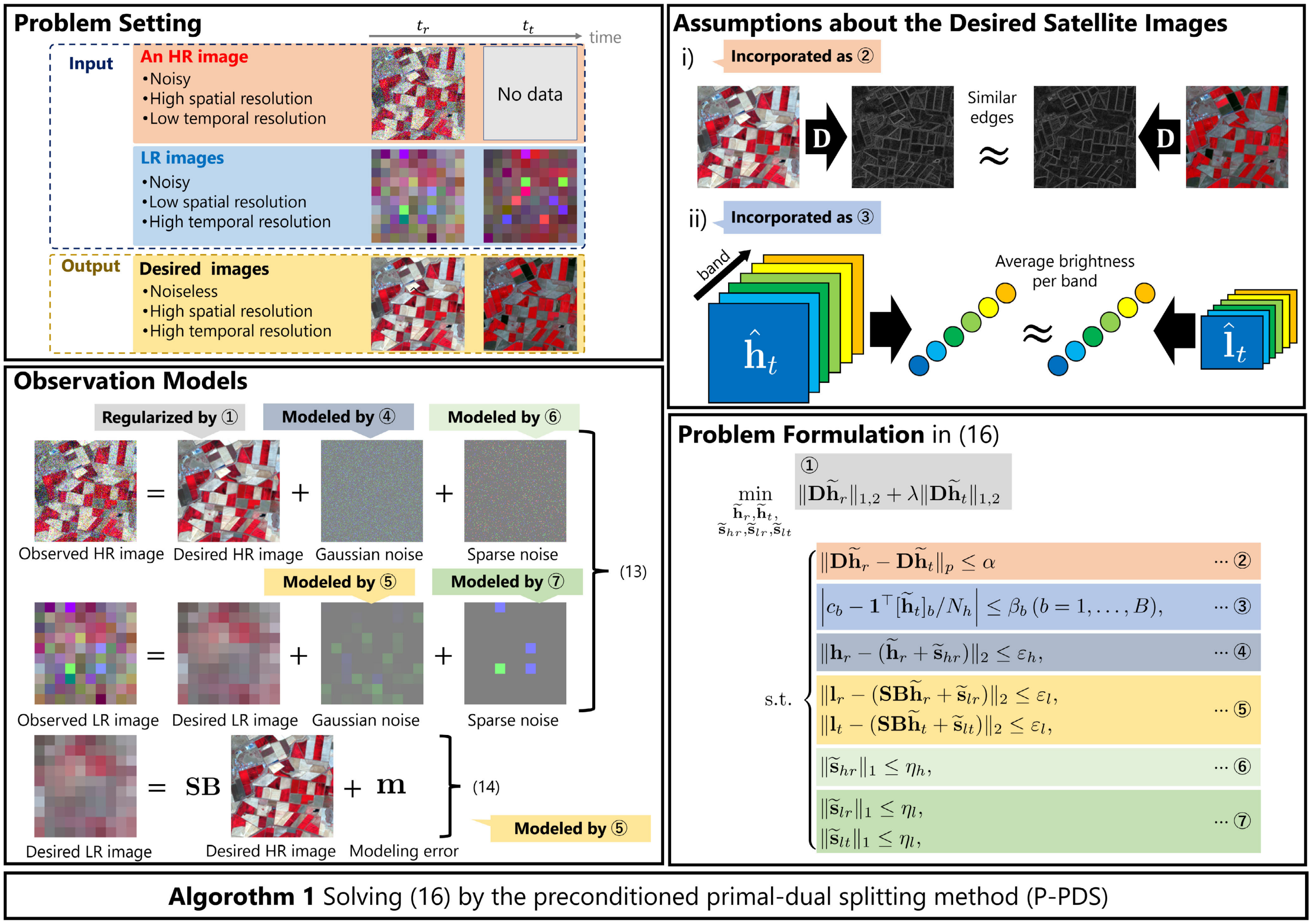
Problem Setting
t_r
t_t
time
Input
An HR image
•Noisy
•High spatial resolution
•Low temporal resolution
No data
LR images
•Noisy
•Low spatial resolution
•High temporal resolution
Output
Desired images
•Noiseless
•High spatial resolution
•High temporal resolution
Assumptions about the Desired Satellite Images
i)
Incorporated as ②
D
Similar edges
≈
D
ii)
Incorporated as ③
band
$\hat{\mathbf{h}}_t$
Average brightness per band
≈
$\hat{\mathbf{l}}_t$
Observation Models
Regularized by ①
Modeled by ④
Modeled by ⑥
=
+
+
Observed HR image
Desired HR image
Gaussian noise
Sparse noise
(13)
Modeled by ⑤
Modeled by ⑦
=
+
+
Observed LR image
Desired LR image
Gaussian noise
Sparse noise
= SB + m
(14)
Modeled by ⑤
Desired LR image
Desired HR image
Modeling error
Problem Formulation in (16)
①
$\min_{\widetilde{\mathbf{h}}_r,\widetilde{\mathbf{h}}_t,\widetilde{\mathbf{s}}_{hr},\widetilde{\mathbf{s}}_{lr},\widetilde{\mathbf{s}}_{lt}} \|\mathbf{D}\widetilde{\mathbf{h}}_r\|_{1,2}+\lambda\|\mathbf{D}\widetilde{\mathbf{h}}_t\|_{1,2}$
s.t.
$\|\mathbf{D}\widetilde{\mathbf{h}}_r-\mathbf{D}\widetilde{\mathbf{h}}_t\|_p\leq\alpha$ ··· ②
$\left|c_b-\mathbf{1}^\top[\widetilde{\mathbf{h}}_t]_b/N_h\right|\leq\beta_b\,(b=1,\ldots,B),$ ··· ③
$\|\mathbf{h}_r-(\widetilde{\mathbf{h}}_r+\widetilde{\mathbf{s}}_{hr})\|_2\leq\varepsilon_h,$ ··· ④
$\|\mathbf{l}_r-(\mathbf{S}\mathbf{B}\widetilde{\mathbf{h}}_r+\widetilde{\mathbf{s}}_{lr})\|_2\leq\varepsilon_l,$
$\|\mathbf{l}_t-(\mathbf{S}\mathbf{B}\widetilde{\mathbf{h}}_t+\widetilde{\mathbf{s}}_{lt})\|_2\leq\varepsilon_l,$ ··· ⑤
$\|\widetilde{\mathbf{s}}_{hr}\|_1\leq\eta_h,$ ··· ⑥
$\|\widetilde{\mathbf{s}}_{lr}\|_1\leq\eta_l,$
$\|\widetilde{\mathbf{s}}_{lt}\|_1\leq\eta_l,$ ··· ⑦
Algorothm 1 Solving (16) by the preconditioned primal-dual splitting method (P-PDS)

Fig. 1. Illustration of our method, that is, ROSTF.

following constrained convex optimization problem:

$$\min_{\substack{\widetilde{\mathbf{h}}_r,\widetilde{\mathbf{h}}_t,\\ \widetilde{\mathbf{s}}_{hr},\widetilde{\mathbf{s}}_{lr},\widetilde{\mathbf{s}}_{lt}}} \left\|\mathbf{D}\widetilde{\mathbf{h}}_r\right\|_{1,2}+\lambda\left\|\mathbf{D}\widetilde{\mathbf{h}}_t\right\|_{1,2}$$

$$\text{s.t.}\begin{cases}\|\mathbf{D}\widetilde{\mathbf{h}}_r-\mathbf{D}\widetilde{\mathbf{h}}_t\|_p\leq\alpha,\\ \left|c_b-\mathbf{1}^\top\left[\widetilde{\mathbf{h}}_t\right]_b/N_h\right|\leq\beta_b\,(b=1,\ldots,B),\\ \|\mathbf{h}_r-\left(\widetilde{\mathbf{h}}_r+\widetilde{\mathbf{s}}_{hr}\right)\|_2\leq\varepsilon_h,\\ \|\mathbf{l}_r-\left(\mathbf{S}\mathbf{B}\widetilde{\mathbf{h}}_r+\widetilde{\mathbf{s}}_{lr}\right)\|_2\leq\varepsilon_l,\\ \|\mathbf{l}_t-\left(\mathbf{S}\mathbf{B}\widetilde{\mathbf{h}}_t+\widetilde{\mathbf{s}}_{lt}\right)\|_2\leq\varepsilon_l,\\ \|\widetilde{\mathbf{s}}_{hr}\|_1\leq\eta_h,\\ \|\widetilde{\mathbf{s}}_{lr}\|_1\leq\eta_l,\\ \|\widetilde{\mathbf{s}}_{lt}\|_1\leq\eta_l\end{cases}\tag{16}$$

where λ is a balancing parameter. The variables $\widetilde{\mathbf{h}}_r$ and $\widetilde{\mathbf{h}}_t$ correspond to the estimates of $\widehat{\mathbf{h}}_r$ and $\widehat{\mathbf{h}}_t$, respectively, and $\widetilde{\mathbf{s}}_{hr}$, $\widetilde{\mathbf{s}}_{lr}$, and $\widetilde{\mathbf{s}}_{lt}$ correspond to the estimates of sparse noise superimposed on $\mathbf{h}_r$, $\mathbf{l}_r$ and $\mathbf{l}_t$, respectively. Each term in the objective function and each constraint have the following roles.

1) The two terms in the objective function promote spatial piecewise smoothness of $\widetilde{\mathbf{h}}_r$ and $\widetilde{\mathbf{h}}_t$, with the hyperspectral total variation (HTV) [50] as regularization.
2) The first constraint encourages $\mathbf{D}\widetilde{\mathbf{h}}_r$ and $\mathbf{D}\widetilde{\mathbf{h}}_t$ to be similar based on Assumption 1). The parameter α controls the degree of similarity. Hereafter, the constraint is referred to as the *edge constraint*.
3) The second constraint is designed based on Assumption 2). Since $\mathbf{l}_t$ is contaminated by noise, we do not use the average brightness of $[\mathbf{l}_t]_b$ itself, that is, $\mathbf{1}^\top[\mathbf{l}_t]_b/N_l$, but the parameter c_b, which is determined based on $\mathbf{1}^\top[\mathbf{l}_t]_b/N_l$ and the noise intensity. The parameter β_b controls the strength of this constraint. Hereafter, the constraint is referred to as the *brightness constraint*.
4) The third constraint serves as data-fidelity based on the observation model in (13). The parameter ε_h depends on the Gaussian noise intensity on the HR image, that is, σ_h.
5) The fourth and fifth constraints are to ensure that the solutions follow the observation model in (14). The parameter ε_l depends on the Gaussian noise intensity on the LR images, that is, σ_l.
6) The last three constraints characterize the sparse noise using the ℓ_1 norms. The parameters η_h and η_l depend on the sparse noise intensity on the HR and the LR images, respectively, that is, r_h and r_l.

