Spectral–Temporal Fusion of Satellite Images via an End-to-End Two-Stream Attention With an Effective Reconstruction Network

Tayeb Benzenati, Yousri Kessentini, and Abdelaziz Kallel

***Abstract*—Due to technical and budget constraints on current optical satellites, the acquisition of satellite images with the best resolutions is not practicable. In this article, aiming to produce products with high spectral (HS) and temporal resolutions, we introduced a two-stream spectral–temporal fusion technique based on attention mechanism called STA-Net. STA-Net aims to combine high spectral and low temporal (HSLT) resolution images with low spectral and high temporal (LSHT) resolution images to generate products with the best characteristics. The proposed technique involves two stages. In the first one, two fused images are generated by a two-stream architecture based on residual attention blocks. The temporal difference estimator stream estimates the temporal difference between HS images at desired and neighboring dates. The reflectance difference estimator is the second stream. It predicts the reflectance difference between the input images (HS–LS) to map LS images into HS products. In the second stage, a reconstruction network combines the latter two-stream outputs via an effective learnable weighted-sum strategy. The two-stage model is trained in an end-to-end fashion using an effective loss function to ensure the best fusion quality. To the best of our knowledge, this work represents the first attempt to address the spectral–temporal fusion using an end-to-end deep neural network model. Experimental results conducted on two actual datasets of Sentinel-2 (HSLT:10 spectral bands and long revisit period) and Planetscope (LSHT: four spectral bands and daily images) images, which proved the effectiveness of the proposed technique with respect to baseline technique.**

***Index Terms*—Attention mechanism, convolutional neural network (CNN), image fusion, multisensor image fusion, Planetscope, Sentinel-2, spectral–temporal fusion.**

I. Introduction

THANKS to the increased requests for satellite images with higher resolution, current spaceborne sensors benefit from the recent technological progress, enabling the acquisition of a wide range of data with different proprieties in terms of spatial, spectral, temporal, and radiometric resolutions. Interpretation and analysis of such data received increasing attention from the remote sensing (RS) community. In particular, image temporal series are playing a significant role for monitoring of land surface dynamics over time for various applications, including monitoring vegetation, detecting and monitoring land-cover changes, and change detection of land cover. Unfortunately, despite the substantial technological progress in optical satellites, the capture of satellite images with the best characteristics in all aspects is not yet feasible due to technical constraints and budget limitation. Researchers proposed powerful image fusion algorithms to combine satellite images with different characteristics into one product [1]. RS image fusion is an effective method aimed at merging one or multiple satellite data to generate a single product with a better interpretability. The RS fusion images are in a continuing evolution, thanks to the growing demands from leading companies, such as google earth and microsoft visual earth [2], which aim to enhance the resolution of their commercial products, and this process can be achieved by effective fusion techniques. Initially, the fusion techniques proposed a solution to enhance the spatial resolution of satellite images or to combine multimodal data. Over the past years, the aim of these methods expanded to include more challenging fusion problems, such as the fusion of different images with a different but complementary spatial and temporal resolution [3], [4]. For instance, on the one hand, satellites, such as Sentinel-2, IKONOS, and Landsat, produce images with spatial resolution varying from 3 to 30 m, which is recommended for dynamic monitoring [5], change detection [6], and land-cover mapping applications [7]. However, the observations of these kinds of satellites in a specific area are characterized by relatively long revisit cycles (Sentinel-2: 5 days and Landsat: 16 days). Besides, this period can increase due to cloud coverage or poor atmospheric conditions. This coarse temporal resolution reduces their application for monitoring and detecting rapid change, in particular, in monitoring plant health and phenology [8]. On the other hand, satellites, such as MODIS and SPOT VEGETATION, can capture daily observation but with a coarse spatial resolution ranging between 250 and 1000 m. Such a spatial resolution does not guarantee a sufficient spatial detail for monitoring and detecting changes for specific areas of

Manuscript received 16 July 2022; revised 23 November 2022 and 24 December 2022; accepted 2 January 2023. Date of publication 6 January 2023; date of current version 19 January 2023. *(Corresponding author: Tayeb Benzenati.)*

Tayeb Benzenati is with the Laboratory of Signals, Systems, Artificial Intelligence and Networks Sfax, Digital Research Centre of Sfax, Sfax 3021, Tunisia, and also with the LIMOSE Laboratory, University of Boumerdes, Boumerdes 35000, Algeria (e-mail: tayeb.benzenati@gmail.com).

Yousri Kessentini and Abdelaziz Kallel are with the Laboratory of Signals, Systems, Artificial Intelligence and Networks Sfax, Digital Research Centre of Sfax, Sfax 3021, Tunisia (e-mail: yousri.kessentini@crns.rnrt.tn; abdelaziz.kallel@crns.rnrt.tn).

This article has supplementary downloadable material available at https://doi.org/10.1109/JSTARS.2023.3234722, provided by the authors.

Digital Object Identifier 10.1109/JSTARS.2023.3234722

interest. The emergence of CubeSats, especially those designed by Planet Lab's, can provide currently daily products at high resolution (3 m) with four bands. Planetscope exploits more than 180 nanosatellites to offer a valuable data source with a great promise to deal with spatial–temporal constraints in current conventional satellite platforms. Despite these satellites' superiority in terms of spatial–temporal aspect, the acquired images have a small number of spectral bands with broad bandwidths that reduces their capabilities in some analyses, such as the monitoring of the vegetation seasonality and the high-sensitive distinguishing and detection of ice and snow [9]. Sentinel-2, on the other hand, includes a larger number of narrower bands (13) mainly in the red-edge domain, making it suitable for vegetation monitoring. As a matter of fact, Planetscope and Sentinel-2 have different but complementary proprieties. Planet has a high spatial–temporal resolution but a low spectral one, while Sentinel-2 has a high spectral resolution but a low temporal resolution. In recent years, many fusion methods have been introduced, attempting to integrate satellite images from various sensors at different resolutions to produce daily satellite images with the best resolutions. For instance, the generation of daily Sentinel-2 images is fruitful for many applications requiring high spectral and temporal resolutions. One can cite disaster monitoring [10], crop growth dynamics monitoring [11], early stage anomaly detection [12], and change detection in vegetation area [13], in general, to permit any kind of early crop monitoring practices. Therefore, the production of time series of high spectral resolution data on a daily basis, thanks to Sentinel 2-Planetscope fusion, is crucial and that can be possible only using an effective multisource multitemporal fusion technique.

A. *Related Works*

The generation of data that includes simultaneously the complementary properties of two kinds of satellite images can offer more informative products suitable for several RS applications, especially for monitoring rapid change areas. For that matter, multisource and multitemporal data fusion techniques have emerged to overcome the limits of a single sensor and introduce a possible solution in a cost-effective manner [4]. Over the past years, several works have been proposed to deal with the multisource and multitemporal fusion problem. According to Chen et al. [14], most of these methods can be grouped into three main categories: reconstruction-based techniques, unmixing-based techniques, and learning-based techniques. Regarding the reconstruction-based techniques, the fused synthetic image is generated via a weighted sum involving appropriate filters of spectrally similar neighboring pixels of the input data. Spatial temporal adaptive reflectance fusion model (STARFM) [15] is the pioneer algorithm. It aimed at combining Landsat and MODIS data to produce daily synthetic Landsat images at 30 m spatial resolution. Since then, several works have been developed to ameliorate STARFMs efficiency [16]. However, this category may lack efficacy in estimating the desirable image when a land-cover change type, occurred as the prediction of such change based on similar pixels of input images, remains difficult. The second category that is related to the unmixing-based methods generally includes the following steps:

1) clustering of available fine resolution images at a prior dates;
2) linear spectral unmixing of the pixels of the coarse images;
3) generating of fused images by substituting the spectral information using the unmixing model at the desired date.

Zhukov et al. [17] introduced the first unmixing-based technique to combine multisensor input images captured at different dates. Based on this work, many works were introduced to ameliorate the fusion performance [3], [18], [19]. Besides, Li et al. [20] introduced a time-effective approach to accelerate the fusion process while maintaining satisfactory accuracy, confirming that the literature approaches are usually time-consuming, which may be improper for practical applications. Nonetheless, these techniques suffer from large errors estimation of endmembers unmixing and insufficiency of within-class variability of the fine-scale pixels inside a single coarse one [3]. Therefore, they may not be effective in detecting endmembers' changes within a coarse pixel due to a land-cover change type. The third category includes learning-based fusion techniques. They were developed based on the machine learning mechanisms, including sparse representation [21], dictionary learning [22], extreme learning [23], and artificial neural networks [24]. This category aims to learn a mapping between prior multisensor and multitemporal image pairs, which will then be used to estimate desired images on the prediction dates. It is worth mentioning that some works [25], [26] proposed an integrated spatio–spectral–temporal fusion framework to combine multisource data with different spatial, spectral, and temporal resolutions. It generally employs a maximum posterior probability to define an inverse fusion problem. However, this approach relies upon a model optimization that, in turn, relies on prior knowledge. This process makes the product quality questionable, which can limit its exploitation for practical RS applications.