Algorithm 1 P-PDS-Based Solver for (16)

```
Require: $\lambda, p, \alpha, \beta'_b, c'_b, \varepsilon_h, \varepsilon_l, \eta_h, \eta_l$
Ensure: $\widetilde{\mathbf{h}}_r^{(n)}, \widetilde{\mathbf{h}}_t^{(n)}, \widetilde{\mathbf{s}}_{hr}^{(n)}, \widetilde{\mathbf{s}}_{lr}^{(n)}, \widetilde{\mathbf{s}}_{lt}^{(n)}$
Initialize $\widetilde{\mathbf{h}}_r^{(0)}, \widetilde{\mathbf{h}}_t^{(0)}, \widetilde{\mathbf{s}}_{hr}^{(0)}, \widetilde{\mathbf{s}}_{lr}^{(0)}, \widetilde{\mathbf{s}}_{lt}^{(0)}, \mathbf{z}_j^{(0)} (j = 1, \dots, 6)$;
Set $\gamma_{1,i} (i = 1, \dots, 3), \gamma_{2,j} (j = 1, \dots, 6)$, as in (20);
while until a stopping criterion is not satisfied do
    $\mathbf{u}_r \leftarrow \mathbf{D}^\top \mathbf{z}_1^{(n)} + \mathbf{D}^\top \mathbf{z}_3^{(n)} + \mathbf{z}_4^{(n)} + \mathbf{B}^\top \mathbf{S}^\top \mathbf{z}_5^{(n)}$;
    $\mathbf{u}_t \leftarrow \mathbf{D}^\top \mathbf{z}_2^{(n)} - \mathbf{D}^\top \mathbf{z}_3^{(n)} + \mathbf{B}^\top \mathbf{S}^\top \mathbf{z}_6^{(n)}$;
    $\widetilde{\mathbf{h}}_r^{(n+1)} \leftarrow \widetilde{\mathbf{h}}_r^{(n)} - \gamma_{1,1} \mathbf{u}_r$;
    $\widetilde{\mathbf{h}}_t^{(n+1)} \leftarrow \widetilde{\mathbf{h}}_t^{(n)} - \gamma_{1,2} \mathbf{u}_t$;
    for $b = 1, \dots, B$ do
        $[\widetilde{\mathbf{h}}_t^{(n+1)}]_b \leftarrow \mathrm{prox}_{\iota_{S_{\beta'_b}^{c'_b}}} ([\widetilde{\mathbf{h}}_t^{(n+1)}]_b)$;
    end for
    $\widetilde{\mathbf{s}}_{hr}^{(n+1)} \leftarrow \mathrm{prox}_{\gamma_{1,3} \iota_{B_{1,\eta_h}^0}} (\widetilde{\mathbf{s}}_{hr}^{(n)} - \gamma_{1,3} \mathbf{z}_4)$;
    $\widetilde{\mathbf{s}}_{lr}^{(n+1)} \leftarrow \mathrm{prox}_{\gamma_{1,4} \iota_{B_{1,\eta_l}^0}} (\widetilde{\mathbf{s}}_{lr}^{(n)} - \gamma_{1,4} \mathbf{z}_5)$;
    $\widetilde{\mathbf{s}}_{lt}^{(n+1)} \leftarrow \mathrm{prox}_{\gamma_{1,5} \iota_{B_{1,\eta_l}^0}} (\widetilde{\mathbf{s}}_{lt}^{(n)} - \gamma_{1,5} \mathbf{z}_6)$;
    $\mathbf{v}_r \leftarrow 2\widetilde{\mathbf{h}}_r^{(n+1)} - \widetilde{\mathbf{h}}_r^{(n)}$;
    $\mathbf{v}_t \leftarrow 2\widetilde{\mathbf{h}}_t^{(n+1)} - \widetilde{\mathbf{h}}_t^{(n)}$;
    $\mathbf{w}_{hr} \leftarrow 2\widetilde{\mathbf{s}}_{hr}^{(n+1)} - \widetilde{\mathbf{s}}_{hr}^{(n)}$;
    $\mathbf{w}_{lr} \leftarrow 2\widetilde{\mathbf{s}}_{lr}^{(n+1)} - \widetilde{\mathbf{s}}_{lr}^{(n)}$;
    $\mathbf{w}_{lt} \leftarrow 2\widetilde{\mathbf{s}}_{lt}^{(n+1)} - \widetilde{\mathbf{s}}_{lt}^{(n)}$;
    $\mathbf{z}_1^{(n)} \leftarrow \mathbf{z}_1^{(n)} + \gamma_{2,1} \mathbf{D}\mathbf{v}_r$;
    $\mathbf{z}_2^{(n)} \leftarrow \mathbf{z}_2^{(n)} + \gamma_{2,2} \mathbf{D}\mathbf{v}_t$;
    $\mathbf{z}_3^{(n)} \leftarrow \mathbf{z}_3^{(n)} + \gamma_{2,3} (\mathbf{D}\mathbf{v}_r - \mathbf{D}\mathbf{v}_t)$;
    $\mathbf{z}_4^{(n)} \leftarrow \mathbf{z}_4^{(n)} + \gamma_{2,4} (\mathbf{v}_r + \mathbf{w}_{hr})$;
    $\mathbf{z}_5^{(n)} \leftarrow \mathbf{z}_5^{(n)} + \gamma_{2,5} (\mathbf{S}\mathbf{B}\mathbf{v}_r + \mathbf{w}_{lr})$;
    $\mathbf{z}_6^{(n)} \leftarrow \mathbf{z}_6^{(n)} + \gamma_{2,6} (\mathbf{S}\mathbf{B}\mathbf{v}_t + \mathbf{w}_{lt})$;
    $\mathbf{z}_1^{(n+1)} \leftarrow \mathbf{z}_1^{(n)} - \gamma_{2,1} \mathrm{prox}_{\frac{1}{\gamma_{2,1}} \|\cdot\|_{1,2}} (\frac{1}{\gamma_{2,1}} \mathbf{z}_1^{(n)})$;
    $\mathbf{z}_2^{(n+1)} \leftarrow \mathbf{z}_2^{(n)} - \gamma_{2,2} \mathrm{prox}_{\frac{\lambda}{\gamma_{2,2}} \|\cdot\|_{1,2}} (\frac{1}{\gamma_{2,2}} \mathbf{z}_2^{(n)})$;
    $\mathbf{z}_3^{(n+1)} \leftarrow \mathbf{z}_3^{(n)} - \gamma_{2,3} \mathrm{prox}_{\iota_{B_{p,\alpha}^{\mathbf{0}}}} (\frac{1}{\gamma_{2,3}} \mathbf{z}_3^{(n)})$;
    $\mathbf{z}_4^{(n+1)} \leftarrow \mathbf{z}_4^{(n)} - \gamma_{2,4} \mathrm{prox}_{\iota_{B_{2,\varepsilon_h}^{\mathbf{h}_r}}} (\frac{1}{\gamma_{2,4}} \mathbf{z}_4^{(n)})$;
    $\mathbf{z}_5^{(n+1)} \leftarrow \mathbf{z}_5^{(n)} - \gamma_{2,5} \mathrm{prox}_{\iota_{B_{2,\varepsilon_l}^{\mathbf{l}_r}}} (\frac{1}{\gamma_{2,5}} \mathbf{z}_5^{(n)})$;
    $\mathbf{z}_6^{(n+1)} \leftarrow \mathbf{z}_6^{(n)} - \gamma_{2,6} \mathrm{prox}_{\iota_{B_{2,\varepsilon_l}^{\mathbf{l}_t}}} (\frac{1}{\gamma_{2,6}} \mathbf{z}_6^{(n)})$;
    $n \leftarrow n + 1$;
end while
```

Using constraints instead of adding terms to the objective function in this way simplifies the parameter setting [37], [38], [39], [40], [41], [42]: we can determine the appropriate parameters for each constraint independently because they are decoupled. The detailed setting of these parameters is discussed in Section IV-C.

C. Optimization

For solving (16) by an algorithm based on P-PDS with OVDP, we need to transform (16) into (10). First, using the hyperslab, the ℓ_p-norm ball, and the indicator function (see (1), (2), and (3) for the definition, respectively), we reformulate our problem in (16) as follows:

$$\begin{aligned}
\min_{\substack{\widetilde{\mathbf{h}}_r, \widetilde{\mathbf{h}}_t, \\ \widetilde{\mathbf{s}}_{hr}, \widetilde{\mathbf{s}}_{lr}, \widetilde{\mathbf{s}}_{lt}}} & \left\| \mathbf{D}\widetilde{\mathbf{h}}_r \right\|_{1,2} + \lambda \left\| \mathbf{D}\widetilde{\mathbf{h}}_t \right\|_{1,2} + \iota_{B_{p,\alpha}^{\mathbf{0}}} \left(\mathbf{D}\widetilde{\mathbf{h}}_r - \mathbf{D}\widetilde{\mathbf{h}}_t \right) \\
& + \sum_{b=1}^{B} \iota_{S_{\beta'_b}^{c'_b}} \left(\left[\widetilde{\mathbf{h}}_t \right]_b \right) + \iota_{B_{2,\varepsilon_h}^{\mathbf{h}_r}} \left(\widetilde{\mathbf{h}}_r + \widetilde{\mathbf{s}}_{hr} \right) \\
& + \iota_{B_{2,\varepsilon_l}^{\mathbf{l}_r}} \left(\mathbf{S}\mathbf{B}\widetilde{\mathbf{h}}_r + \widetilde{\mathbf{s}}_{lr} \right) + \iota_{B_{2,\varepsilon_l}^{\mathbf{l}_t}} \left(\mathbf{S}\mathbf{B}\widetilde{\mathbf{h}}_t + \widetilde{\mathbf{s}}_{lt} \right) \\
& + \iota_{B_{1,\eta_h}^{\mathbf{0}}} \left(\widetilde{\mathbf{s}}_{hr} \right) + \iota_{B_{1,\eta_l}^{\mathbf{0}}} \left(\widetilde{\mathbf{s}}_{lr} \right) + \iota_{B_{1,\eta_l}^{\mathbf{0}}} \left(\widetilde{\mathbf{s}}_{lt} \right) \qquad (17)
\end{aligned}$$

where $\beta'_b = \beta_b N_h$ and $c'_b = c_b N_h$ for $b = 1, \dots, B$. Introducing auxiliary variables $\mathbf{z}_1$, $\mathbf{z}_2$, $\mathbf{z}_3$, $\mathbf{z}_4$, $\mathbf{z}_5$, and $\mathbf{z}_6$, we can transform (17) into the following equivalent problem:

$$\begin{aligned}
\min_{\substack{\widetilde{\mathbf{h}}_r, \widetilde{\mathbf{h}}_t, \\ \widetilde{\mathbf{s}}_{hr}, \widetilde{\mathbf{s}}_{lr}, \widetilde{\mathbf{s}}_{lt}}} & \|\mathbf{z}_1\|_{1,2} + \lambda \|\mathbf{z}_2\|_{1,2} + \iota_{B_{p,\alpha}^{\mathbf{0}}} (\mathbf{z}_3) + \sum_{b=1}^{B} \iota_{S_{\beta'_b}^{c'_b}} \left(\left[\widetilde{\mathbf{h}}_t \right]_b \right) \\
& + \iota_{B_{2,\varepsilon_h}^{\mathbf{h}_r}} (\mathbf{z}_4) + \iota_{B_{2,\varepsilon_l}^{\mathbf{l}_r}} (\mathbf{z}_5) + \iota_{B_{2,\varepsilon_l}^{\mathbf{l}_t}} (\mathbf{z}_6) \\
& + \iota_{B_{1,\eta_h}^{\mathbf{0}}} \left(\widetilde{\mathbf{s}}_{hr} \right) + \iota_{B_{1,\eta_l}^{\mathbf{0}}} \left(\widetilde{\mathbf{s}}_{lr} \right) + \iota_{B_{1,\eta_l}^{\mathbf{0}}} \left(\widetilde{\mathbf{s}}_{lt} \right) \\
\text{s.t.} & \begin{cases} \mathbf{z}_1 = \mathbf{D}\widetilde{\mathbf{h}}_r, \\ \mathbf{z}_2 = \mathbf{D}\widetilde{\mathbf{h}}_t, \\ \mathbf{z}_3 = \mathbf{D}\widetilde{\mathbf{h}}_r - \mathbf{D}\widetilde{\mathbf{h}}_t, \\ \mathbf{z}_4 = \widetilde{\mathbf{h}}_r + \widetilde{\mathbf{s}}_{hr}, \\ \mathbf{z}_5 = \mathbf{S}\mathbf{B}\widetilde{\mathbf{h}}_r + \widetilde{\mathbf{s}}_{lr}, \\ \mathbf{z}_6 = \mathbf{S}\mathbf{B}\widetilde{\mathbf{h}}_t + \widetilde{\mathbf{s}}_{lt}. \end{cases} \qquad (18)
\end{aligned}$$

Then, by defining

$$\begin{aligned}
& g_1 \left(\widetilde{\mathbf{h}}_r \right) = 0, \; g_2 \left(\widetilde{\mathbf{h}}_t \right) = \sum_{b=1}^{B} \iota_{S_{\beta'_b}^{c'_b}} \left(\left[\widetilde{\mathbf{h}}_t \right]_b \right), \\
& g_3 \left(\widetilde{\mathbf{s}}_{hr} \right) = \iota_{B_{1,\eta_h}^{\mathbf{0}}} \left(\widetilde{\mathbf{s}}_{hr} \right), \; g_4 \left(\widetilde{\mathbf{s}}_{lr} \right) = \iota_{B_{1,\eta_l}^{\mathbf{0}}} \left(\widetilde{\mathbf{s}}_{lr} \right), \\
& g_5 \left(\widetilde{\mathbf{s}}_{lt} \right) = \iota_{B_{1,\eta_l}^{\mathbf{0}}} \left(\widetilde{\mathbf{s}}_{lt} \right), \; h_1(\mathbf{z}_1) = \|\mathbf{z}_1\|_{1,2}, \\
& h_2(\mathbf{z}_2) = \lambda \|\mathbf{z}_2\|_{1,2}, \; h_3(\mathbf{z}_3) = \iota_{B_{p,\alpha}^{\mathbf{0}}} (\mathbf{z}_3), \\
& h_4(\mathbf{z}_4) = \iota_{B_{2,\varepsilon_h}^{\mathbf{h}_r}} (\mathbf{z}_4), \; h_5(\mathbf{z}_5) = \iota_{B_{2,\varepsilon_l}^{\mathbf{l}_r}} (\mathbf{z}_5), \\
& h_6(\mathbf{z}_6) = \iota_{B_{2,\varepsilon_l}^{\mathbf{l}_t}} (\mathbf{z}_6) \qquad (19)
\end{aligned}$$

we reduce (10) to (18), that is, (16).

The algorithm for solving (16) is summarized in Algorithm 1. The stepsizes are determined based on OVDP [36] as follows:

$$\begin{aligned}
\gamma_{1,1} &= \frac{1}{2\|\mathbf{D}\|_{\mathrm{op}}^2 + \|\mathbf{I}\|_{\mathrm{op}}^2 + \|\mathbf{S}\mathbf{B}\|_{\mathrm{op}}^2} = \frac{1}{18}, \\
\gamma_{1,2} &= \frac{1}{2\|\mathbf{D}\|_{\mathrm{op}}^2 + \|\mathbf{S}\mathbf{B}\|_{\mathrm{op}}^2} = \frac{1}{17}, \\
\gamma_{1,3} &= \gamma_{1,4} = \gamma_{1,5} = \frac{1}{\|\mathbf{I}\|_{\mathrm{op}}^2} = 1, \\
\gamma_{2,i} &= \frac{1}{5}, \quad \text{for } i = 1, \dots, 6. \qquad (20)
\end{aligned}$$

D. Detailed Computations and Their Complexity

Table I shows the computational complexity (in $\mathcal{O}$-notation) of each operation used in Algorithm 1, where the operated vector is $\mathbf{x} = ([\mathbf{x}]_1^\top, \dots, [\mathbf{x}]_B^\top)^\top \in \mathbb{R}^{NB}$. According to Table I,

TABLE I
COMPUTATIONAL COMPLEXITY OF EACH OPERATION

Operation	$\mathcal{O}$-notation
$\mathbf{Dx}$	$\mathcal{O}(NB)$
$\mathbf{D}^\top\mathbf{x}$	$\mathcal{O}(NB)$
$\mathbf{Sx}$	$\mathcal{O}(NB/k)$, where k is the window size of $\mathbf{S}$
$\mathbf{S}^\top\mathbf{x}$	$\mathcal{O}(NB/k)$, where k is the window size of $\mathbf{S}$
$\mathbf{Bx}$	$\mathcal{O}(kNB)$, where k is the filter size of $\mathbf{B}$
$\mathbf{B}^\top\mathbf{x}$	$\mathcal{O}(kNB)$, where k is the filter size of $\mathbf{B}$
$\mathrm{prox}_{\iota_{S_\alpha^\omega}}([\mathbf{x}]_b)$ in (8)	$\mathcal{O}(N)$
$\mathrm{prox}_{\iota_{B_{1,\eta}^c}}(\mathbf{x})$ in [46]	$\mathcal{O}(NB\log(NB))$
$\mathrm{prox}_{\iota_{B_{2,\varepsilon}^c}}(\mathbf{x})$ in (9)	$\mathcal{O}(NB)$
$\mathrm{prox}_{\gamma\Vert\cdot\Vert_{1,2}}(\mathbf{x})$ in (7)	$\mathcal{O}(NB)$

TABLE II
EXISTING METHODS

	Minimum number of reference dates	Robustness to Gaussian noise	Robustness to sparse noise
STARFM [9]	1	–	–
VIPSTF [33]	1	–	–
RSFN [28]	2	✓	–
RobOt [23]	1	–	–
SwinSTFM [31]	1	–	–
ROSTF	1	✓	✓

TABLE III
PARAMETER SETTING

α	$\begin{cases} 2\times 10^{-3}N_hB\ (p=1), \\ 1\times 10^{-6}N_hB\ (p=2), \end{cases}$
β_b	$\left\lvert \frac{1}{N_l}\mathbf{1}^\top[\mathbf{l}_r]_b - \frac{1}{N_h}\mathbf{1}^\top[\mathbf{h}_r]_b \right\rvert$
c_b	$\frac{1}{N_l}\mathbf{1}^\top[\mathbf{l}_t]_b - r_l(0.5 - \frac{1}{N_l}\mathbf{1}^\top[\mathbf{l}_t]_b)$
ε_h	$0.98\sigma_h\sqrt{N_hB(1-r_h)}$
ε_l	$\Vert\mathbf{l}_r - \mathbf{SBh}_r\Vert_2$

the computational complexity of each step in one iteration of Algorithm 1 is as follows.

1) Steps 4, 5, 23, and 24: $\mathcal{O}(kN_hB)$.
2) Steps 6, 7, 14, 15, 16, 19, 20, 21, 22, 25, 26, 27, and 28: $\mathcal{O}(N_hB)$ when $p = 2$.
3) Step 9: $\mathcal{O}(N_h)$.
4) Steps 17, 18, 29, and 30: $\mathcal{O}(N_lB)$.
5) Steps 11 and 27: $\mathcal{O}(N_hB\log(N_hB))$ when $p = 1$.
6) Steps 12 and 13: $\mathcal{O}(N_lB\log(N_lB))$.

After all, one iteration of the algorithm has an overall computational complexity of $\mathcal{O}(N_hB\log(N_hB))$.

IV. EXPERIMENTS

We demonstrate the effectiveness of our ST fusion method, ROSTF, through comprehensive experiments using simulated and real data for two sites. Our experiments aim to verify the following three items.

1) ROSTF is as effective as state-of-the-art ST fusion methods in noiseless cases and outperforms them in noisy cases. We conducted comparative experiments on four cases of noise contamination. The experimental results for simulated data and real data are presented in Sections IV-D and IV-E, respectively.
2) Key components of ROSTF, such as the assumption-based constraints and the denoising mechanism, operate as expected. To measure their influence, ROSTF without each key component is compared with the original ROSTF in terms of fusion accuracy and convergence speed in Section IV-F.
3) ROSTF is practical in terms of computational time. For a fair comparison, only nondeep-learning-based methods are compared to ROSTF in Section IV-G.

Note that there are two options for ROSTF: $p = 1$ or 2, where p corresponds to the choice of the norm in the first constraint in (16), that is, the edge constraint. Hereafter, ROSTF with $p = 1$ and ROSTF with $p = 2$ are referred to as ROSTF-1 and ROSTF-2, respectively.

A. Data Description

We tested our methods both on real data and simulated data. In the case of satellite observations, radiometric and geometric inconsistencies exist between two different image sensors. This means that the fusion capability of each method cannot be accurately evaluated in experiments using real data because these inconsistencies affect performance, as also addressed in [32]. Therefore, we generated simulated data based on the observation models and verified the performance of each method using this data in addition to the real data. Specifically, in experiments using simulated data, the simulated LR images were generated from the corresponding real HR images according to (14) with $\mathbf{m} = \mathbf{0}$ while the real HR images were used as HR images.

We used MODIS and Landsat time-series images for the following two sites in our experiments.

Site1: The first site is situated in the Daxing district in the south of Beijing city (39.0009° N, 115.0986° E) [16]. For Site1, we employed MODIS and Landsat time-series images acquired on May 29, 2019 (a reference date) and June 14, 2019 (a target date).

Site2: The second site is located in southern New South Wales, Australia (34.0034° S, 145.0675° E) [51]. For Site2, MODIS and Landsat time-series images acquired on January 4, 2002 (a reference date) and February 12, 2002 (a target date) were used.

B. Compared Methods

Our method was compared with STARFM [9], VIPSTF [33], RSFN [28], RobOt [23], and SwinSTFM [31]. Table II shows the characteristics of these methods and ROSTF. Note that, unlike the other methods, RSFN requires input images obtained on two reference dates. Since our experiments assume a scenario with only one reference date, the same HR-LR image pair was input as two different reference image pairs for RSFN.

TABLE IV

RMSE, SAM, MSSIM, and CC Results in the Experiments With Simulated Data

Site	Noise	Metrics	STARFM [9]	VIPSTF [33]	RSFN [28]	RobOt [23]	SwinSTFM [31]	**ROSTF-1 (Ours)**	**ROSTF-2 (Ours)**
Site1	Case1	RMSE	0.0235	**0.0234**	0.0541	0.0246	0.0289	0.0245	0.0238
		SAM	**0.0711**	0.0742	0.1763	0.0748	0.0970	0.0762	0.0740
		SSIM	**0.9755**	0.9752	0.8610	0.9734	0.9636	0.9734	0.9746
		CC	0.9794	**0.9795**	0.8997	0.9776	0.9698	0.9778	0.9789
	Case2	RMSE	0.0554	0.0510	0.0672	0.0546	0.0391	**0.0298**	0.0299
		SAM	0.2431	0.2254	0.2390	0.2380	0.1453	**0.0966**	0.0968
		SSIM	0.8484	0.8652	0.7882	0.8515	0.9284	**0.9543**	0.9541
		CC	0.8984	0.9098	0.8198	0.8982	0.9490	**0.9667**	0.9666
	Case3	RMSE	0.1402	0.1288	0.0542	0.1390	0.0504	0.0314	**0.0278**
		SAM	0.2133	0.2017	0.1769	0.2167	0.1676	0.1173	**0.0896**
		SSIM	0.5193	0.5578	0.8556	0.5225	0.8767	0.9510	**0.9666**
		CC	0.6134	0.6374	0.8937	0.6167	0.9080	0.9633	**0.9711**
	Case4	RMSE	0.1484	0.1360	0.0682	0.1469	0.0544	0.0385	**0.0370**
		SAM	0.3502	0.3230	0.2453	0.3467	0.1918	0.1297	**0.1210**
		SSIM	0.4826	0.5227	0.7799	0.4836	0.8587	0.9282	**0.9346**
		CC	0.5911	0.6156	0.8130	0.5890	0.8964	0.9435	**0.9483**
Site2	Case1	RMSE	0.0367	**0.0312**	0.0499	0.0341	0.0375	0.0398	0.0428
		SAM	0.1091	**0.1040**	0.1583	0.0989	0.1290	0.1231	0.1323
		SSIM	0.9116	**0.9343**	0.8712	0.9223	0.9057	0.8946	0.8816
		CC	0.9372	**0.9523**	0.9065	0.9441	0.9308	0.9289	0.9182
	Case2	RMSE	0.0628	0.0462	0.0677	0.0527	0.0449	**0.0328**	0.0364
		SAM	0.2943	0.2091	0.2816	0.2365	0.1672	**0.1023**	0.1132
		SSIM	0.7281	0.8332	0.7529	0.7908	0.8566	**0.9240**	0.9100
		CC	0.8408	0.8978	0.7868	0.8741	0.9006	**0.9482**	0.9371
	Case3	RMSE	0.1389	0.0909	0.0520	0.1103	0.0543	**0.0391**	0.0420
		SAM	0.2668	0.2100	0.1767	0.2280	0.2040	**0.1244**	0.1258
		SSIM	0.4037	0.5863	0.8575	0.5019	0.7853	**0.8964**	0.8878
		CC	0.5585	0.7048	0.8981	0.6395	0.8517	**0.9281**	0.9188
	Case4	RMSE	0.1476	0.0969	0.0683	0.1189	0.0569	**0.0367**	0.0376
		SAM	0.4138	0.2970	0.2828	0.3326	0.2219	0.1300	**0.1160**
		SSIM	0.3635	0.5463	0.7494	0.4542	0.7650	0.9065	**0.9068**
		CC	0.5319	0.6767	0.7832	0.6040	0.8374	**0.9335**	0.9302