Over the past few years, deep learning, mainly, convolutional neural networks (CNN), have achieved impressive success in many computer vision applications, including image segmentation, image denoising, and super-resolution (SR). SR aims to increase the spatial resolution of low-resolution images to produce high-resolution images, which is almost the same goal of the multisensor fusion. Inspired by the state-of-the-art SR-CNN [27], Dong et al. proposed a two-stage CNN-based approach [28] to learn a complex mapping from MODIS coarse images to Landsat fine images. Liu et al. [29] introduced a CNN-based technique called spatial–temporal fusion two-stream network (StfNet), which employs residual learning of the difference between available and desired dates using SR-CNN architecture. Later, Tan et al. [30] introduced an effective generative adversarial networks spatiotemporal fusion model, termed GAN-STFM, aiming to limit the model inputs to only one pair of coarse–fine resolution images. It is clear that learning-based approaches, particularly CNN-based ones, have boosted the fusion performance with respect to the traditional fusion methods. However, the latter deal with the multisensor multitemporal fusion as an SR task and this strategy would involve significant drawbacks from the point of view of fusion quality. One can mention, in multisensor fusion, the reconstruction scale generally ranges between 8 and 16, which is considered as a large gap compared with SR (ranging from 2 and 4). Consequently, these methods cannot effectively extract texture details required to reconstruct the fused images [31]. Besides, CNN-based methods are borrowed from the pioneer SR-CNN architecture, which is shown to be insufficient for generating enough high-frequency detail due to its shallowness (includes only three layers) [32]. Moreover, SR-CNN was significantly outperformed by advanced architectures, such as deep residual networks [33] and attention mechanism [34]. It should be stressed that this kind of approach deals mainly with the spatial–temporal fusion problematic, which aims to combine satellite images of high spatial but low temporal resolution, such as Landsat and Sentinel-2, with images of lower spatial but higher temporal resolution images, such as MODIS and Sentinel-3, to synthesize high spatial–temporal data. Currently, spatial–temporal fusion represents the main approach to generate daily Sentinel-2 images, as it exploits freely available public satellite data. However, such a fusion category is considered a different and challenging task compared with SR and traditional RS fusion approaches (e.g., pansharpening) due to the following factors.

1) *Resolution factor:* The scale ratio in the spatial–temporal fusion ranges from 8 to 16, which is higher than the resolution ratio in SR and pansharpening that generally ranges from 2 and 4. Such a high ration can be problematic and leads to less fusion performance, in particular, when a borrowed SR model is explicitly applied to learn the end-to-end mapping.
2) *Temporal factor:* In spatial–temporal fusion, the inputs are captured at different dates, which make the problem even more complex, contrary to SR and pansharpening, where the images are acquired at the same time by the different modalities.
3) *Spectral factor:* Contrary to natural images used in the traditional SR problem that include only three bands, satellite images may include multiple bands covering different regions of the optical electromagnetic spectrum.

To the best of the authors' knowledge, it should be noted that there is no work in the literature addressing the spectral–temporal satellite image fusion capable of generating products with high spectral resolution on a daily basis.

B. Motivation

To tackle the drawbacks of the spatial–temporal category and benefit from the complementary spectral–temporal relationship between Planetscope and Sentinel-2 satellites, in this article, we propose the pioneer effort to deal with the spectral–temporal fusion of Planetscope and Sentinel-2 images to produce daily Sentinel-2 products using a novel deep two-stream spectral–temporal fusion technique, residual attention mechanism, and a reconstruction network using a learned weighted-sum strategy, called STA-Net. STA-Net mainly aims at integrating high temporal low spectral resolution Planetscope images and low temporal high spectral resolution Sentinel-2 images to produce high spectral and temporal products. This process can generate daily Sentinel-2 data with high accuracy. More specifically, this article makes the following contributions.

1) MODIS that has widely used in the state-of-the-art has a high-frequency coverage but a coarse spatial resolution of 250 m, which makes the estimation of Sentinel-2 at 10 m a complex task. In contrast, Planetscope can produce daily images with 3 m resolution. Its resolution can ease the fusion process and generate more accurate Sentinel-2 data.
2) Instead of using the basic SR-CNN [27], which is outperformed by deeper CNN, we adapt a deep CNN to boost the fusion performance, allowing the network to learn more complex structures at multiple levels of abstractions [35].
3) An end-to-end two-stream architecture based on residual attention blocks (RABs) is proposed to extract relevant features from a Sentinel-2 image in prior date and Planetscope one at prediction date, separately. The temporal difference estimator (TDE) focuses on learning the temporal difference, whereas the reflectance difference estimator (RDE) concentrates on learning the reflectance difference between Planetscope and Sentinel-2 images. Next, a reconstruction block is introduced to generate the final Sentinel-2 image via a learned weighting-sum manner.
4) A novel loss is developed to ensure that the estimated output is as close as possible to the target involving the two-stream outputs. Also, it penalizes bias error in the predicted image to guarantee high spectral quality.
5) The generated Sentinel-2-like data can be exploited in several agricultural contexts for monitoring different phenomena that require a high spectral resolution with dense time series.

C. Article Outline

The rest of this article is organized as follows. Section II provides the background of the attention mechanism. Section III presents the proposed spectral–temporal fusion method STA-Net. Section IV describes the considered datasets and gives the results and discussions. Finally, Section V concludes this article.

II. Background

A. Attention Mechanism

Over the past few years, after the successful application in machine translation task [36], attention mechanism has received great attention from the machine learning community, and it is now considered as a vital part for various deep neural network models for several applications of machine translation [37], speech recognition [38], and computer vision [39]. The intuition behind the attention mechanism can be understood using human biological systems, as the human visual system has the tendency to focus on adequate information while ignoring the irrelevant one in a way that can help in perception [40]. Besides the improvement of performance on several applications, attention

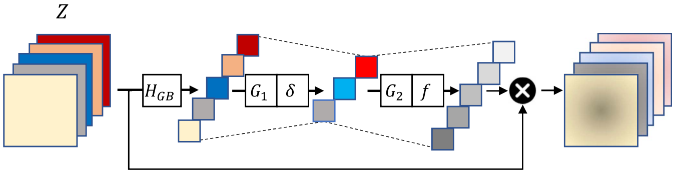

Fig. 1. Flowchart of CA block. ⊗ indicates the elementwise product.

mechanism has been widely used for enhancing the interpretability of neural networks, which are treated mostly as a black-box model [41] since it is challenging to interpret precisely how the output is inferred from the input.

1) Channel Attention (CA): CA [42] represents a meaningful application of attention mechanism in which each feature map is associated by a specific weight that defines the degree of relevancy of each feature map. CA was employed on RS pansharpening [43] to generate high resolution multispectral images, which allows the network to focus on the pertinent features from the multispectral and panchromatic images. CA can assist the CNN to pay more attention to important features and less focus on the less relevant ones, which leads to more effective feature extraction.

Let $X = [x_1, \ldots, x_c, \ldots, x_C]$ be a feature maps with C channels with size of $H \times W$. A global average information operation (H_{GP}) is first applied to the feature maps to aggregate spatial information of each channel, which can be calculated as follows:

$$z_c = H_{GP} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j), \; c = 1, \ldots, C \quad (1)$$

where $x_c(i, j)$ denotes the pixel value at (i, j) in the cth channel x_c. Next, the outputs pass through a gating mechanism that includes two fully connected layers (G_1 and G_2), which can be expressed as follows:

$$B_{CA} = f(G_2 \delta(G_1 Z)) \quad (2)$$

where $f()$ and $\delta()$ denote the sigmoid and ReLU activation function, respectively. Sigmoid is applied to define the importance degree of each channel of the feature maps by assigning a weight value between 0 and 1. CA is illustrated in Fig. 1.

III. Proposed Method

In this work, Sentinel-2 and Planetscope images are considered to validate the proposed approach. Let S be a Sentinel-2 image of b_s bands and P be a Planetscope image of b_p bands. Both images were captured within the same geographic region. We proposed a spectral–temporal fusion technique aiming at estimating a Sentinel image (S_t) captured at time t from an associated P_t image captured at the same time, and a pair of Sentinel-Planet image (S_{t-1} and P_{t-1}) captured at a prior date $t-1$. As a result, we generate products with a high spectral resolution with frequent coverage. It should be noted that Sentinel-2 and Planetscope have different resolutions. Besides the spatial difference, there is also a spectral difference not only in band numbers but also in spectral wavelength range even for the similar overlapped bands (i.e., RGB and NIR). From theoretical perspectives, the fusion process is expected to be easier for the similar bands than the nonoverlapped ones, which are supposed to be more challenging due to the difference on both spatial and spectral proprieties. Indeed, the proposed method and the integrated spatial–spectral–temporal framework [25] are similar in nature as both can produce images with high spectral and temporal resolutions. However, the former has two significant characteristics that make it different from the latter. First, on the one hand, the integrated framework requires three different kinds of data with complementary spatial, spectral, and temporal properties. On the other hand, the proposed method necessitates only two different modalities with complementary spectral and temporal resolutions, which makes it easier and more suitable for real-life applications. Second, the integrated framework needs the definition of a complex spatial–temporal–spectral relationship for different input data, as it is based on the maximum posterior probability criterion. However, since such prior knowledge is not always available, the proposed approach is deep-learning-based; hence, it does not require establishing such a complicated relationship as it tries to learn it automatically.