As the parameters of these existing methods, we used the values recommended in each reference. It should be noted that RSFN and SwinSTFM require significantly more data than our and other existing methods due to training and validation processes. For our experiments, we trained and validated RSFN and SwinSTFM using a different set of images from those used for the tests described in Section IV-A. Specifically, 24 groups from Site1 and two groups from Site2 were used for training, and one group from Site1 and two groups from Site2 were used for validation.

C. Experimental Setup

Our method, ROSTF, is implemented using MATLAB. The source code is available on the GitHub[1] platform. For these experiments, the spatial spread transform matrix $\mathbf{B}$ in (14) was implemented as a blurring operator with an averaging filter of size k. The downsampling matrix $\mathbf{S}$ in (14) was designed to take the top-left pixel value in the $k \times k$ window. The parameter k was set to 20 to account for the difference in spatial resolution between the Landsat and MODIS images. The balancing parameter λ in (16) was set to 1 since $\widetilde{\mathbf{h}}_r$ and $\widetilde{\mathbf{h}}_t$ have the same number of elements and are expected to be piecewise smooth to the same degree. The parameters in the constraints in (16) were set as shown in Table III. We set the maximum number of iterations for Algorithm 1 to 50 000 and the stopping criterion of Algorithm 1 was set to $\|\widetilde{\mathbf{h}}_r^{(n)} - \widetilde{\mathbf{h}}_r^{(n-1)}\|_2/\|\widetilde{\mathbf{h}}_r^{(n-1)}\|_2 < 10^{-5}, \|\widetilde{\mathbf{h}}_t^{(n)} - \widetilde{\mathbf{h}}_t^{(n-1)}\|_2/\|\widetilde{\mathbf{h}}_t^{(n-1)}\|_2 < 10^{-5}$, $\|\mathbf{l}_r - \mathbf{SB}\widetilde{\mathbf{h}}_r^{(n)}\|_2 \leq \varepsilon_l$, and $\|\mathbf{l}_t - \mathbf{SB}\widetilde{\mathbf{h}}_t^{(n)}\|_2 \leq \varepsilon_l$.

To verify both the pure fusion capability and the robustness against the noise of the existing methods and ours, we conducted the following four combinations of Gaussian noise with different standard deviations and sparse noise (salt-and-pepper noise) with different rates.

Case1: The observed HR and LR images are noiseless, that is, $\sigma_h = \sigma_l = r_h = r_l = 0$ in (13).

Case2: The observed HR images are contaminated with Gaussian noise with a standard deviation $\sigma_h = 0.05$ while the observed LR images are noiseless, that is, $\sigma_h = 0.05, \sigma_l = r_h = r_l = 0$ in (13).

Case3: The observed HR images are contaminated with sparse noise with a superimposition rate $r_h = 0.05$, while the observed LR images are noiseless, that is, $r_h = 0.05, \sigma_h = \sigma_l = r_l = 0$ in (13).

Case4: The observed HR images are contaminated with Gaussian noise with a standard deviation $\sigma_h = 0.05$ and sparse noise with a superimposition rate $r_h = 0.05$,

[1] https://github.com/MDI-TokyoTech/ROSTF

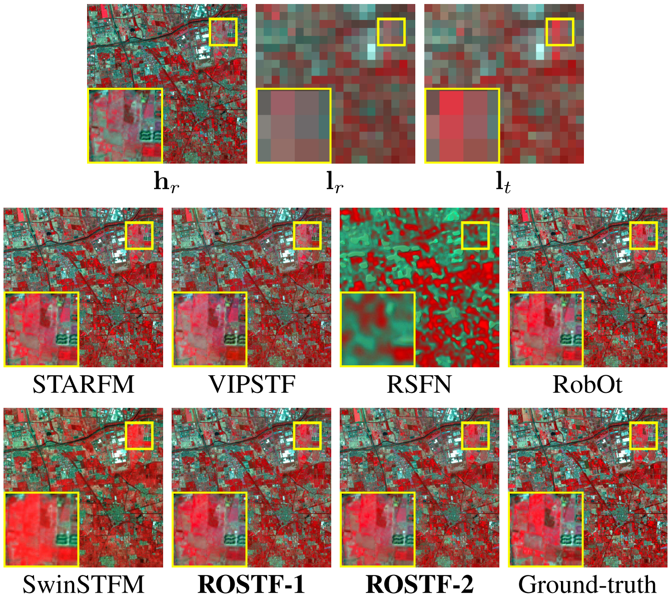

Fig. 2. ST fusion results for the Site1 simulated data in Case1.

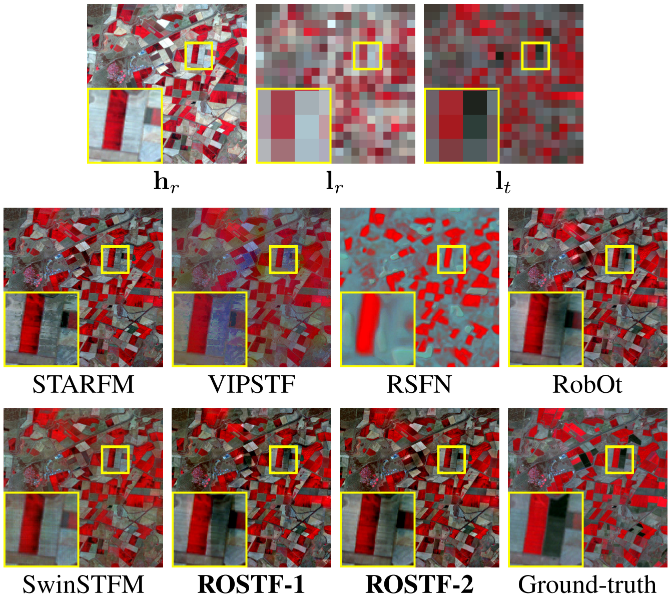

Fig. 3. ST fusion results for the Site2 simulated data in Case1.

while the observed LR images are noiseless, that is, $\sigma_h = r_h = 0.05, \sigma_l = r_l = 0$ in (13).

For the quantitative evaluation, we used the following four metrics: the root mean square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_h B}\left\|\widetilde{\mathbf{h}}_t - \widehat{\mathbf{h}}_t\right\|_2^2} \tag{21}$$

where $\widetilde{\mathbf{h}}_t$ and $\widehat{\mathbf{h}}_t$ denote an estimated HR image and a ground-truth HR image, respectively, on the target date; the spectral angle mapper (SAM) [52]

$$\mathrm{SAM} = \frac{1}{N_h}\sum_{n=1}^{N_h} \arccos\left(\frac{<\widetilde{\mathbf{e}}_n, \widehat{\mathbf{e}}_n>}{\|\widetilde{\mathbf{e}}_n\|_2 \cdot \|\widehat{\mathbf{e}}_n\|_2}\right) \tag{22}$$

where $\widetilde{\mathbf{e}}_n \in \mathbb{R}^B$ and $\widehat{\mathbf{e}}_n \in \mathbb{R}^B$ represent the spectral vectors of the nth pixel of $\widetilde{\mathbf{h}}_t$ and $\widehat{\mathbf{h}}_t$, respectively; the mean structural similarity overall bands (MSSIM) [53]

$$\mathrm{MSSIM} = \frac{1}{B}\sum_{b=1}^{B} \mathrm{SSIM}\left(\left[\widetilde{\mathbf{h}}_t\right]_b, \left[\widehat{\mathbf{h}}_t\right]_b\right) \tag{23}$$

and the correlation coefficient (CC)

$$\mathrm{CC} = \frac{s_{\widetilde{\mathbf{h}}_t \widehat{\mathbf{h}}_t}}{s_{\widetilde{\mathbf{h}}_t} s_{\widehat{\mathbf{h}}_t}} \tag{24}$$

where $s_{\widetilde{\mathbf{h}}_t \widehat{\mathbf{h}}_t}$ denotes the covariance of $\widetilde{\mathbf{h}}_t$ and $\widehat{\mathbf{h}}_t$, and $s_{\widetilde{\mathbf{h}}_t}$ and $s_{\widehat{\mathbf{h}}_t}$ denote the standard deviations of $\widetilde{\mathbf{h}}_t$ and $\widehat{\mathbf{h}}_t$, respectively. RMSE was calculated to measure the difference between the estimated image and the ground-truth at the pixel level. SAM was used to measure spectral fidelity. Lower RMSE and SAM values indicate better estimation performance. We used MSSIM to evaluate the similarity of the overall structure. CC shows the strength of the linear relationship between the estimated image and the ground truth. Higher MSSIM and CC values indicate better estimation performance.

D. Experimental Results With Simulated Data

Table IV shows the RMSE, SAM, MSSIM, and CC results in experiments with the simulated data. In Case1, STARFM, VIPSTF, RobOt, SwinSTFM, ROSTF-1, and ROSTF-2 perform equally well. In contrast, the results of RSFN are not good for both the Site1 and Site2 data. This may be because the size of the training data described above was insufficient to effectively train RSFN. Thus, if more training data were used, RSFN might have produced better results. However, collecting noise-free training data is challenging in real-world applications, and this is the scenario considered in these experiments. Next, we focus on the results in Case2, Case3, and Case4, where the observed reference images are contaminated with noise. While STARFM, VIPSTF, RobOt, SwinSTFM, and RSFN demonstrate significantly worse performance due to the influence of noise, ROSTF-1 and ROSTF-2 show no significant performance degradation in Case2, Case3, and Case4. This is because ROSTF estimates the target HR image while simultaneously denoising the reference HR image.