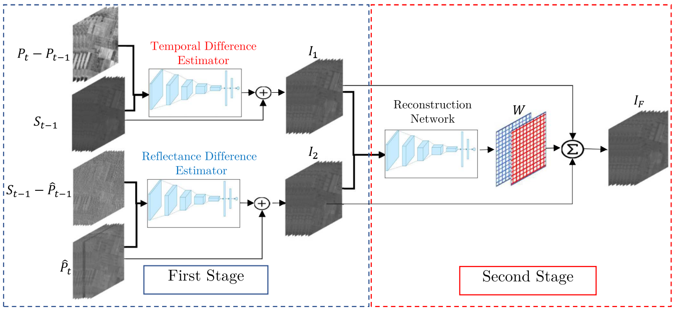

Fig. 2. Overall framework of STA-Net. $+$ and $\sum$ denote the pixelwise addition and the weighted-sum operations, respectively.

Unlike most learning-based techniques, the proposed two-stream spectral–temporal based on attention mechanism, referred to as STA-Net, requires only one pair of images at a prior date rather than two pairs of images at prior and posterior dates [28], [29], which makes the proposed approach more suitable to generate fused products without waiting for posterior date, in particular, for estimating a Sentinel-2 image at the current date. To make the most of the available information, STA-Net predicts the unknown Sentinel-2 image in a two-steam manner involving two stages. On the one hand, the first stream estimates S_t by learning the unavailable temporal changes between S_t and S_{t-1}. On the other hand, the second one estimates S_t by learning the unknown difference between S_t and P_t to map the Planestcope image into Sentinel-2 product. Next, a reconstruction network ingrates the two-stream outputs to produce the final fused product via a learned weighted sum. The general flowchart of STA-Net is illustrated in Fig. 2.

A. First Stage

Two-stream architectures were successfully applied to several tasks [44], [45], including image fusion [46], [47], which have access to two kinds of information characterized by different and complementary proprieties. Inspired by this strategy and believing that S_{t-1} and P_t contain different and complementary information as they were acquired by different sensors and included different spatial, spectral, and temporal resolutions, by which it is possible to produce a Sentinel-2-like images by learning a complex mapping using an appropriate CNN, we introduced a two-stream CNN based on attention mechanism. The two streams have the same objective, i.e., producing a Sentinel-2-like images but using different concepts.

1) First Stream: TDE: TDE aims to generate the first intermediate fused product (I_1) by learning a complex mapping (ϕ_1) of the temporal changes. Instead of taking roughly P_t and S_{t-1} as inputs, which does not make a full use of the available information since P_{t-1} is also available, which, combined with P_t, can assist the CNN to learn valuable features, this stream includes two inputs: S_t and $P_t - P_{t-1}$ ($D_T^{t,t-1}$), which are concatenated to act as a single input. Since S_t and S_{t-1} can be highly correlated, a residual learning is employed to learn only the residual difference between S_t and S_{t-1}. This difference represents the temporal change within the study area, which should be added to S_{t-1} to reconstruct the first intermediate fused image I_1. The residual learning is proven to improve the accuracy and ease the training [48] compared with the traditional stacked convolutional layers. Besides, it provides more interpretability to the fusion algorithm. The image generated by the first stream can be summarized by the following formula:

$$I_1 = S_{t-1} + \underbrace{\phi_1(S_{t-1}, D_T^{t,t-1}; \theta_1)}_{I_{D_T}} \tag{3}$$

where I_{D_T} indicates the temporal difference image that needs to be inserted into S_{t-1} to produce I_1, and θ_1 denotes the network's parameters to be trained.

2) Second Stream: RDE: The majority of works predict the desired image using a mapping into S_t or by estimating the difference image that need to be injected into S_{t-1}. This strategy alone may not lead to an effective performance, especially when considerable changes occur within the area. In our work, assuming that Sentinel-2 and Planetscope images are captured within the same region but having different reflectance responses as they are acquired via different sensors, the second stream, called RDE, aims to reconstruct a Sentinel-2-like image (I_2) by learning a complex mapping (ϕ_2) from P_t. In other words, this network learns the difference between Planetscope and Sentinel-2 images with the aim to transform the Planetscope image into a Sentinel-2-like image. However, such a strategy cannot be applied explicitly due to the difference in band number between both constellations, in particular, for the additional red-edge ones of Sentinel-2 that are unavailable in Planetscope products. Therefore, aiming to adjust the equivalent spectral bands, the latter Sentinel-2 bands are estimated using the closest Planetscope bands in terms of root mean squared error (RMSE) forming eight band versions of Planetscope images ($\hat{P}_t$ and $\hat{P}_{t-1}$). Sentinel bands: B2, B3, B4, and B8 are estimated by the associated Planetscope bands Blue, Green, Red, and NIR, respectively, whereas Sentinel-2 red-edge bands: B5 and B6 are estimated based on the Planetscope red band, and B7 and B8a are predicted via the Planetscope NIR band. Aiming to ease the learning process for the network, we provide two inputs for the CNN: $\hat{P}_t$ and $S_{t-1} - P_{t-1}$ (D_R^{t-1}). D_R^{t-1} provides additional accessible knowledge to the network, as it includes the reflectance difference at a prior date, which can assist the network in learning the mapping from the inputs to S_t. The image produced by this stream can be expressed as follows:

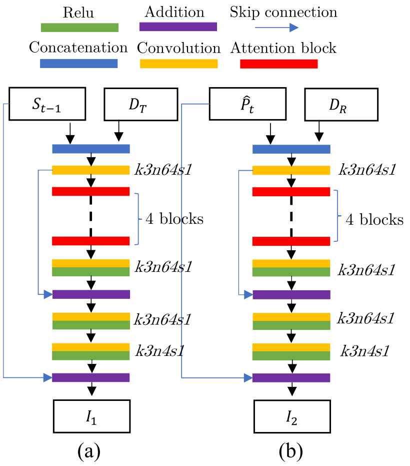

Fig. 3. Detailed architecture of the first stage two streams: k denotes the filter's size, n indicates the number of output filters, and s represents the stride size. (a) Temporal Difference Estimator. (b) Reflectance Difference Estimator.

$$I_2 = \hat{P}_t + \underbrace{\phi_2(\hat{P}_t, D_R^{t-1}, \theta_2)}_{I_{D_R}} \tag{4}$$

where I_{D_R} denotes the radiometric difference that must be injected into $\hat{P}_t$ to produce I_2, and θ_2 represents the network parameters to be optimized. The detailed architecture of each stream is shown in Fig. 3. Each stream has the same architecture but different weights as it was trained using different inputs. Each stream includes three main parts: shallow feature extraction, deep feature extraction, and difference reconstruction part. First, one convolution is performed to extract shallow features from the corresponding inputs of each stream. Next, four RABs are applied for deep feature extraction. The used attention block is described in Section III-A3. Finally, from the elementwise sum of shallow and deep feature, two convolutions are used to estimate the residual difference that needs to be inserted into S_{t-1}/P_t to produce I_1/I_2 for TDE/RDE streams, respectively.

3) RABs: It has been shown that residual blocks can be utilized to develop effective deep CNN [49]. However, since the traditional residual blocks apply equal attention to all features, this kind of network is generally difficult to train and reconstruct the high-frequency details [34]. To overcome, the attention mechanism has been proposed in [42], and it offers complementary characteristics. It can focus on more informative features and ignore the useless ones, which help the networks to easily capture the important features and reconstruct finer texture details. Inspired by this trend [34], [50], we proposed an RAB, which combines the effectiveness of attention mechanism and

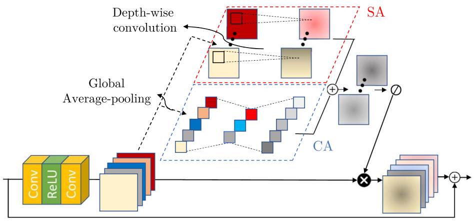

Fig. 4. Structure of the attention block: $\oplus$ and $\otimes$ indicate the elementwise addition and multiplication, respectively, and $\oslash$ represents the sigmoid activation function. SA and CA are the spatial and channel attention blocks, respectively.

the regular residual blocks. The architecture of a single attention block is shown in Fig. 4. The latter includes two parts to model two kinds of information suitable for spectral–temporal fusion: CA and spatial attention (SA), as satellite images include low- and high-frequency components. The latter provides valuable information that represents edges, texture, and other kinds of details. Focusing on such components can be beneficial for the network to reconstruct the desired fused product. Accordingly, to pay more attention to high-frequency information, CA, as described in Section II-A1, is introduced to exploit better each channel of feature maps. This strategy can prioritize the channels with more relevant information. It is common knowledge that channels in each feature map can include different representations based on the applied filter's objective. For instance, some filters can capture horizontal edges, and other filters extract vertical ones, and obviously, each of them plays a significant role in reconstructing the fused product. Trying to separate the spatial information and depthwise one, we performed an SA using a depthwise convolution [51] to exploit spatial interdependencies of each channel while maintaining channel-specific characteristics. Contrary to regular convolutions, which are applied over multiple channels, depthwise convolution traits each channel individually to produce two-dimensional feature maps for each one. This part can be expressed as follows:

$$B_{\mathrm{SA}} = f_{\mathrm{depth}}(X) \tag{5}$$

where f_{depth} denotes the depthwise convolution operation via three kernels. The final output of an attention block combines the spatial and spectral attention outputs and can be expressed as follows:

$$\hat{X} = H_{AB}(X) = f(B_{\mathrm{SA}} \oplus B_{\mathrm{CA}}) \otimes X \tag{6}$$

where $\hat{X}$ represents the final output of the RAB, B_{CA} and B_{SA} denote the outputs of CA and SA, respectively, and $\oplus$ and $\otimes$ indicate the elementwise sum and product, respectively. The RAB is composed of successive stacked attention blocks. Assuming that the output of the ith RAB is F_i, the latter can be calculated as follows:

$$F_{i+1} = H_{AB}(F_{i+1}^{1}) + F_i \tag{7}$$

where $H_{AB}(.)$ indicates the operation of RAB, and F_{i+1}^{1} is the feature maps output of C channels after the application of convolution, ReLU, and convolution on F_i.

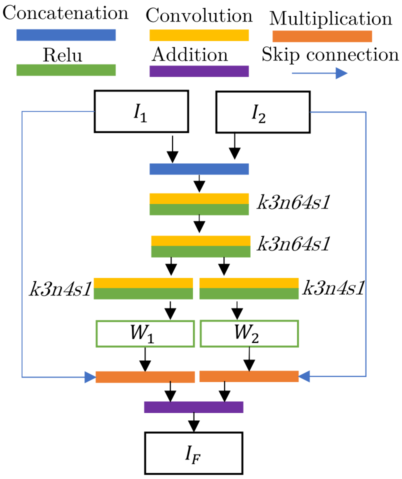

Fig. 5. Detailed architecture of the reconstruction network: k denotes the filter's size, n indicates the number of output filters, and s represents the stride size.

B. Second Stage: Reconstruction Network

At the end of the first stage, two outputs are generated, each of them has complementary features. To make the most of the latter, the fusion stage aims to extract the hierarchical characteristics of the two outputs I_1 and I_2 to catch complimentary properties and produce the final fused product (I_F). To this end, inspired by CNNs learning capacity, we introduced a reconstruction block to merge the two outputs to recover the desired image. Instead of predicting the latter directly in a black-box manner lacking physical interpretability, this stage blends the two inputs by learning the appropriate pixelwise weighted sum, which aims to guide the network to select the best pixels that boost the fusion performance. The intuition behind this strategy is that the performances of two-stream's outputs vary depending on the spatio-temporal features. For instance, areas with minor changes are better preserved by the STD as S_t is almost equal to S_{t-1}, whereas the ones with significant changes and low spatial variation are better reconstructed by the second stream (RDE) since S_t and P_t are highly correlated. The final fused product can be expressed by the following formula:

$$I_F = \sum_{j=1}^{2} \underbrace{\phi_F(I_1, I_2, \theta_F)_j}_{W_j} \cdot I_j \tag{8}$$

where W represents the outputs of the network that represents the learned weights, $\cdot$ denotes the pixelwise product, I_i indicates the intermediate fused product of the ith stream, I_F represents the final fused product, and θ_F indicates the network's parameters to be optimized. The architecture of the second stage is illustrated in Fig. 5. First, the two-stream outputs are concatenated to form a single input to pass through two convolution layers to extract features that encourage the network to select the best pixels of the input. Next, two parallel convolutions are applied to estimate the appropriate weights for the associated intermediate fused products I_1 and I_2, respectively. Aiming at blending the latter products via a weighted sum, the estimated weights are multiplied pixel-by-pixel by their associated fused images to choose the best value for each pixel. The results are added together at pixel level to produce the final fused image I_F.

C. Proposed Loss

The design of the loss function is vital for network training and prediction. Therefore, unlike some CNN-based techniques [28], [29] that are often time-consuming during the training process since they train each part of the network separately, STA-Net employed a combined loss function to optimize the network's parameters in an end-to-end manner. This strategy updates all the network parameters simultaneously via a single loss function, which leads to fast and accurate fusion results. The objective of the training is to optimize the following loss function:

$$L(\theta) = \alpha_1 L_1 + \alpha_2 L_2 + \alpha_3 L_3 \tag{9}$$

where α_1, α_2, and α_3 represent the loss weights used to equilibrate the contribution of each part, and each part is defined as follows:

$$L_1 = \frac{1}{N_s} \sum_{i=1}^{N_s} \left\| S_t^i - I_f^i \right\|_1$$

$$L_2 = \frac{1}{N_s} \sum_{i=1}^{N_s} \left\| S_t^i - I_1^i \right\|_2 + \left\| S_t^i - I_2^i \right\|_2$$

$$L_3 = \frac{1}{N_s} \sum_{i=1}^{N_s} \left\| \bar{I}_1^i - \bar{S}_t^i \right\|_2 + \left\| \bar{I}_1^i - \bar{S}_t^i \right\|_2$$

where S_t indicates the reference Sentinel-2 image, i denotes the sample index of minibatch of N_s, and the bar operation $(\bar{.})$ indicates the mean value. The first part is the mean absolute error (known as l_1 or MAE) between the final predicted and the reference images, aiming to obtain fused images as close as possible to the reference images. As the final output combines the two-stream intermediate results, the second part encourages the first stage networks to produce Sentinel-2-like products similar to the reference ones via a mean squared error loss (l_2) as well as to allow an end-to-end training. Concerning the third part, as Planetscope and Sentinel-2 images have different radiometric responses, it is used to ensure that the intermediate predicted images have the same mean as the reference images to preserve their spectral information. The MAE (l_1) is used for the first part, as it provides better performance than l_2 loss and better convergence behavior [49], which guarantees the best fusion result for the final fused product. At the same time, l_2 is utilized for the other parts because it is less sensitive to the variation than l_1. This is mainly the case when the two images are very similar to each other, which allows the network to give more focus on the final predicted product while ensuring competitive performance for the two-stream outputs.

IV. Experimental Results

A. Datasets

To assess the proposed technique's fusion performance, two datasets acquired by Planetscope and Sentinel-2 satellites, within the same area and the same date, are used for the training and the evaluation procedure, respectively. The first dataset (denoted as Sfax dataset) was captured over the region of Sfax city, Tunisia (35°06 N, 10°54 E), located about 30 km north of the city, which covers a complex area that includes an agricultural area from Jebiniana town. The training dataset consists of two pairs of Planetscope and Sentinel-2 images captured on March 2nd, 2017 and November 12th, 2017, respectively. The test dataset includes two pairs acquired on June 5th, 2018 and December 12th, 2018, respectively. The second dataset (denoted as Coleambally dataset) was acquired over the irrigation area located in Coleambally (34°54 S, 146°1 E) in southern New South Wales, Australia. The training images include two pairs captured on May 23, 2019 and December 14, 2019. The test dataset comprises two pairs of images acquired on May 15, 2020 and November 21, 2020. This period between the acquisition allows the apparition of significant phenological changes due to the growth of plants and different types of vegetation as well as shadow variation due to the sun inclination variation. For each dataset, the first date represents the images at a prior date $(t-1)$, and the second date indicates the desired image at t that needs to be estimated, which is used as a ground-truth image for evaluation of the fusion product. The size of each training dataset is 2100×2100 pixels at Sentinel-2 10 m scale. Regarding the evaluation process, for each test dataset, 25 images of size 256×256 pixels are chosen to assess quantitatively the performance of the proposed approach. Furthermore, a qualitative evaluation was carried out visually using one scene from the selected ones.