Figs. 2 and 3 show the estimated results in Case1 for the Site1 and Site2 simulated data, respectively. In the zoomed-in areas, there are significant temporal changes in brightness between the reference HR image $\mathbf{h}_r$ and the target HR image, ground truth. For the Site1 data, it is visually apparent that ROSTF-1 and ROSTF-2 capture these changes most accurately and estimate the brightness closest to that of the ground truth compared with the other methods. This is thanks to the brightness constraint and the fifth constraint in (16), which promote the brightness of the estimated image to be close to that of the observed LR image $\mathbf{l}_t$ on the target date, based on Assumption 2) and the observation model in (14). Following ROSTF, RobOt also captures the temporal changes effectively. STARFM and VIPSTF are significantly influenced by the reference HR image $\mathbf{h}_r$, resulting in estimates that are closer in brightness to that of $\mathbf{h}_r$ rather than to that of the ground truth. The result of RSFN is not good due to a lack of

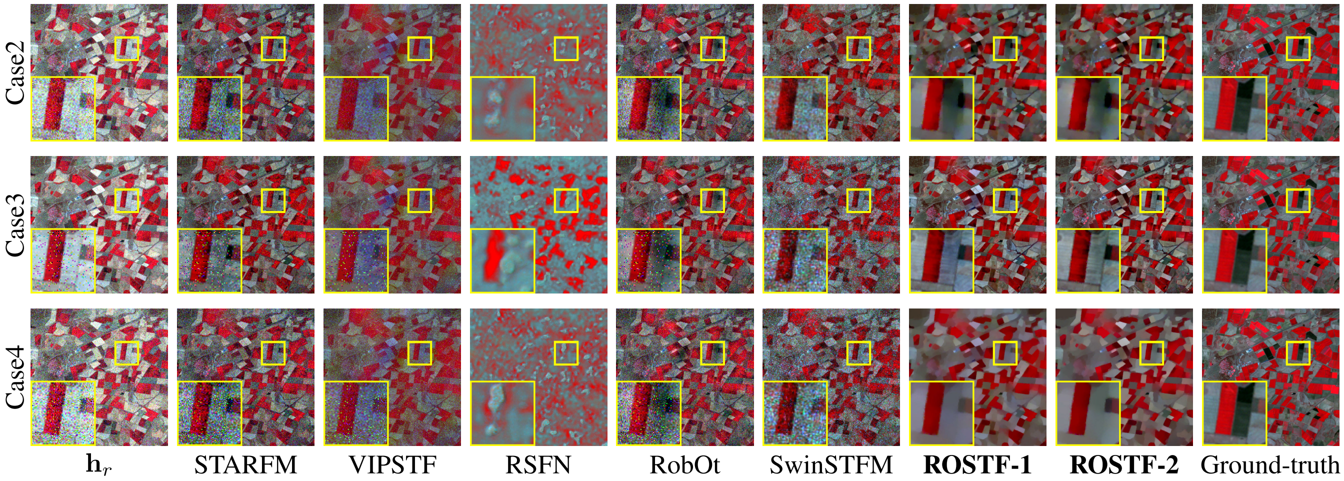

Fig. 4. ST fusion results for the noisy Site2 simulated data. (Top, Middle, and Bottom) Results in Case2, Case3, and Case4, respectively.

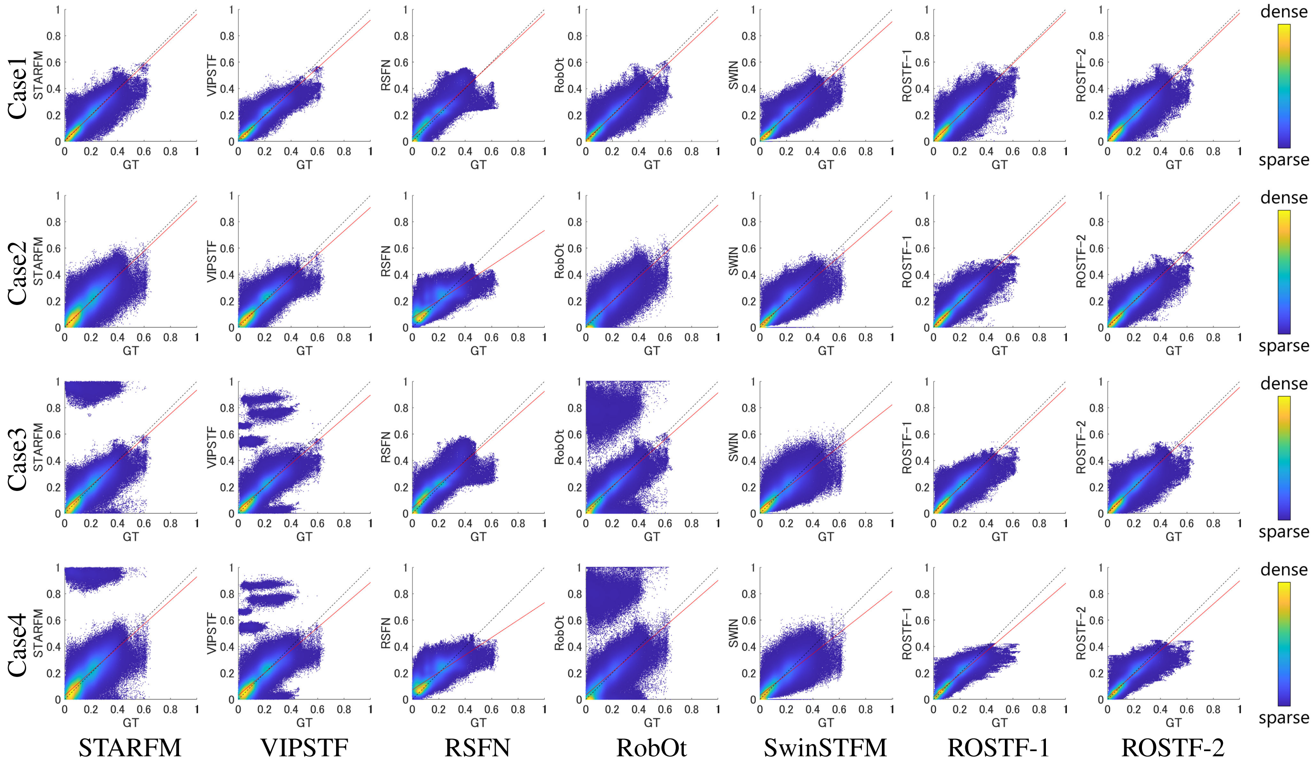

Fig. 5. Scatter plots of the ground truth and the estimated values for the Site2 simulated data.

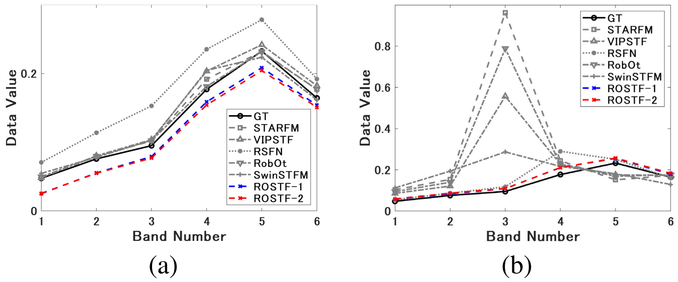

Fig. 6. Spectral profiles of a specific pixel in the results of each method for the Site2 data in (a) Case1 and (b) Cas 4.

training data. The SwinSTFM result appears to show spectral distortion. For the Site2 data, Fig. 3 illustrates that ROSTF-1 most accurately captures the changes in brightness within the zoomed-in area. ROSTF-2 and RobOt have good estimates, followed by ROSTF-1. STARFM and SwinSTFM do not capture the large temporal changes. The result of VIPSTF has different tints throughout. RSFN estimates brightness closer to that of $\mathbf{h}_r$ than to that of the ground truth.

Next, we focus on the results in noisy cases. Fig. 4 shows the estimated results for the Site2 data in noisy cases, that is, Case2, Case3, and Case4. The results of STARFM, VIPSTF, and RobOt are contaminated with a noise similar to that of $\mathbf{h}_r$ because they estimate the target image pixel-wise based solely on the noisy reference image $\mathbf{h}_r$. The results of RSFN are corrupted for the unexpected inputs because RSFN was not trained on noisy data. Furthermore, using noisy data to train RSFN may not have produced good results. This is because RSFN is robust to noise only in situations where

TABLE V

RMSE, SAM, MSSIM, AND CC RESULTS IN THE EXPERIMENTS WITH REAL DATA

Site	Noise	Metrics	STARFM [9]	VIPSTF [33]	RSFN [28]	RobOt [23]	SWIN [31]	**ROSTF-1 (Ours)**	**ROSTF-2 (Ours)**
Site1	Case1	RMSE	**0.0263**	0.0275	0.0523	0.0281	0.0297	0.0278	0.0266
		SAM	0.0838	0.0897	0.1601	0.0906	0.0987	0.0937	**0.0829**
		SSIM	0.9701	0.9666	0.8700	0.9667	0.9625	0.9652	**0.9704**
		CC	**0.9742**	0.9719	0.9029	0.9710	0.9680	0.9711	0.9735
	Case2	RMSE	0.0565	0.0439	0.0692	0.0562	0.0398	0.0363	**0.0333**
		SAM	0.2466	0.1826	0.2455	0.2436	0.1471	0.1277	**0.1119**
		SSIM	0.8448	0.9024	0.7957	0.8465	0.9271	0.9347	**0.9463**
		CC	0.8942	0.9282	0.8089	0.8927	0.9465	0.9501	**0.9582**
	Case3	RMSE	0.1409	0.0991	0.0539	0.1402	0.0512	0.0362	**0.0300**
		SAM	0.2234	0.1955	0.1696	0.2301	0.1699	0.1434	**0.0998**
		SSIM	0.5165	0.6604	0.8679	0.5181	0.8742	0.9289	**0.9608**
		CC	0.6095	0.7278	0.8998	0.6120	0.9039	0.9515	**0.9664**
	Case4	RMSE	0.1490	0.1046	0.0700	0.1481	0.0551	0.0447	**0.0382**
		SAM	0.3533	0.2706	0.2493	0.3521	0.1938	0.1608	**0.1252**
		SSIM	0.4804	0.6321	0.7849	0.4800	0.8563	0.8988	**0.9318**
		CC	0.5875	0.7085	0.8012	0.5848	0.8919	0.9238	**0.9448**
Site2	Case1	RMSE	0.0458	**0.0371**	0.0503	0.0428	0.0388	0.0483	0.0487
		SAM	0.1406	**0.1283**	0.1585	0.1310	0.1285	0.1478	0.1345
		SSIM	0.8762	**0.9136**	0.8713	0.8911	0.9041	0.8707	0.8725
		CC	0.9051	**0.9326**	0.9064	0.9127	0.9267	0.8950	0.8957
	Case2	RMSE	0.0677	0.0485	0.0742	0.0588	0.0466	0.0470	**0.0463**
		SAM	0.3067	0.2146	0.3121	0.2514	0.1668	0.1649	**0.1355**
		SSIM	0.7047	0.8229	0.7184	0.7641	0.8527	0.8684	**0.8821**
		CC	0.8191	0.8856	0.7425	0.8460	0.8944	0.8939	**0.8996**
	Case3	RMSE	0.1419	0.0914	0.0539	0.1140	0.0568	**0.0490**	0.0494
		SAM	0.2906	0.2248	0.1842	0.2533	0.2097	0.1558	**0.1380**
		SSIM	0.3891	0.5789	0.8513	0.4821	0.7730	0.8618	**0.8690**
		CC	0.5448	0.6911	0.8922	0.6180	0.8411	0.8920	**0.8929**
	Case4	RMSE	0.1500	0.0963	0.0755	0.1222	0.0595	0.0509	**0.0484**
		SAM	0.4229	0.2972	0.3177	0.3461	0.2266	0.2072	**0.1546**
		SSIM	0.3526	0.5439	0.7096	0.4374	0.7524	0.8291	**0.8724**
		CC	0.5214	0.6686	0.7338	0.5844	0.8252	0.8725	**0.8863**

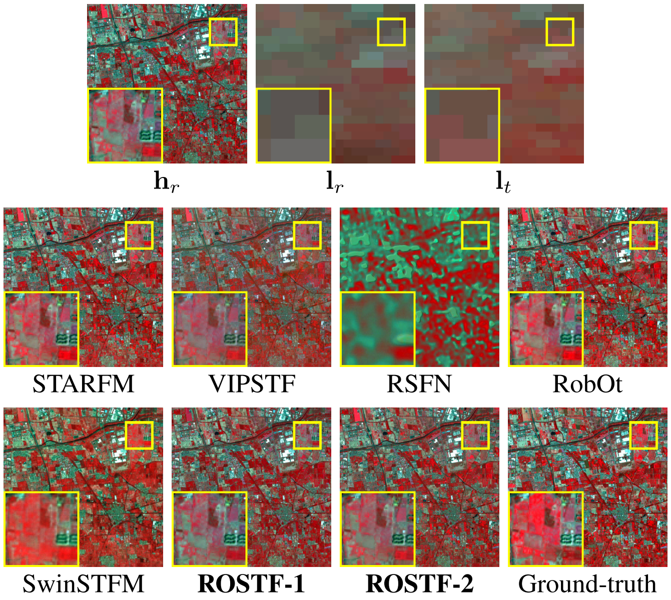

Fig. 7. ST fusion results for the Site1 real data in Case1.

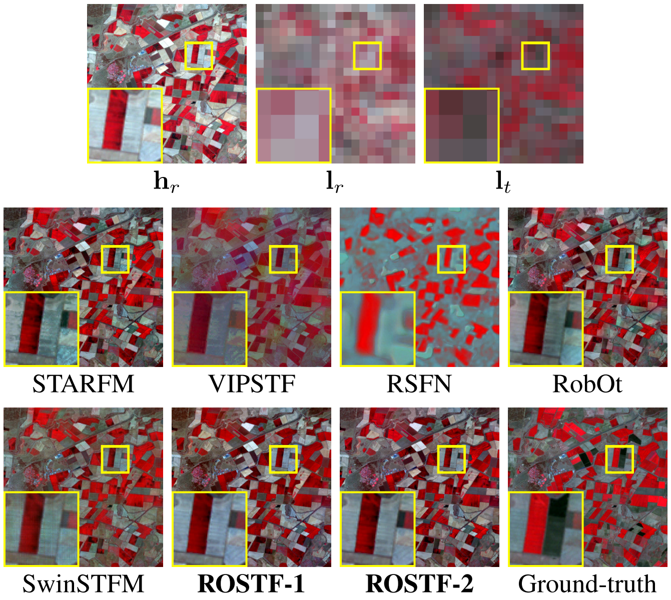

Fig. 8. ST fusion results for the Site2 real data in Case1.

noisy pixels in the two reference HR images do not appear at the same location. SwinSTFM generates unstable noisy results due to noisy unexpected inputs. On the other hand, ROSTF-1 and ROSTF-2 provide good estimates even when the observed reference images are noisy, without much loss of accuracy compared to the noiseless case. This is thanks

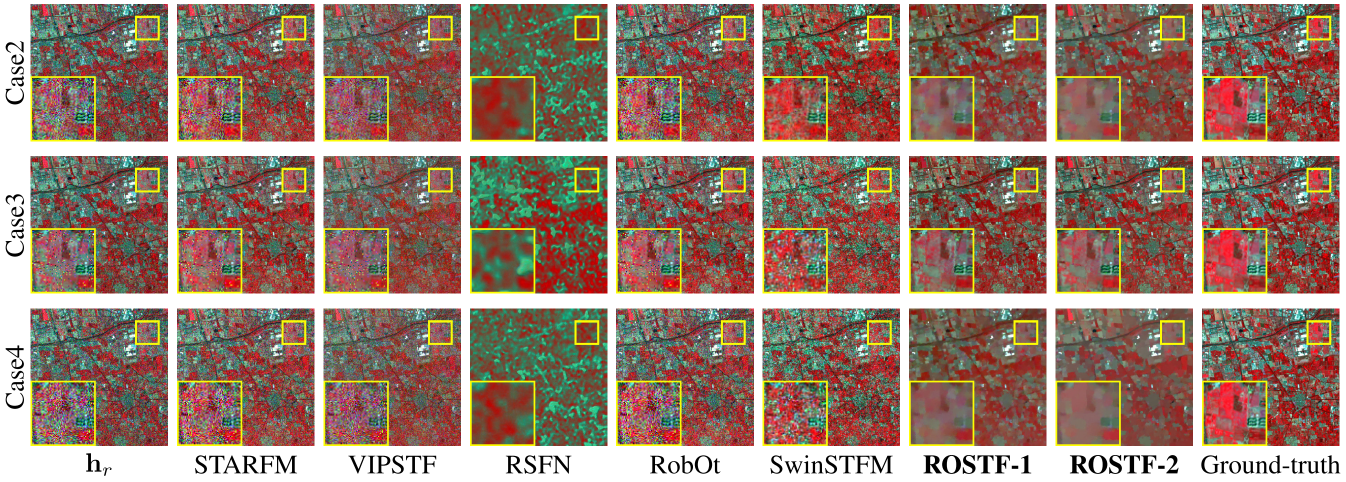

Fig. 9. ST fusion results for the noisy Site1 simulated data. (Top, Middle, and Bottom) Results in Case2, Case3, and Case4, respectively.

to the framework that simultaneously performs noise removal and ST fusion.

The impact of noise on each method is also visually evident in Figs. 5 and 6. The scatter plots in Fig. 5 reveal the difference between the ground-truth and the estimated values of each method for the simulated Site2 data. In Case2, STARFM, VIPSTF, RSFN, RobOt, and SwinSTFM exhibit greater variance compared to Case1 due to the influence of Gaussian noise. In Case3, STARFM, VIPSTF, and RobOt estimate the wrong values close to 0 or 1 affected by sparse noise while the RSFN and SwinSTFM results have no such outliers but show greater variance. Furthermore, in Case4, the distributions indicate that STARFM, VIPSTF, RSFN, RobOt, and SwinSTFM are affected by both Gaussian and sparse noise. In contrast, the results of ROSTF-1 and ROSTF-2 have minimal variance and no outliers, indicating their robustness to Gaussian, sparse, and mixed noise. Spectral profiles of a specific pixel in the results of each method for the Site2 data in Case1 and Case4 are depicted in Fig. 6. STARFM, VIPSTF, RobOt, SwinSTFM, and ROSTF estimate similarly accurate values in Case1, that is, they are comparable in the noiseless case. In Case4, STARFM, VIPSTF, RobOt, and SwinSTFM estimate completely wrong values for the third band affected by sparse noise and perform worse for the other bands due to the influence of Gaussian noise. In contrast, ROSTF-1 and ROSTF-2 have accurate estimates for all bands, even in the noisy case.

E. Experimental Results With Real Data

Table V shows the RMSE, SAM, MSSIM, and CC results in experiments with the real data. Compared to the results for the simulated data in Table IV, the performance of ROSTF-1 and ROSTF-2 degrades due to radiometric and geometric inconsistencies between the Landsat and MODIS sensors. Despite the performance degradation, ROSTF-1 and ROSTF-2 perform as well as the existing methods in the noiseless case, that is, Case1, and outperform them in the noisy cases, that is, Case2, Case3, and Case4, as in the experiments with the real data. Thus, it can be concluded that ROSTF is robust to noise even for the simulated data.

Figs. 7 and 8 show the estimated results in Case1 for the Site1 and Site2 real data, respectively. Compared to the results for the simulated data in Figs. 2 and 3, ROSTF-1 and ROSTF-2 perform worse due to modeling errors between the real HR images $\mathbf{h}_r, \mathbf{h}_t$ and the real LR images $\mathbf{l}_r, \mathbf{l}_t$ caused by radiometric and geometric inconsistencies.

Fig. 9 shows the estimated results for the Site2 data in the noisy cases, that is, Case2, Case3, and Case4. The results of the existing methods are not good, especially STARFM, VIPSTF, and RobOt generate noisy outputs. The estimated images of ROSTF seem to be blurred in the zoomed area in Fig. 9. This is due to oversmoothing by the HTV regularization terms in (16), which might have undesirable effects on some applications. Nevertheless, according to the difference map in Fig. 10, the results of ROSTF exhibit the least error, and the accuracy evaluation in Table V also shows that ROSTF achieves the best performance in all metrics.

F. Ablation Study

We conducted ablation experiments focusing on the following three items.

1) The edge constraint, $\|\mathbf{D}\widetilde{\mathbf{h}}_r - \mathbf{D}\widetilde{\mathbf{h}}_t\|_p \leq \alpha$, to encourage similarity in the land structure, specifically the edges, between the reference and target HR images based on Assumption 1).
2) The brightness constraint, $|c_b - \mathbf{1}^\top[\widetilde{\mathbf{h}}_t]_b/N_h| \leq \beta_b$ $(b = 1, \ldots, B)$, which is designed based on Assumption 2) to ensure that the estimated target HR image has a similar average brightness to the target LR image.
3) The denoising mechanism for the reference HR image $\mathbf{h}_r$, that is, the first regularization term $\|\mathbf{D}\widetilde{\mathbf{h}}_r\|_{1,2}$ and the third, fourth, sixth, and seventh constraints, $\|\mathbf{h}_r - (\widetilde{\mathbf{h}}_r + \widetilde{\mathbf{s}}_{hr})\|_2 \leq \varepsilon_h$, $\|\mathbf{l}_r - (\mathbf{SB}\widetilde{\mathbf{h}}_r + \widetilde{\mathbf{s}}_{lr})\|_2 \leq \varepsilon_l$, $\|\widetilde{\mathbf{s}}_{hr}\|_1 \leq \eta_h$, $\|\widetilde{\mathbf{s}}_{lr}\|_1 \leq \eta_l$.

We tested ROSTF with each of the three components mentioned above removed. In the following, we present the ablation studies on the two constraints, the edge constraint and the brightness constraint, followed by the ablation study on the denoising mechanism. The hyperparameters in each optimiza-

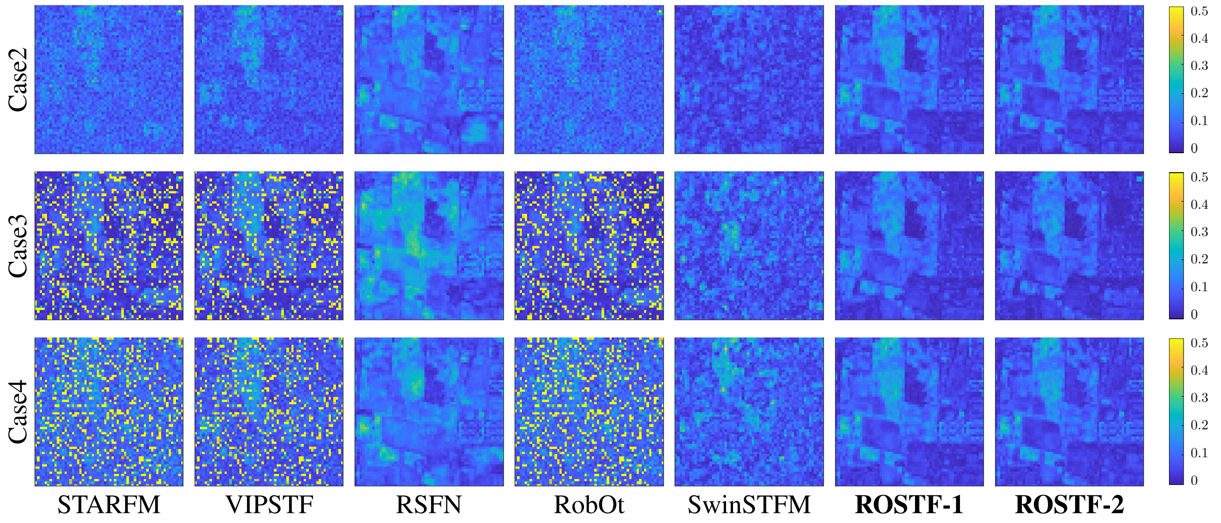

Fig. 10. Difference map (absolute errors) of the fusion results in the zoomed-in area in the noisy Site1 real data. (Top, Middle, and Bottom) Results in Case2, Case3, and Case4, respectively.