For two constellations, Sentinel-2 and Planetscope, the products with high processing levels are considered; level L2A with Bottom of Atmosphere reflectance for Sentinel-2, which includes an atmospheric correction, the Analytic Ortho Scene (3B) for Planetscope. Regarding Sentinel-2 products, in this work, eight spectral bands are considered. The broad spectral bands: B2 (Blue 458–523 nm), B3 (Green, 543–578 nm), B4 (Red, 650–680 nm), and B8 (NIR, 785–900 nm) with 10 m ground sampling distance and the vegetation red-edge bands: B5 (Red-Edge 1, 698–713 nm), B6 (Red-Edge 2, 733–748 nm), B7 (Red-Edge 3, 773–793 nm), and B8a (Narrow NIR, 855–875 nm) with 20 m, which are valuable for several vegetation study applications, such as identifying vegetation types [52] and detection of crop disease [53]. For Planetscope, the accessible four spectral bands (blue, green, red and near infrared) at 3 m are used. All selected datasets are cloud-free, geometrically corrected images. Sentinel-2 bands at 20 m were resampled to fit the resolution of 10 m bands. Besides, Planetscope images at 3 m were upscaled to 10 m to fit the Sentinel-2 resolution. Although the downsampling of Planetscope images to 10 m may lose some spatial details, it is more suitable for our approach to process with such a resolution for three reasons. First, our

TABLE I
CHARACTERISTICS AND REQUIRED PROCESSING OF PLANETSCOPE AND SENTINEL-2 USED IN THIS STUDY

Satellite	**Planetscope**	**Sentinel-2**
Spatial resolution	3 m	10, 20, and 60 m
Number of bands	4	13
Spectral range	0.44 -2.19 nm	0.45-0.67 nm
Considered bands	All	10 m : B2, B3, B4, B8 20 m : B5, B6, B7, and B8a
Temporal resolution	Daily	5 days
Preprocessing	Resampling to 10 m	Resampling the 20 m bands to 10 m
Post-processing	—-	Downsampling of upsampled bands to the original resolution

model aims to generate Sentinel-2 at 10 m, which is the highest spatial resolution of Sentinel-2. Second, our technique is not supposed to be sensitive to minor changes that are smaller than 10 m resolution, and they should be ignored. Third, the fusion performance remained approximately the same for both resolutions 10 m and 3 m. As postprocessing, the resampled bands of Sentinel-2 (i.e., B5, B6, B7, and B8a) will be downsampled using a Gaussian low-pass kernel that mimics the modulation transfer function to restore their original resolution of 20 m. Table I summarizes the spatial, spectral, temporal resolutions and the required preprocessing of Sentinel-2 and Planetscope used in this work.

B. Implementation Details

For the training stage, the training images were cropped into patches of size 41×41 pixels and generating, therefore, 3000 samples for allowing the training process. A convolutional filter of size 3×3 was set in all weight layers of the network. Regarding the optimization, the network was trained for 1200 epochs (25 818 iterations) and optimized via Adam optimizer [54] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch size was set to 64 during the training process. The loss's weights α_1, α_2, and α_3 were empirically set to 1 in the present work. These weights are set empirically instead of being learned, as we noticed that this strategy can lead to more stable training and best fusion performance. The learning rate was first initialized to 10^{-4} and divided by 10 every 300 epochs. The network was implemented and tested through NVIDIA Titan Xp GPU with 32 GB of RAM. The training stage is achieved when the loss does not improve for 50 epochs. In the prediction phase, as our STA-Net processes images of arbitrary size respecting the limit of GPU's memory, the tested image was predicted without the need for cropping, unlike the training phase.

C. Quality Assessment

Quantitative validation represents an indispensable step for evaluating and comparing each fusion technique. Thanks to the existence of Sentinel-2 images at the desired dates that serve as reference images, it is possible to evaluate the fused images with their associated target ones in a full-reference method. For this reason, several full-reference metrics have been proposed to measure the spectral and spatial quality of fused products. In this work, the fusion performances have been evaluated through four highly used metrics, including the RMSE, the correlation coefficient (CC) [55], the spectral angle mapper (SAM) [2], the structure similarity (SSIM) [56], and the universal image quality index (UIQI) [57]. In addition to the quantitative validation, a qualitative assessment was performed via a visual inspection to visually evaluate the fused product, which helps identify other kinds of spectral and spatial distortions, which may not be noticed in a quantitative manner.

TABLE II
COMPARISON OF FUSION PERFORMANCE ON COLEAMBALLY DATASET DEPENDING ON THE EMPLOYED ARCHITECTURE

Metric	Reference	Band	Single-stream	Two-stream
RMSE	0	Mean	0.0123	**0.0079**
CC	1	Mean	0.9530	**0.9739**
SAM	0	Mean	0.0795	**0.0594**
SSIM	1	Mean	0.9497	**0.9717**

TABLE III
COMPARISON OF FUSION PERFORMANCE ACHIEVED BY DIFFERENT NETWORK'S WIDTH

Metric	Reference	Band	2	4	6	8
RMSE	0	Mean	0.0105	**0.0079**	0.0089	0.0090

D. Ablation Study

Aiming to investigate the influence of the network's components, an ablation study was performed to show the effectiveness of the proposed method as well as to select the optimal parameters that ameliorate the fusion accuracy. More precisely, such a study intends to assess the direct impact of two-stream architecture, weighted-sum strategy, loss function, and the employed attention blocks to gain in fusion efficiency.

1) Influence of Two-Stream Architecture: The use of two-stream architecture is one of the main contributions of this work. Therefore, trying to show the effectiveness of the proposed two-stream architecture over a one-stream one, we implemented a one-stream network by stacking the inputs of each stream on the original method into a single input. Table II describes the achieved quantitative fusion results on Coleambally dataset depending on the employed architecture. The proposed two-stream architecture shows a better fusion ability than the single-stream one in all aspects. It achieved the best scores in all metrics, proving the suitability of a two-stream architecture in the proposed method.

2) Influence of the Network's Depth: An ablation study was also conducted to investigate the impact of the network's depth (i.e., the number of attention blocks) on the fusion performance. It is known that the number of the network's parameters grows linearly with the depth. Therefore, we should carefully find the best tradeoff between the fusion performance and the network's depth. The quantitative results illustrated on Table III show that fusion performance and the depth of the model have a positive relationship. However, this trend was downward after

TABLE IV
COMPARISON OF FUSION PERFORMANCE ACHIEVED BY WEIGHTED-SUM AND TYPICAL STRATEGIES ON COLEAMBALLY DATASET

Metric	Reference	Band	Typical	I_1	I_2	Proposed
RMSE	0	Mean	0.0099	0.0097	0.0126	**0.0079**
CC	1	Mean	0.9683	0.9656	0.9443	**0.9739**
SAM	0	Mean	0.0653	0.0680	0.0854	**0.0594**
SSIM	1	Mean	0.9633	0.9586	0.9285	**0.9717**

TABLE V
VARIATION OF FUSION SCORES BASED ON THE EMPLOYED LOSS FUNCTION ON COLEAMBALLY DATASET

Metric	Reference	Band	l_1	l_2	Proposed
RMSE	0	Mean	0.0097	0.0100	**0.0079**
CC	1	Mean	0.9606	0.9588	**0.9739**
SAM	0	Mean	0.0664	0.0683	**0.0594**
SSIM	1	Mean	0.9545	0.9519	**0.9717**

four blocks. For this reason, a depth of four blocks has been chosen to develop the proposed technique.

3) Influence of Weighted-Sum Strategy: Employing a weighted-sum strategy to produce the desired fused images represents another originality of the present work. Attempting to examine the impact of this strategy to improve the fusion accuracy of the proposed method, we compared the latter with a typical strategy that does not employ any weighted-sum strategy. In other words, it produces the desired image directly from the intermediate fused products via the second stage's reconstruction network. Besides, it was compared with the intermediate images: I_1 and I_2, generated by TDE and RDE, respectively. Table IV illustrates the obtained fusion score on Coleambally dataset using the considered mechanism. It may be seen that the proposed weighted-sum strategy is highly advantageous over the typical ones since the former produces the best fusion accuracy in all considered aspects. Furthermore, such a mechanism can further boost the fusion performance by combining the intermediate images in a learned weighted-sum manner into a single, more accurate product. Consequently, including a weighted-sum strategy in the last layer can effectively enhance the fusion results, which further proves the effectiveness of the proposed method.

4) Influence of Loss Function: Aiming to evaluate the influence of loss function to enhance the fusion quality, we compared the proposed loss function with l_1 and l_2 loss functions, which were largely used in the literature in several image enhancement applications, especially satellite images fusion. Table V illustrates the fusion scores obtained by the compared loss function on Coleambally dataset. It can be seen that l_1 achieves higher fusion scores than l_2, which shows the significance of employing l_1 as a principal part of the proposed loss. However, the proposed loss function offers the best fusion results in all bands in terms of RMSE. The obtained scores prove the effectiveness of the proposed loss function to gain in fusion performance. In particular, it shows the importance of optimizing the network's parameters using a combined loss function that considers the output of each stream.

5) Influence of RABs: In this experiment, we analyze the impact of the chosen blocks as they play a significant role in boosting the proposed method's fusion quality. Table VI presents the obtained fusion results by the employed RABs, residual blocks [48], and residual dense blocks [58] on Coleambally dataset. Such blocks are widely used in computer vision applications, thanks to their ability to develop deeper neural networks. The achieved fusion scores by the compared residual blocks are very close to each other. However, the employed residual blocks yielded the highest scores compared with other blocks with respect to all considered metrics. These results prove that RABs represent the ideal choice for the proposed method, given its high fusion accuracy.

TABLE VI
COMPARISON OF FUSION PERFORMANCE DEPENDING ON THE USED RESIDUAL BLOCKS ON COLEAMBALLY DATASET

Metric	Reference	Band	Residual	Dense	Attention
RMSE	0	Mean	0.0090	0.0080	**0.0074**
CC	1	Mean	0.9610	0.9555	**0.9694**
SAM	0	Mean	0.0728	0.0629	**0.0622**
SSIM	1	Mean	0.9660	0.9704	**0.9750**

E. Quantitative Validation

To evaluate the fusion performance of the proposed technique, the latter was compared with the reconstruction-based approach STARFM [15], the common CNN baseline for image processing, SRCNN [27], trained for spectral–temporal fusion, and the well-established spatial–temporal fusion methods based on deep learning: two-stream convolutional neural network for spatiotemporal image fusion (StfNet) [29] and GAN-STFM [30], adapted to deal with spectral–temporal fusion. Both SRCNN and StfNet were implemented and trained by ourselves, and all the parameters of these techniques are set as described in their original papers to ensure the optimal performance. As STARFM combines satellite images with similar spectral properties and requires the same number of input and output bands, the additional Sentinel-2 bands are estimated via the closest Planetscope bands in terms of RMSE. Sentinel-2 bands: B2, B3, B4, and B8 are estimated by the corresponding Planetscope bands Blue, Green, Red, and NIR, respectively. On the other hand, the remaining Sentinel-2 red-edge bands: B5 and B6 are estimated based on the Planetscope red band, and B7 and B8a are predicted via the Planetscope NIR band. Tables VII and VIII describe the quantitative scores of the considered fusion techniques and the associated S_{t-1} on Sfax and Coleambally datasets, respectively. As it can be seen, the correlation between the observations at the two dates is low, for both datasets, because of the long period between the two acquisitions, which can make the fusion task more complex for the CNN-based approaches to learn from the data. As expected, the traditional approach: STARFM achieves the weakest scores in terms of quantitative quality as it cannot address the phenology changes in the homogeneous regions. SRCNN, StfNet, and GAN-STFM obtain an RMSE mean of around 0.01 on both datasets. STA-Net enhances the accuracy obviously, with an RMSE mean of roughly 0.007. SSIM mean scores calculated from the eight estimated bands reached the highest values of 0.97 for the proposed approach, indicating

TABLE VII
QUANTITATIVE SCORES OF THE FUSED PRODUCTS ON SFAX DATASET

Metric	Ref	Band	S_{t-1}	STARFM	SR-CNN	StfNet	GAN-STFM	STA-Net
RMSE	0	B2	0.0165	0.0396	0.0111	0.0066	0.0054	**0.0042**
		B3	0.0262	0.0321	0.0127	0.0086	0.0082	**0.0060**
		B4	0.0390	0.0241	0.0161	0.0101	0.0158	**0.0078**
		B8	0.0592	0.0381	0.0176	0.0121	0.0227	**0.0081**
		B5	0.0423	0.0230	0.0133	0.0093	0.0169	**0.0066**
		B6	0.0520	0.0264	0.0152	0.0104	0.0227	**0.0070**
		B7	0.0561	0.0346	0.0161	0.0106	0.0229	**0.0071**
		B8a	0.0602	0.0376	0.0168	0.0109	0.0251	**0.0073**
		Mean	0.0439	0.0319	0.0149	0.0099	0.0175	**0.0068**
CC	1	B2	0.6984	0.7286	0.8949	0.9450	0.9578	**0.9739**
		B3	0.7155	0.7819	0.9215	0.9520	0.9648	**0.9766**
		B4	0.8021	0.7836	0.9187	0.9689	0.9677	**0.9835**
		B8	0.7362	0.7996	0.9086	0.9547	0.9641	**0.9803**
		B5	0.7861	0.8037	0.9406	0.9700	0.9789	**0.9858**
		B6	0.7520	0.7992	0.9296	0.9647	0.9748	**0.9848**
		B7	0.7467	0.8289	0.9297	0.9637	0.9726	**0.9841**
		B8a	0.7254	0.8179	0.9241	0.9586	0.9711	**0.9821**
		Mean	0.7453	0.7929	0.9195	0.9597	0.9690	**0.9814**
SAM	0	B2	0.1306	0.1140	0.0708	0.0515	0.0425	**0.0334**
		B3	0.1216	0.0897	0.0605	0.0447	0.0380	**0.0309**
		B4	0.1049	0.0922	0.0585	0.0369	0.0370	**0.0268**
		B8	0.0940	0.0704	0.0495	0.0351	0.0316	**0.0233**
		B5	0.0950	0.0757	0.0439	0.0314	0.0265	**0.0217**
		B6	0.0976	0.0720	0.0445	0.0317	0.0276	**0.0211**
		B7	0.0932	0.0631	0.0438	0.0305	0.0273	**0.0204**
		B8a	0.0855	0.0594	0.0420	0.0300	0.0262	**0.0198**
		Mean	0.1028	0.0796	0.0516	0.0365	0.0321	**0.0247**
SSIM	1	B2	0.9132	0.9044	0.9538	0.9747	0.9805	**0.9773**
		B3	0.8722	0.9087	0.9346	0.9639	0.9718	**0.9802**
		B4	0.8646	0.8810	0.8996	0.9675	0.9612	**0.9796**
		B8	0.8190	0.8737	0.9042	0.9461	0.9561	**0.9730**
		B5	0.8684	0.8966	0.9461	0.9688	0.9733	**0.9810**
		B6	0.8389	0.891	0.9408	0.9578	0.9690	**0.9782**
		B7	0.8342	0.9051	0.9383	0.9555	0.9678	**0.9769**
		B8a	0.8336	0.9027	0.9394	0.9535	0.9659	**0.9754**
		Mean	0.8555	0.8954	0.9321	0.9610	0.9682	**0.9790**
UIQI	1	B2	0.9757	0.9209	0.9854	0.9981	0.9989	**0.9993**
		B3	0.9711	0.6134	0.9937	0.9986	0.9987	**0.9994**
		B4	0.9679	0.6332	0.9973	0.9991	0.9971	**0.9995**
		B8	0.9671	0.6294	0.9958	0.9992	0.9960	**0.9997**
		B5	0.9693	0.6344	0.9985	0.9993	0.9970	**0.9997**
		B6	0.9694	0.6341	0.9964	0.9993	0.9954	**0.9997**
		B7	0.9694	0.6310	0.9954	0.9994	0.9958	**0.9998**
		B8a	0.9691	0.6303	0.9951	0.9994	0.9954	**0.9998**
		Mean	0.9699	0.9716	0.9947	0.9991	0.9968	**0.9996**

TABLE VIII
QUANTITATIVE SCORES OF THE FUSED PRODUCTS ON COLEAMBALLY DATASET

Metric	Ref	Band	S_{t-1}	STARFM	SR-CNN	StfNet	GAN-STFM	STA-Net
RMSE	0	B2	0.0270	0.0234	0.0208	0.0163	0.0102	**0.0054**
		B3	0.0276	0.0184	0.0141	0.0170	0.0109	**0.0062**
		B4	0.0612	0.0215	0.0138	0.0146	0.0148	**0.0100**
		B8	0.1341	0.0410	0.0134	0.0180	0.0142	**0.0103**
		B5	0.0497	0.0357	0.0138	0.0111	0.0094	**0.0077**
		B6	0.1093	0.1305	0.0118	0.0170	0.0098	**0.0084**
		B7	0.1259	0.0434	0.0114	0.0126	0.0098	**0.0084**
		B8a	0.1271	0.0350	0.0160	0.0191	0.0120	**0.0092**
		Mean	0.0765	0.0436	0.0151	0.0116	0.0114	**0.0079**
CC	1	B2	0.3478	0.8237	0.9369	0.9211	0.9341	**0.9535**
		B3	0.3547	0.7746	0.9532	0.9429	0.9572	**0.9600**
		B4	0.4955	0.5933	0.9536	0.9513	0.9061	**0.9691**
		B8	0.2359	0.7355	0.9841	0.9559	0.9769	**0.9845**
		B5	0.3533	0.8637	0.9735	0.9750	0.9368	**0.9852**
		B6	0.2452	0.6349	0.9764	0.9659	0.9749	**0.9827**
		B7	0.2576	0.6585	0.9803	0.9717	0.9810	**0.9864**
		B8a	0.2482	0.674	0.9857	0.9727	0.9695	**0.9902**
		Mean	0.3207	0.7198	0.9645	0.9531	0.9546	**0.9739**
SAM	0	B2	0.3196	0.1546	0.1007	0.1087	0.1001	**0.0874**
		B3	0.2573	0.2086	0.0760	0.0825	0.0735	**0.0705**
		B4	0.3351	0.2137	0.0973	0.0927	0.1473	**0.0739**
		B8	0.4361	0.1567	0.0596	0.0788	0.0551	**0.0461**
		B5	0.2599	0.1631	0.0840	0.0605	0.1048	**0.0466**
		B6	0.4171	0.1906	0.0564	0.0622	0.0567	**0.0452**
		B7	0.4307	0.1665	0.0516	0.0582	0.0474	**0.0409**
		B8a	0.4225	0.1611	0.0653	0.0749	0.0629	**0.0369**
		Mean	0.3553	0.1769	0.0768	0.0808	0.0810	**0.0594**
SSIM	1	B2	0.8192	0.8948	0.9303	0.9433	0.9637	**0.9802**
		B3	0.8580	0.9062	0.9634	0.9537	0.9742	**0.9738**
		B4	0.7090	0.8737	0.9532	0.9438	0.9215	**0.9586**
		B8	0.3914	0.7482	0.9564	0.8971	0.9573	**0.9578**
		B5	0.7947	0.8440	0.9652	0.9611	0.9400	**0.9758**
		B6	0.4884	0.5313	0.9652	0.9404	0.9629	**0.9731**
		B7	0.4446	0.7724	0.9656	0.9371	0.9656	**0.9727**
		B8a	0.4493	0.8116	0.9637	0.9318	0.9532	**0.9734**
		Mean	0.6415	0.7978	0.9548	0.9391	0.9548	**0.9717**
UIQI	1	B2	0.7591	0.8395	0.9019	0.9337	0.9702	**0.9887**
		B3	0.9082	0.7239	0.9765	0.9678	0.9948	**0.9949**
		B4	0.7467	0.6519	0.9933	0.9922	0.9533	**0.9936**
		B8	0.7883	0.7316	0.9968	0.9958	0.9937	**0.9980**
		B5	0.9232	0.7391	0.9946	0.9966	0.9760	**0.9978**
		B6	0.7930	0.7230	0.9956	0.9925	0.9966	**0.9985**
		B7	0.7838	0.7258	0.9968	0.9966	0.9974	**0.9988**
		B8a	0.8128	0.7247	0.9956	0.9962	0.9912	**0.9978**
		Mean	0.8144	0.7324	0.9814	0.9839	0.9841	**0.9960**

that the fused and reference images have the best structural similarity. Besides, STA-Net offers the highest scores in terms of CC and UIQI, which denotes an effective reconstruction of small-size structures [59]. In terms of spectral fidelity measured using SAM index, STA-Net conserves better spectral signature, as it produces the best SAM results among the compared techniques. Surprisingly, the fusion accuracy of B5 and B6 bands surpasses the one of B4 and B8 that have corresponding bands in Planetscope (RGB and NIR). This phenomenon may be due to the difference of the spectral characteristics of B4 and B8, and the corresponding Red and NIR bands of Planetscope, as there is only a partial overlap between them (see Section IV-A). All the aforementioned results indicate that the proposed technique offers the best fused products in terms of spatial, spectral, and radiometric properties.

F. Qualitative Validation

The visual inspection is considered a fundamental step to validate each fusion approach along with the quantitative validation. It can highlight different kinds of noticeable distortions and artifacts on the fused images, which help compare the performance between the considered fusion techniques. Figs. 6 and 7 illustrate the fusion results from the considered techniques on the Sfax and Coleambally datasets, respectively, along with the reference Sentinel-2 image (S_t) and its associated Planetscope product at the same date P_t in addition to Sentinel-2 image at the prior date (S_{t-1}). It can be seen at first sight that the visual results are in line with the quantitative observations from Tables VII and VIII. We can observe that all the methods are able to estimate the phenological changes occurred between the desired and prior dates. STARFM suffers from a serious blurring effect and lost a lot of detail in some heterogeneous areas. Besides, high spectral distortion is noticed in some parts of the image, as the color appears different from the reference image (red rectangle). Regarding SR-CNN, its fused product suffers from a serious blurring effect on the whole image and lost a lot of detail in some heterogeneous areas. Besides, a significant spectral distortion is noticed in some parts of the fused image, as the color appears dissimilar from the reference image (blue rectangle). StfNet, on the other hand, yields better performances than SR-CNN but lacks spatial details in some

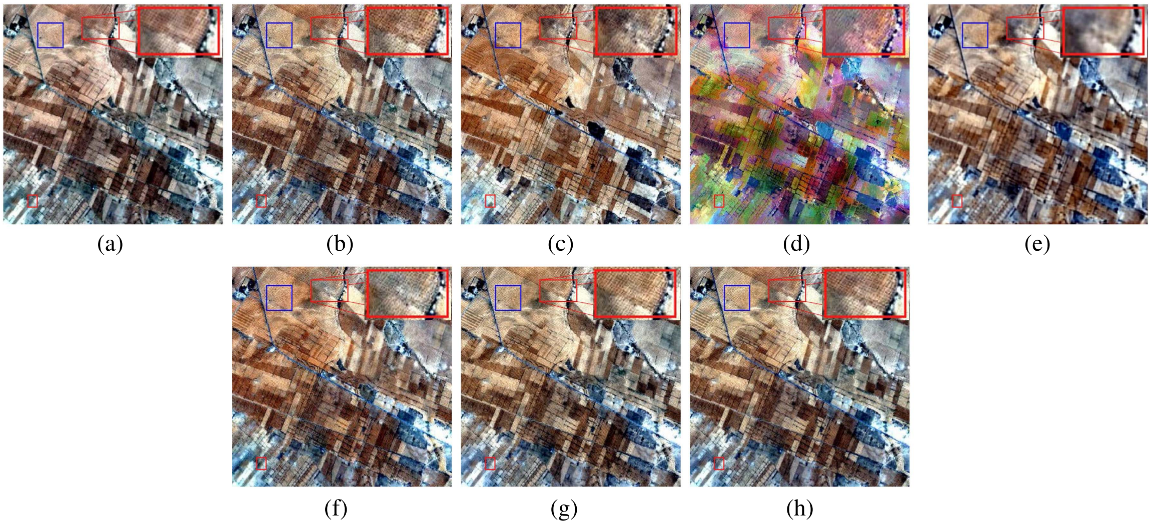

Fig. 6. Comparison of fused images from the considered methods on Sfax dataset. All images are at 10 m resolution. The images are displayed in the natural image composite (RGB: B4, B3, and B2). (a) Planetscope (P_t). (b) Reference (S_t). (c) S_{t-1}. (d) STARFM. (e) SR-CNN. (f) StfNet. (g) GAN-STFM. (h) STA-Net (Ours).

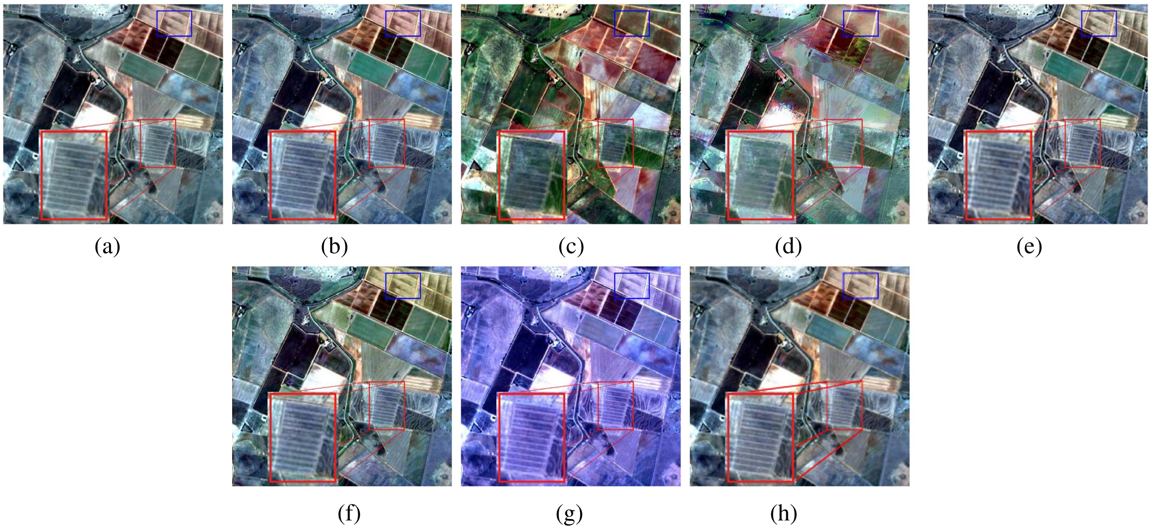

Fig. 7. Comparison of fused images from the considered methods on Coleambally dataset. All images are at 10 m resolution. The images are displayed in the natural image composite (RGB: B4, B3, and B2). (a) Planetscope (P_t). (b) Reference (S_t). (c) S_{t-1}. (d) STARFM. (e) SR-CNN. (f) StfNet. (g) GAN-STFM. (h) STA-Net (Ours).

regions (blue rectangle) even if the color is better preserved than the latter on both datasets. GAN-STFM offers better details reconstruction on Sfax dataset (highlighted in red) but generates a significant spectral distortion on Coleambally dataset as the color is bluer on the whole image compared with the reference image. The proposed method STA-Net achieves better fusion accuracy than the aforementioned methods in terms of spatial and spectral quality, as our fusion result is the closest to the reference image without any noticeable distortion or artifacts within the fused product in terms of visual quality. Besides, the colors are well-preserved by the proposed method than SR-CNN and StfNet, as we can see that the color in StfNet is different from the original image (blue rectangle). In terms of spatial information, StfNet lacks structural detail because its shallow architecture does not allow capturing sufficient high-frequency details, in particular, in heterogeneous areas and the edges of the areas highlighted in red. Regarding the proposed technique, since the network is deeper and benefits from the attention mechanism to select the best features, the details and contours are well reconstructed with sharper edges (e.g., regions highlighted in red). From the above-mentioned comparisons, it can be concluded that the two-stream strategy via attention mechanism can boost the fusion performance to produce more accurate products.

G. Limitations of the Proposed Method

The proposed method provides very competitive results not only for spectral–temporal fusion but also for generating data with high spatial and temporal resolutions (cf. Section 1.2 of the supplementary material). These results show the extensibility aspect of our work to deal with different satellite image fusion problems. However, as with the majority of works, the proposed method is subject to some limitations. Mainly regarding the applicability of the proposed method to fuse other modalities and different kinds of data. One can cite the behavior of our method to generate dense time series of nonreflected data, such as land surface temperature (LST) one, that includes thermal bands [60], [61], as it is recommended for climate change monitoring applications.

V. Conclusion

In this article, we introduced an STA-Net, an end-to-end two-stream fusion technique based on RABs via an effective loss function to integrate Planetscope and Sentinel-2 images. The proposed approach includes two stages. In the first stage, based on RABs, TDE predicts the temporal residual between the actual Sentinel-2 at the desired and prior dates. Simultaneously, the RDE estimates reflectance difference between Sentinel-2 and Planetscope images. Hence, two intermediate fused images are produced by injecting the corresponding temporal and reflectance differences, respectively. The second stage aimed to reconstruct the desired fused product via a learned weighting sum to combine the two-stream outcomes. An effective loss is introduced that involves the two-stream outputs to guarantee the best performances. To the best of our knowledge, this is the first attempt to fuse Planetscope and Sentinel-2 images to produce daily Sentinel-2 images using such a network.

The experiments have been conducted on Planetscope and Sentinel-2 images using quantitative and qualitative evaluations on two datasets; it was shown that the proposed approach yielded the best fusion performances in terms of spatial and spectral information compared with the considered state-of-the-art techniques. In our future work, we intend to explore more advanced deep-learning models to ameliorate the fusion quality further while making the product more realistic. Besides, we plan to extend our approach to be capable of generating fused images at Planetscope 3 m resolution. Also, we intend to extend STA-Net applicability to produce LST data for dynamic monitoring and prediction in climate change tasks.

References

[1] H. Ghassemian, "A review of remote sensing image fusion methods," *Inf. Fusion*, vol. 32, pp. 75–89, 2016.

[2] G. Vivone et al., "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015.

[3] X. Zhu, F. Cai, J. Tian, and T. K.-A. 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.

[4] M. Belgiu and A. Stein, "Spatiotemporal image fusion in remote sensing," *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 818.

[5] Q. Zhou et al., "Monitoring landscape dynamics in central U.S. grasslands with harmonized landsat-8 and sentinel-2 time series data," *Remote Sens.*, vol. 11, no. 3, 2019, Art. no. 328.

[6] Z. Zhu and C. E. Woodcock, "Continuous change detection and classification of land cover using all available landsat data," *Remote Sens. Environ.*, vol. 144, pp. 152–171, 2014.

[7] M. E. D. Chaves, M. C. A. Picoli, and I. D. Sanches, "Recent applications of landsat 8/OLI and sentinel-2/MSI for land use and land cover mapping: A systematic review," *Remote Sens.*, vol. 12, no. 18, 2020, Art. no. 3062.

[8] M. González-Sanpedro, T. Le Toan, J. Moreno, L. Kergoat, and E. Rubio, "Seasonal variations of leaf area index of agricultural fields retrieved from landsat data," *Remote Sens. Environ.*, vol. 112, no. 3, pp. 810–824, 2008.

[9] R. Houborg and M. F. McCabe, "A cubesat enabled spatio-temporal enhancement method (CESTEM) utilizing planet, landsat and MODIS data," *Remote Sens. Environ.*, vol. 209, pp. 211–226, 2018.

[10] G. Panteras and G. Cervone, "Enhancing the temporal resolution of satellite-based flood extent generation using crowdsourced data for disaster monitoring," *Int. J. Remote Sens.*, vol. 39, no. 5, pp. 1459–1474, 2018.

[11] J. M. A. Duncan, J. Dash, and P. M. Atkinson, "The potential of satellite-observed crop phenology to enhance yield gap assessments in smallholder landscapes," *Front. Environ. Sci.*, vol. 3, 2015, Art. no. 56.

[12] D. S. Candra, "Deforestation detection using multitemporal satellite images," in *IOP Conf. Series, Earth Environ. Sci.*, vol. 500, no. 1, 2020, Art. no. 012037.

[13] Z. Wang et al., "Monitoring land surface albedo and vegetation dynamics using high spatial and temporal resolution synthetic time series from landsat and the MODIS BRDF/NBAR/albedo product," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 59, pp. 104–117, 2017.

[14] B. Chen, B. Huang, and B. Xu, "Comparison of spatiotemporal fusion models: A review," *Remote Sens.*, vol. 7, no. 2, pp. 1798–1835, 2015.

[15] 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.

[16] 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.

[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. G. P. W. 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] 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] Y. Li, J. Li, and Z. Shaoquan, "A extremely fast spatio-temporal fusion method for remotely sensed images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 4452–4455.

[21] 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.

[22] 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.

[23] X. Liu, C. Deng, S. Wang, G.-B. 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.

[24] 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, 2015.

[25] 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.

[26] X. Meng, H. Shen, L. Zhang, Q. Yuan, and H. Li, "A unified framework for spatio-temporal-spectral fusion of remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 2584–2587.

[27] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[28] 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.

[29] 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.

[30] 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, Jan. 2021, Art. no. 5601413.

[31] L. Zhou, Z. Wang, Y. Luo, and Z. Xiong, "Separability and compactness network for image recognition and superresolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3275–3286, Nov. 2019.

[32] J. Lee et al., "FBRNN: Feedback recurrent neural network for extreme image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 2021–2028.

[33] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 105–114.

[34] Y. Zhang, K. Li, Kai Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 286–301.

[35] R. Eldan and O. Shamir, "The power of depth for feedforward neural networks," in *Proc. 29th Annu. Conf. Learn. Theory*, 2016, pp. 907–940.

[36] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[37] S. Garg, S. Peitz, U. Nallasamy, and M. Paulik, "Jointly learning to align and translate with transformer models," 2019, *arXiv:1909.02074*.

[38] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 4960–4964.

[39] F. Wang and D. M. Tax, "Survey on the attention based RNN model and its applications in computer vision," 2016, *arXiv:1601.06823*.

[40] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature Rev. Neurosci.*, vol. 3, no. 3, pp. 201–215, 2002.

[41] S. Chaudhari, G. Polatkan, R. Ramanath, and V. Mithal, "An attentive survey of attention models," 2019, *arXiv:1904.02874*.

[42] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[43] W. Zhang, J. Li, and Z. Hua, "Attention-based tri-UNet for remote sensing image pan-sharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3719–3732, 2021.

[44] D. Chung, K. Tahboub, and E. J. Delp, "A two stream siamese convolutional neural network for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1992–2000.

[45] Y. Zhu, Z. Lan, S. Newsam, and A. Hauptmann, "Hidden two-stream convolutional networks for action recognition," in *Proc. Asian Conf. on Comput. Vis.*, New York, NY, USA: Springer, 2018, pp. 363–378.

[46] X. Liu, Q. Liu, and Y. Wang, "Remote sensing image fusion based on two-stream fusion network," *Inf. Fusion*, vol. 55, pp. 1–15, 2020.

[47] J. Wei, Y. Xu, W. Cai, Z. Wu, J. Chanussot, and Z. Wei, "A two-stream multiscale deep learning architecture for pan-sharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5455–5465, Sep. 2020.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[49] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 136–144.

[50] J.-H. Kim, J.-H. Choi, M. Cheon, and J.-S. Lee, "Mamnet: Multi-path adaptive modulation network for image super-resolution," *Neurocomputing*, vol. 402, pp. 38–49, 2020.

[51] A. G. Howard et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[52] A. Fernández-Manso, O. Fernández-Manso, and C. Quintano, "Sentinel-2a red-edge spectral indices suitability for discriminating burn severity," *Int. J. Appl. Earth Observation Geoinf.*, vol. 50, pp. 170–175, 2016.

[53] M. Liu, T. Wang, A. K. Skidmore, and X. Liu, "Heavy metal-induced stress in rice crops detected using multi-temporal sentinel-2 satellite images," *Sci. Total Environ.*, vol. 637–638, pp. 18–29, 2018.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[55] A. G. Mahyari and M. Yazdi, "Panchromatic and multispectral image fusion based on maximization of both spectral and spatial similarities," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 6, pp. 1976–1985, Jun. 2011.

[56] Z. Wang, 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, no. 4, pp. 600–612, Apr. 2004.

[57] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[58] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2472–2481.

[59] M. Deshmukh et al., "Image fusion and image quality assessment of fused images," *Int. J. Image Process.*, vol. 4, no. 5, 2010, Art. no. 484.

[60] Z. Yin et al., "Spatiotemporal fusion of land surface temperature based on a convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1808–1822, Feb. 2021.

[61] P. Wu et al., "Spatially continuous and high-resolution land surface temperature product generation: A review of reconstruction and spatiotemporal fusion techniques," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 3, pp. 112–137, Sep. 2021.