TABLE VI
Average Number of Iterations and the Average Performance Results for All the Situations

	ROSTF Without		
	The Edge Constraint	The Brightness Constraint	The Original ROSTF
iter	47788	28217	**27208**
RMSE	0.0578	0.0581	**0.0381**
SAM	0.2071	0.1511	**0.1195**
SSIM	0.8393	0.8780	**0.9151**
CC	0.8484	0.9314	**0.9341**

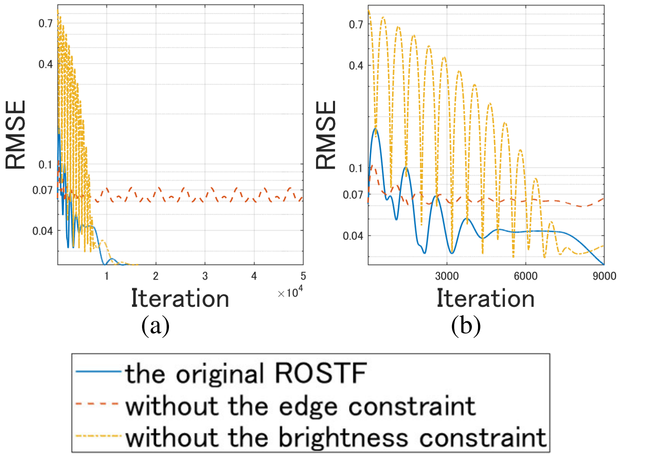

Fig. 11. Behavior of the original ROSTF-2, ROSTF-2 without the edge constraint, and ROSTF-2 without the brightness constraint in the experiment on the Site1 simulated data in Case1. The transition of the RMSE values (a) until the algorithms stopped and (b) in early iterations.

tion problem and the stopping criterion of each P-PDS-based algorithm to solve it were set as in the original ROSTF. On the other hand, the stepsizes in each algorithm were set as the values computed according to the operator-norm-based design method of variable-wise diagonal preconditioning in (11).

1) Edge and Brightness Constraints: First, we measure the effectiveness of the two constraints based on Assumptions 1) and 2), that is, the edge constraint and the brightness constraint, in terms of convergence speed and fusion performance.

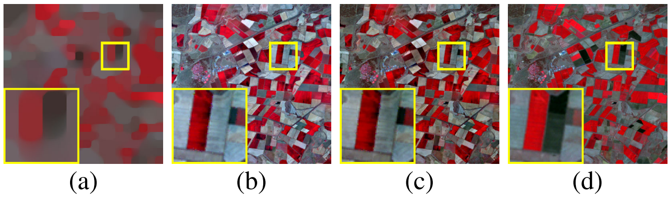

Fig. 12. ST fusion results of (a) ROSTF-2 without edge constraint, (b) ROSTF-2 without brightness constraint, and (c) original ROSTF-2 for the Site2 simulated data in Case3. (d) Ground truth.

Table VI shows the average number of iterations spent before each algorithm stopped and the average performance results for all sites (Site1 or Site2), data types (real or simulated data), and noise cases (from Case1 to Case4). Note that each algorithm always stopped at the maximum number of iterations, 50 000, even if the stopping criterion was not met. The original ROSTF performs best with the fewest number of iterations on average, indicating that both of the two constraints contribute to achieving higher fusion performance with fewer iterations.

Fig. 11 illustrates the transition of the RMSE values for the original ROSTF-2, ROSTF-2 without the edge constraint, and ROSTF-2 without the brightness constraint in the experiment on the Site1 simulated data in Case1. ROSTF-2 without the edge constraint does not meet the stopping criterion until 50 000 iterations, possibly because the solution space of the optimization problem without the edge constraint is too large to efficiently reach an optimal solution. This suggests that the edge constraint has the effect of making the solution space moderately small and speeding up convergence. On the other hand, ROSTF-2 without the brightness constraint converges faster than ROSTF-2 without the edge constraint but exhibits unstable behavior, especially in the early iterations. This may be because the variable $\widetilde{\mathbf{h}}_t$ is no longer directly updated as a primal variable when the brightness constraint is removed. The update equation corresponding to the brightness constraint

TABLE VII
AVERAGE PERFORMANCE RESULTS FOR EACH NOISE CASE IN THE ABLATION STUDY OF THE DENOISING MECHANISM

Noise	Metrics	Without denoising	ROSTF
Case1	RMSE	**0.0355**	0.0363
	SAM	**0.1049**	0.1101
	SSIM	**0.9244**	0.9212
	CC	**0.9404**	0.9384
Case2	RMSE	0.0614	**0.0382**
	SAM	0.2702	**0.1218**
	SSIM	0.7772	**0.9154**
	CC	0.8596	**0.9356**
Case3	RMSE	0.1391	**0.0370**
	SAM	0.2891	**0.1137**
	SSIM	0.4597	**0.9192**
	CC	0.5976	**0.9364**
Case4	RMSE	0.1471	**0.0408**
	SAM	0.4140	**0.1324**
	SSIM	0.4238	**0.9046**
	CC	0.5767	**0.9259**

is implemented as the only update equation for the primal variable $\widetilde{\mathbf{h}}_t$, as in the lines 8–10 of Algorithm 1, while the update equations corresponding to the other components for $\widetilde{\mathbf{h}}_t$ update the dual variables $\mathbf{z}_2$, $\mathbf{z}_3$, and $\mathbf{z}_6$. Directly updating the primal variable leads to faster convergence of the algorithm and more stable behavior than updating the primal variable via the update of the dual variables. Introducing the brightness constraint allows the variable $\widetilde{\mathbf{h}}_t$ to be updated directly as the primal variable, thereby enhancing convergence speed and stability.

Fig. 12 provides a visual comparison of the original ROSTF-2 and ROSTF-2 without these constraints. The image estimated by ROSTF-2 without the edge constraint (a) loses spatial structure, and ROSTF-2 without the brightness constraint and (b) produces an image with incorrect brightness. On the other hand, the original ROSTF-2 (c) estimates brightness close to that of the ground truth while still preserving spatial structure, indicating that both constraints work effectively.

2) Denoising Mechanism: Next, we move on to the ablation study of the denoising mechanism of ROSTF. The optimization problem of ROSTF without the denoising mechanism is formulated as

$$\min_{\widetilde{\mathbf{h}}_t, \widetilde{\mathbf{s}}_{lt}} \|\mathbf{D}\widetilde{\mathbf{h}}_t\|_{1,2}$$
$$\text{s.t.} \begin{cases} \|\mathbf{D}\mathbf{h}_r - \mathbf{D}\widetilde{\mathbf{h}}_t\|_p \leq \alpha, \\ |c_b - \mathbf{1}^\top \left[\widetilde{\mathbf{h}}_t\right]_b / N_h| \leq \beta_b \, (b = 1, \ldots, B), \\ \|\mathbf{l}_t - \left(\mathbf{S}\mathbf{B}\widetilde{\mathbf{h}}_t + \widetilde{\mathbf{s}}_{lt}\right)\|_2 \leq \varepsilon_l, \\ \|\widetilde{\mathbf{s}}_{lt}\|_1 \leq \eta_l. \end{cases} \tag{25}$$

Since the objective function contains only one regularization term for $\widetilde{\mathbf{h}}_t$, no balancing parameter is needed. Also, note that the edge constraint of (25) is different from that of the original ROSTF in (16) because this optimization problem has no variable $\widetilde{\mathbf{h}}_r$.

Table VII displays the average RMSE, SAM, MSSIM, and CC results for each noise case. As expected, in the noisy cases, that is, Case2, Case3, and Case4, ROSTF without the denoising mechanism performs significantly worse than the original ROSTF due to the direct impact of noise. This shows that the denoising mechanism works effectively to make ROSTF robust to noise. On the other hand, we newly found that in the noiseless case, that is, Case1, ROSTF without the denoising mechanism achieves slightly better fusion results than the original one. This result indicates that in noiseless cases, the observed HR image $\mathbf{h}_r$ can be used as the reference as it is, and there is no need to estimate $\widetilde{\mathbf{h}}_r$. Furthermore, by removing the variable $\widetilde{\mathbf{h}}_r$, ROSTF does not need the reference LR image $\mathbf{l}_r$ as input in (25). Thus, fortunately, if noiseless images can be observed, ROSTF can be applied in situations where only two input images are available, a reference HR image $\mathbf{h}_r$ and a target LR image $\mathbf{l}_t$.

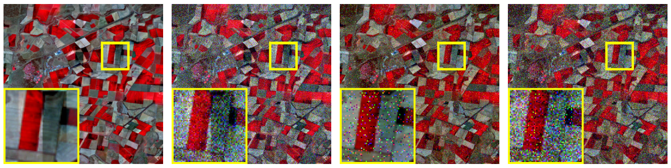
Case1 Case2 Case3 Case4

Fig. 13. ST fusion results of ROSTF-1 without the denoising mechanism for the Site2 simulated data.

Fig. 13 illustrates the fusion results of ROSTF-1 without the denoising mechanism for the Site2 simulated data. It is also visually apparent that in Case1, ROSTF-1 without the denoising mechanism produces a satisfactory result using only two input images. On the other hand, in Case2, Case3, and Case4, the results of ROSTF without the denoising mechanism are contaminated with noise. This is because the edge constraint copies not only the true edge or spatial structure, but also the noise in the reference HR image to the estimated target HR image. The results of this ablation study confirm that the denoising mechanism plays an effective role in avoiding such noise effects.

G. Computational Cost

We measured the actual running times using MATLAB (R2022b) on a Windows 11 computer equipped with an Intel Core i9-13900 1.0 GHz processor, 32 GB of RAM, and NVIDIA GeForce RTX 4090. We used the Site1 and Site2 data with 400 × 400 pixels and six bands. The measurement results show that the average number of iterations for Algorithm 1 was 59.20 per second. Table VIII displays the average running times of the nondeep-learning-based comparison methods (STARFM, VIPSTF, and RobOt) and our methods (ROSTF-1 and ROSTF-2). For ROSTF, the average number of iterations for Algorithm 1 is also given in parentheses. Note that only the nondeep-learning-based methods were compared with ROSTF for a fair comparison because deep-learning-based methods require training processes in addition to ST fusion processes.

ROSTF-1 and ROSTF-2 each took about 4–10 min. STARFM and VIPSTF took longer than ROSTF because they estimated the target HR image pixel by pixel. On the other hand, RobOt ran much faster than ROSTF, possibly because the Least Absolute Shrinkage and Selection Operator (LASSO) problem in RobOt has a closed-form solution in our

TABLE VIII
AVERAGE RUNNING TIME [S]

	Site1	Site2
STARFM	1.841×10^3	6.237×10^2
VIPSTF	1.279×10^4	1.301×10^4
RobOt	2.750×10^{-1}	2.740×10^{-1}
ROSTF-1	4.536×10^2 (26048 ite)	5.743×10^2 (33796 ite)
ROSTF-2	2.358×10^2 (13893 ite)	6.003×10^2 (35093 ite)

experiment with only one reference date. This result indicates that ROSTF is slower than RobOt, but we believe that ROSTF remains practical in terms of computational cost.

H. Summary

We summarize the insights from the experiments as follows.

1) From the results of the experiments in Case1, we see that ROSTF is comparable to state-of-the-art ST fusion methods in noiseless cases. Therefore, the observation model in (14) and the assumptions introduced in Section III-B are valid for ST fusion.
2) The results of the experiments in Case2, Case3, and Case4 confirm that ROSTF has good performance even when observed HR images are degraded by random noise, missing values, and outliers.
3) The ablation studies demonstrate that the key components, such as the edge constraint, the brightness constraint, and the denoising mechanism, work effectively as expected.

V. CONCLUSION

We have proposed an optimization-based ST fusion method, named ROSTF, which is robust to mixed Gaussian and sparse noise in observed satellite images. We have formulated the fusion problem as a constrained optimization problem and have developed the optimization algorithm based on P-PDS with OVDP. ROSTF was tested through experiments using both simulated and real data. The experimental results demonstrate that ROSTF achieves performance comparable to state-of-the-art ST fusion methods in noiseless cases and significantly better in noisy cases. ROSTF will have a strong impact on the field of remote sensing, including the estimation of satellite image series with high spatial and temporal resolution from observed image series taken in measurement environments with severe degradation.

REFERENCES

[1] M. Zhang, H. Lin, H. Sun, and Y. Cai, "Estimation of vegetation productivity using a Landsat 8 time series in a heavily urbanized area, Central China," *Remote Sens.*, vol. 11, no. 2, p. 133, Jan. 2019.

[2] K. R. Knipper et al., "Evapotranspiration estimates derived using thermal-based satellite remote sensing and data fusion for irrigation management in California vineyards," *Irrigation Sci.*, vol. 37, no. 3, pp. 431–449, May 2019.

[3] Y. Pan, F. Shen, and X. Wei, "Fusion of Landsat-8/OLI and GOCI data for hourly mapping of suspended particulate matter at high spatial resolution: A case study in the Yangtze (Changjiang) estuary," *Remote Sens.*, vol. 10, no. 2, p. 158, Jan. 2018.

[4] X. Yang and C. P. Lo, "Using a time series of satellite imagery to detect land use and land cover changes in the Atlanta, Georgia metropolitan area," *Int. J. Remote Sens.*, vol. 23, no. 9, pp. 1775–1798, Jan. 2002.

[5] V. Heimhuber, M. G. Tulbure, and M. Broich, "Addressing spatio-temporal resolution constraints in Landsat and MODIS-based mapping of large-scale floodplain inundation dynamics," *Remote Sens. Environ.*, vol. 211, pp. 307–320, Jun. 2018.

[6] N. Pastick, B. Wylie, and Z. Wu, "Spatiotemporal analysis of Landsat-8 and Sentinel-2 data to support monitoring of dryland ecosystems," *Remote Sens.*, vol. 10, no. 5, p. 791, May 2018.

[7] M. Chiesi et al., "Spatio-temporal fusion of NDVI data for simulating soil water content in heterogeneous Mediterranean areas," *Eur. J. Remote Sens.*, vol. 52, no. 1, pp. 88–95, Jan. 2019.

[8] X. Li, Y. Zhou, G. R. Asrar, J. Mao, X. Li, and W. Li, "Response of vegetation phenology to urbanization in the conterminous United States," *Global Change Biol.*, vol. 23, no. 7, pp. 2818–2830, Jul. 2017.

[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] 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, p. 527, Mar. 2018.

[11] Y. Chen, S. Liu, and X. Wang, "Learning continuous image representation with local implicit image function," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, Jul. 2021, pp. 8628–8638.

[12] X. Hu, H. Mu, X. Zhang, Z. Wang, T. Tan, and J. Sun, "Meta-SR: A magnification-arbitrary network for super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1575–1584.

[13] 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.

[14] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and *k*-means clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, Oct. 2009.

[15] S. Takemoto, K. Naganuma, and S. Ono, "Graph spatio-spectral total variation model for hyperspectral image denoising," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[16] J. Li, Y. Li, L. He, J. Chen, and A. Plaza, "Spatio-temporal fusion for remote sensing data: An overview and new benchmark," *Sci. China Inf. Sci.*, vol. 63, no. 4, pp. 1–17, Apr. 2020.

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

[18] R. Zurita-Milla, J. 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.

[19] 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, Nov. 2010.

[20] J. Xue, Y. Leung, and T. Fung, "A Bayesian data fusion approach to spatio-temporal fusion of remotely sensed images," *Remote Sens.*, vol. 9, no. 12, p. 1310, Dec. 2017.

[21] 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, Aug. 2013.

[22] B. Huang and H. Song, "Spatiotemporal reflectance fusion via sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 10, pp. 3707–3716, Oct. 2012.

[23] S. Chen, J. Wang, and P. Gong, "ROBOT: A spatiotemporal fusion model toward seamless data cube for global remote sensing applications," *Remote Sens. Environ.*, vol. 294, Aug. 2023, Art. no. 113616.

[24] Y. Ke, J. Im, S. Park, and H. Gong, "Downscaling of MODIS one kilometer evapotranspiration using Landsat-8 data and machine learning approaches," *Remote Sens.*, vol. 8, no. 3, p. 215, Mar. 2016.

[25] 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.

[26] V. Moosavi, A. Talebi, M. H. Mokhtari, S. R. F. Shamsi, and Y. Niazi, "A wavelet-artificial intelligence fusion approach (WAIFA) for blending Landsat and MODIS surface temperature," *Remote Sens. Environ.*, vol. 169, pp. 243–254, Nov. 2015.

[27] X. Liu, C. Deng, S. Wang, G. Huang, B. Zhao, and P. Lauren, "Fast and accurate spatiotemporal fusion based upon extreme learning machine," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 2039–2043, Dec. 2016.

[28] Z. Tan, M. Gao, J. Yuan, L. Jiang, and H. Duan, "A robust model for MODIS and Landsat image fusion considering input noise," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5407217.

[29] 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, 2022, Art. no. 5601413.

[30] 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.

[31] G. Chen, P. Jiao, Q. Hu, L. Xiao, and Z. Ye, "SwinSTFM: Remote sensing spatiotemporal fusion using Swin transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5410618.

[32] 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, Jan. 2016.

[33] Q. Wang, Y. Tang, X. Tong, and P. M. Atkinson, "Virtual image pair-based spatio-temporal fusion," *Remote Sens. Environ.*, vol. 249, Nov. 2020, Art. no. 112009.

[34] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: A technical overview," *IEEE Signal Process. Mag.*, vol. 20, no. 3, pp. 21–36, May 2003.

[35] T. Pock and A. Chambolle, "Diagonal preconditioning for first order primal-dual algorithms in convex optimization," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1762–1769.

[36] K. Naganuma and S. Ono, "Variable-wise diagonal preconditioning for primal-dual splitting: Design and applications," *IEEE Trans. Signal Process.*, vol. 71, pp. 3281–3295, 2023.

[37] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 681–695, Mar. 2011.

[38] G. Chierchia, N. Pustelnik, J.-C. Pesquet, and B. Pesquet-Popescu, "Epigraphical projection and proximal tools for solving constrained convex optimization problems," *Signal, Image Video Process.*, vol. 9, no. 8, pp. 1737–1749, Nov. 2015.

[39] S. Ono and I. Yamada, "Signal recovery with certain involved convex data-fidelity constraints," *IEEE Trans. Signal Process.*, vol. 63, no. 22, pp. 6149–6163, Nov. 2015.

[40] S. Ono, "Primal-dual plug-and-play image restoration," *IEEE Signal Process. Lett.*, vol. 24, no. 8, pp. 1108–1112, Aug. 2017.

[41] S. Ono, "L_0 gradient projection," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1554–1564, Jun. 2017.

[42] K. Naganuma and S. Ono, "A general destriping framework for remote sensing images using flatness constraint," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 5525016.

[43] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imag. Vis.*, vol. 40, no. 1, pp. 120–145, May 2011.

[44] R. Isono, K. Naganuma, and S. Ono, "Robust spatiotemporal fusion of satellite images via convex optimization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[45] P. L. Combettes and N. N. Reyes, "Moreau's decomposition in Banach spaces," *Math. Program.*, vol. 139, nos. 1–2, pp. 103–114, 2013.

[46] L. Condat, "Fast projection onto the simplex and the L_1 ball," *Math. Program.*, vol. 158, nos. 1–2, pp. 575–585, 2016.

[47] J. Farrell, F. Xiao, and S. Kavusi, "Resolution and light sensitivity tradeoff with pixel size," *Proc. SPIE*, vol. 6069, pp. 211–218, Feb. 2006.

[48] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, Feb. 2011.

[49] 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.

[50] Q. Yuan, L. Zhang, and H. Shen, "Hyperspectral image denoising employing a spectral–spatial adaptive total variation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 10, pp. 3660–3677, Oct. 2012.

[51] I. V. Emelyanova, T. R. McVicar, T. G. Van Niel, L. T. Li, and A. I. J. M. van Dijk, "Assessing the accuracy of blending Landsat–MODIS surface reflectances in two landscapes with contrasting spatial and temporal dynamics: A framework for algorithm selection," *Remote Sens. Environ.*, vol. 133, pp. 193–209, Jun. 2013.

[52] L. Alparone, L. Wald, J. Chanussot, C. Thomas, P. Gamba, and L. M. Bruce, "Comparison of pansharpening algorithms: Outcome of the 2006 GRS-S data-fusion contest," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3012–3021, Oct. 2007.

[53] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 600–612, 2004.

Ryosuke Isono (Graduate Student Member, IEEE) received the B.E. degree in information and computer sciences from Osaka University, Suita, Japan, in 2022, and the M.E. degree in information and computer sciences from Tokyo Institute of Technology, Yokohama, Japan, in 2024, where he is currently pursuing the Ph.D. degree with the Department of Computer Science.

His research interests include signal and image processing, optimization theory, and remote sensing.

Kazuki Naganuma (Graduate Student Member, IEEE) received the B.E. degree in information and computer sciences from Kanagawa Institute of Technology, Atsugi, Japan, in 2020, and the M.E. degree in information and computer sciences from Tokyo Institute of Technology, Yokohama, Japan, in 2022, where he is currently pursuing the Ph.D. degree with the Department of Computer Science.

Since October 2023, he has been a Research Fellow (DC2) of Japan Society for the Promotion of Science (JSPS) and a Researcher of ACT-X of Japan Science and Technology Agency (JST), Tokyo, Japan. His research interests include signal and image processing and optimization theory.

Mr. Naganuma received the Student Conference Paper Award from the IEEE SPS Japan Chapter in 2023.

Shunsuke Ono (Senior Member, IEEE) received the B.E. degree in computer science and the M.E. and Ph.D. degrees in communications and computer engineering from Tokyo Institute of Technology, Yokohama, Japan, in 2010, 2012, and 2014, respectively.

From April 2012 to September 2014, he was a Research Fellow (DC1) of Japan Society for the Promotion of Science (JSPS), Tokyo, Japan. From October 2016 to March 2020 and since October 2021, he was/is a Researcher of the Precursory Research for Embryonic Science and Technology (PRESTO), Japan Science and Technology Agency (JST), Tokyo. He is currently an Associate Professor with the Department of Computer Science, School of Computing, Tokyo Institute of Technology. His research interests include signal processing, image analysis, remote sensing, mathematical optimization, and data science.

Dr. Ono received the Young Researchers' Award and the Excellent Paper Award from the IEICE in 2013 and 2014, respectively, the Outstanding Student Journal Paper Award and the Young Author Best Paper Award from the IEEE SPS Japan Chapter in 2014 and 2020, the Funai Research Award from the Funai Foundation in 2017, the Ando Incentive Prize from the Foundation of Ando Laboratory in 2021, the Young Scientists' Award from MEXT in 2022, and the Outstanding Editorial Board Member Award from IEEE SPS in 2023. Since 2019, he has been an Associate Editor of IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS.