Modality Translation in Remote Sensing Time Series

Xun Liu, *Student Member, IEEE*, Danfeng Hong, *Senior Member, IEEE*, Jocelyn Chanussot, *Fellow, IEEE*, Baojun Zhao, *Member, IEEE*, and Pedram Ghamisi, *Senior Member, IEEE*

***Abstract*— Modality translation, which aims to translate images from a source modality to a target one, has attracted a growing interest in the field of remote sensing recently. Compared to translation problems in multimedia applications, modality translation in remote sensing often suffers from inherent ambiguities, i.e., a single input image could correspond to multiple possible outputs, and the results may not be valid in the following image interpretation tasks, such as classification and change detection. To address these issues, we make the attempt to utilizing time-series data to resolve the ambiguities. We propose a novel multimodality image translation framework, which exploits temporal information from two aspects: 1) by introducing a guidance image from given temporally neighboring images in the target modality, we employ a feature mask module and transfer semantic information from temporal images to the output without requiring the use of any semantic labels and 2) while incorporating multiple pairs of images in time series, a temporal constraint is formulated during the learning process in order to guarantee the uniqueness of the prediction result. We also build a multimodal and multitemporal dataset that contains synthetic aperture radar (SAR), visible, and short-wave length infrared band (SWIR) image time series of the same scene to encourage and promote research on modality translation in remote sensing. Experiments are conducted on the dataset for two cross-modality translation tasks (SAR to visible and visible to SWIR). Both qualitative and quantitative results demonstrate the effectiveness and superiority of the proposed model.**

***Index Terms*— Feature mask module, image time series, inherent ambiguities, modality translation, remote sensing.**

I. Introduction

A. Background

IN REAL-WORLD data processing problems, it is always possible to have access to images from various modalities associated with a certain content [1]–[3], including RGB images, infrared images, or multispectral images, and so forth. These images often share some common presentations, such as main edges, textures, and other structure primitives [4], and, thus, represent the opportunity to simulate corresponding images from each other [5], [6]. To this end, modality translation aiming at learning a mapping to translate images from a source modality to a target one has become a quite promising topic and been successfully applied in a wide range of tasks, such as image super-resolution [7]–[10], denoising [11], and feature learning [2], [12].

Due to the increased availability of remote sensing data in recent years [13]–[15], modality translation has also shown considerable success in remote sensing. These approaches transform images from one modality to another using paired or unpaired data, which achieves the availability of multiple image modalities of the same scene, and, to some extent, help to overcome the limitations of satellite revisiting cycles and sensor capabilities [16]–[21]. For instance, Peng *et al.* [16] extended the swath of hyperspectral data by predicting hyperspectral images from the corresponding large-coverage multispectral ones. In [17], the effects of cloud cover in multispectral images were mitigated through synthetic aperture radar (SAR) to optical image translation. Furthermore, modality translation represents an opportunity to achieve appropriate interpretation of heterogeneous images at the data level. For example, SAR images have the ability to provide day and night Earth observations and overcome various undesired weather conditions [22]. However, since human eyes are not familiar with the distance-dependent radar imaging mechanism [23], even experts still have limited ability to interpret these images. Translating SAR to optical images, therefore, facilitates human visual interpretation and largely supports the analysis of the original data [18]–[20].

To date, modality translation has attracted much research interest in the field of remote sensing, and the results have been widely used in the following remote sensing image processing tasks, such as land cover classification [24]–[27] and change detection [28].

B. Motivation

Despite the success of modality translation, it should be mentioned that it still faces the great challenge of translation ambiguity [29]. Given one image in the source modality, modality translation may not provide sufficient information to reconstruct the corresponding image in the target one, and naturally, a single input image may correspond to multiple

Manuscript received August 2, 2020; revised February 27, 2021; accepted April 1, 2021. Date of publication May 21, 2021; date of current version January 5, 2022. This work was supported in part by MIAI@Grenoble Alpes under Grant ANR-19-P3IA-0003. *(Corresponding author: Danfeng Hong.)*

Xun Liu and Baojun Zhao are with the Beijing Key Laboratory of Embedded Real-Time Information Processing Technology, School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: liuxun@bit.edu.cn; zbj@bit.edu.cn).

Danfeng Hong is with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Weßling, Germany, and also with the GIPSA-lab, Centre national de la recherche scientifique (CNRS), Grenoble Institute of Technology (Grenoble INP), Université Grenoble Alpes, 38000 Grenoble, France (e-mail: danfeng.hong@dlr.de).

Jocelyn Chanussot is with the Laboratoire Jean Kuntzmann (LJK), Inria, Centre national de la recherche scientifique (CNRS), Grenoble Institute of Technology (Grenoble INP), Université Grenoble Alpes, 38000 Grenoble, France (e-mail: jocelyn@hi.is).

Pedram Ghamisi is with the Machine Learning Group, Helmholtz Institute Freiberg for Resource Technology, Helmholtz-Zentrum Dresden-Rossendorf, 09599 Freiberg, Germany, and also with the Institute of Advanced Research in Artificial Intelligence (IARAI), 1030 Vienna, Austria (e-mail: p.ghamisi@gmail.com).

Digital Object Identifier 10.1109/TGRS.2021.3079294

possible outputs. As a result, many modality translation problems are generally ill-posed and can have infinitely many solutions [30].

To address this issue, the most common approach is to distill the ambiguity of the mapping step in random noise or a low-dimensional latent vector and model a distribution of possible outputs instead of a single output [31]. At interference time, this distribution is randomly sampled to generate diverse outputs, while it still remains faithful to the input. For example, by introducing some random noise or latent codes, a sketch of shoes can be mapped to photographs of shoes with a variety of colors and textures [32], which do not actually exist in the input image. In this way, translation models, which produce results that are both diverse and realistic, largely address the ambiguities, especially for some multimedia applications [30], [31], [33].

The abovementioned strategy, in general, is still problematic for modality translation in remote sensing. The reason could be that, to achieve a reliable image interpretation, it is crucial to translate images as accurately as possible in remote sensing. Nevertheless, the stochastic effect (e.g., a sample from the distribution), which leads to uncertainties in translation, is not desirable and can be harmful to the subsequent analysis (e.g., decision-making) [2], [34]. Hence, it is still a challenging problem to resolve the ambiguities in modality translation. One possible solution to this issue would be incorporating some additional information.

On the other hand, we know that, in remote sensing, there is always a capability to acquire multitemporal images from the Earth's surface with repeated observations by satellites. These multitemporal observations, which aim at monitoring land cover dynamics, present similar contexts at the same geographic area and follow smooth variations if no abrupt change occurs [35], [36]. For a remote sensing image translation task, it is valid to assume that we have side information in the form of temporal correlation among image time series. Instead of only exploiting the input image to generate the corresponding outputs, it might be beneficial to incorporate time-series data to aid the learning process to achieve more precise results in translation. For instance, in the popular SAR to optical translation problem, besides the input SAR image, additional optical images covering the same geographic region yet captured at different dates may also be provided in the historical dataset, and one can possibly leverage such side information to achieve better modality translation.

C. Proposed Model

Motivated by the above realization, in this work, we make the attempt to resolve the ambiguities with temporal side information. Specifically, we incorporate multimodality image time series in translation and propose a temporal information-guided remote sensing image translation (TIRSIT) framework. The main contributions of this article are given as follows:

1) We explore a new problem, i.e., modality translation in remote sensing image time series, as opposed to traditional translation approaches that only rely on input images to generate cross-modality outputs. The proposed method, instead, has access to the knowledge from multitemporal images in source and target modalities, and therefore, it addresses the performance barrier in existing algorithms.
2) We propose a TIRSIT framework to make use of remote sensing time-series data in translation. In the framework, we adopt a feature mask module to capture semantic information in a guidance image for translation. Moreover, given multiple temporal images, we formulate a uniqueness constraint in network learning, leading to a more convincing final result.
3) We build a multimodal and multitemporal remote sensing image dataset, which contains multiple SAR, visible, and short-wave length infrared band (SWIR) images of the same scene taken on the same dates. Extensive experiments are performed on the dataset for two cross-modality translation tasks (SAR to visible and visible to SWIR), and both visual and quantitative results demonstrate that the proposed model can successfully benefit from temporal information and outperform the existing approaches.

This work is an extension of the conference version [37] presented earlier.[1] This work adds values to the initial version in several significant ways. First, we investigate more analysis of the ambiguities problem in image translation both for multimedia and remote sensing and, thus, further clarify the motivation of this work. Second, we improve the proposed framework by introducing a semantic guidance module, in which the feature mask module is elaborately designed to utilize the domain knowledge of the output while avoiding overfitting. Third, the network architecture is modified to learn the cross-modality mapping in translation, including introducing the popular U-Net architecture and an adversarial learning module. In addition, more experiments are conducted to confirm the superiority of the proposed model, and considerable new analyses and intuitive explanations are also added to the results.

D. Article Outline

The remainder of this article is organized as follows. In Section II, we review the related works. Section III formulates the problem of modality translation in remote sensing time series and gives detailed descriptions and analyses of the proposed framework. Section IV includes the experimental results and discussions. The conclusion is drawn in Section V.

II. Related Work

A. Generative Adversarial Networks

With the advancement in deep learning [38], recent years have witnessed great progress in generative adversarial networks (GANs) [39]–[42]. Proposed by Goodfellow *et al.* [43], GANs consist of two main competing

[1]The work was presented earlier in the Student Paper Competition during the 2019 IGARSS conference in Yokohama, Japan, and, eventually, won the First Place (IEEE Mikio Takagi Prize) from over 300 applications.

components, i.e., generator and discriminator, which plays a two-player minimax game to generate samples from a data distribution. On the one hand, the generator learns to map from a latent space to the data distribution of interest and then samples from the distribution as outputs. On the other hand, a discriminator is trained to distinguish between real data drawn in the true data distribution from fake samples produced by the generator. The learning procedure continues until the generator can produce samples that are very close to the real data that can "fool" the discriminator, whereas the discriminator can only make a nearly random decision [39]. In this way, the distribution of the generated images is enforced to match that of the real data through learning, and therefore, more realistic fake images can be achieved. Conditional GANs [44], as a particular case, generate fake data from a distribution conditioned on the input contexts, such as text [45], audio [46], image [47], and video [48]. Compared to the previous generative models, GANs have achieved in synthesizing visually appealing images and been successfully applied in many image generation tasks, including text-to-image synthesis [45], image super-resolution [47], domain adaption [49], and image translation [50].

B. Image Translation

Image translation is the task of learning a procedure to map images from a source domain to a target one. A milestone work called *pix2pix* was proposed by Isola *et al.* [51], where a translation function is learned from input to output image domains with conditional GANs. Inspired by *pix2pix*, some works further adapted it to a variety of relevant tasks and achieved high-quality results, including labels to facades, sketches to photographs, and gray to color images [52]–[55].

In the remote sensing community, a number of works related to image translation have also been developed recently. Ghamisi and Yokoya [56] and Paoletti *et al.* [57] simulated elevation data, i.e., digital surface model (DSM), which contain rich height information from a single color image with paired or unpaired approaches. Ao *et al.* [58] proposed a dialectical GAN combining Wasserstein GAN-gradient penalty and spatial gram matrices to synthesize a high-resolution TerraSAR-X image from the corresponding Sentinel-1 SAR image. Bermudez *et al.* [17] presented an algorithm for the translation from SAR to optical images in order to recover regions that are covered by clouds. Similarly, Wang *et al.* [19] proposed a supervised cycle-consistent GAN to generate large optical images from the SAR images. Fuentes Reyes *et al.* [20] explored the value of the empirical knowledge for initialization of a conditional GAN in SAR to optical image translation and gave a detailed discussion about the opportunities and drawbacks related to the application.

C. Image Translation With Side Information

Learning with side information means that there is additional knowledge available in training and testing, which can be used to improve the performance of the model. In image translation tasks, side information always appears as a guidance image. In this way, an input image is translated into another while respecting the constraints specified by an external guidance image, such as generating a photograph from a sketch, guided with a texture patch, and facial landmark available in a face expression translation task.

Nowadays, various forms of utilizing the guidance image in translation have been proposed in the literature. The most straightforward way is to directly concatenate the source and guidance images at the input along the channel dimension, followed by a conventional image translation model. One can also concatenate the source and guidance images at the feature level for subsequent processing. Moreover, several efforts have also been made using other strategies. Featurewise linear modulation (FiLM) [59] utilizes conditioning information to generate scaling and shifting parameters and applies the featurewise affine transformation to the norm layers in the original translation network (input to output). Inspired by this work, Dumoulin *et al.* [65] proposed a conditional instance normalization (CIN) model, in which the scaling and shifting parameters are learned from the guidance image. To better utilize the information from guidance images, AlBahar *et al.* [60] further proposed a bidirectional feature transformation (bFT) scheme, in which information could also flow from the input image back to the guidance.

III. METHODOLOGY

In this section, we present the details of the proposed TIRSIT framework. The goal is to realize modality translation in remote sensing time series, which is the translation of images from two modalities while leveraging multiple available image pairs covering the same geographic area. We first consider high redundancies among image time series and focus on the temporal evolution of information in the translation. Then, we develop a multistream-guided translation network, which exploits the temporal information with semantic guidance and uniqueness constraint. Finally, with the learned translation network, we predict output images from the corresponding inputs and reconstruct the final target image.

A. Problem Formulation

In this work, we aim at translating an image from one modality to another while respecting constraints from the time-series data. As we know, in remote sensing, sensors scan the same geographical region repeatedly. There are always image pairs available in the source and target modalities acquired at different dates in the historical dataset. These images present similar contexts in the same geographic area and show considerable temporal correlation. The goal of the presented method is to exploit the extra guidance knowledge from temporal images to translate images covering the same geographical area from the source modality to the target one. To be specific, we mainly consider a typical case (as shown in Fig. 1), in which two pairs of images in the source and target modalities (S_j and T_j; S_k and T_k) acquired at different dates t_k and t_j are utilized as side information to translate an input image S_i in the source modality that is captured at date t_i to the corresponding image T_i in the target modality, which is captured on the same date t_i.

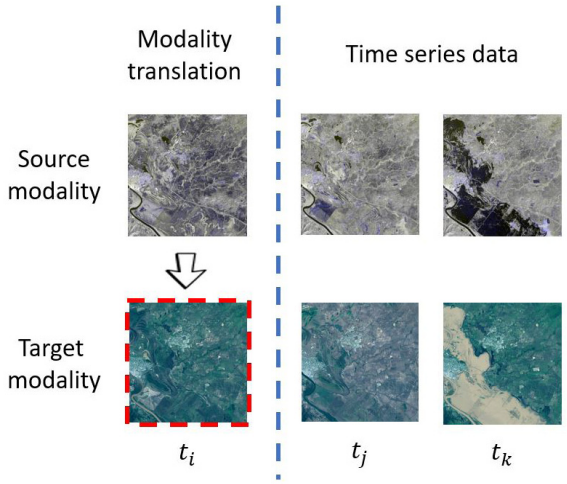

Fig. 1. Illustration of modality translation in remote sensing time series. The goal is to predict the image acquired at t_i in the target modality from the corresponding image in the source modality, as well as two pairs of external images in the source and target modalities acquired at t_k and t_j.

To tackle this problem, we assume that a limited part of the image changes in time series and mainly focus on the temporal evolution of the information instead of the original data in translation. The proposed method will establish a complex mapping between a difference image in the source modality and its corresponding difference image in the target modality at the same period (e.g., the difference image S_{kj} and T_{kj} from t_k to t_j in the source and target modalities). Then, given the available difference images S_{ij} and S_{ik} in the source modality, we translate them to their corresponding difference images T_{ij} and T_{ik} through the learned mapping and reconstruct the final result T_i with the original data (i.e., T_j and T_k) in the target modality.

B. Multistream-Guided Translation Network

In this section, we propose a multistream-guided translation network, as shown in Fig. 2, which can be used to build a cross-modality mapping between difference images, with the guidance of a temporally neighboring image in the target modality. In what follows, we will discuss two highlights in the network design, i.e., semantic guidance and uniqueness constraint to utilize the time-series data, and also present the network architecture and the learning procedure.

1) Semantic Guidance: To translate a difference image from the source modality to the target one in remote sensing, it could be beneficial for us to involve some temporally neighboring images in the target modality since they typically provide prior information in the same geographic area and reflect the desired visual effects or constraints of images in the target modality. The main goal is to incorporate such additional guidance into the translation model and leverage the prior information to achieve controllable translation.

As mentioned in Section II-C, there are various strategies to utilize guidance images for the translation task in the literature. The main technical question is how the guidance image is used to affect the processing of the input source image. For the concentration schemes (concentrate input and guidance images at the data or feature level), it is assumed that the guidance and input images are equally important in translation, and the concentration schemes allow a free information flow from the guidance to the output. While being parameter efficient, these approaches generally suffer from the overfitting problem if directly concentrating the difference image in the source modality with the temporal image in the target modality to generate a difference image in the target modality. The reason is that the correlation between images in the same modality (covering the same geographic area) is much stronger than the cross-modality correlation. Therefore, it is always easier for the network to learn a mapping between the guidance and the output (both in the target modality), and thus, the information flow from input to the output might not be activated. With regard to other conditioning schemes (e.g., Film, CIN, and bFT), they mainly focus on applying a featurewise affine transformation conditioned on the guidance image, and thus, the external information could be formulated as the scale and shift parameters in the transformation. All these works having in common that the guidance image influences the visual effects of the output image instead of fine-grained spatial details are reasonable in the context of style transfer. However, in the task of guided translation with temporal images, it is nontrivial to exploit the structural information in the guidance image, and directly applying such conditioning methods may not give satisfying results.

In this study, we propose to explore semantic guidance in the translation model to better utilize the available information from the temporal guidance image. Unlike the existing schemes, the proposed approach explicitly supports the communication of semantic information from the guidance to the output and leverages semantic labels as an additional form of supervision in translation. This kind of semantic label enables the generative model to receive the prior knowledge (i.e., semantics about the scene understanding) in the target modality from the given guidance image, rather than only gain information from the cross-modality input.

As we know, for image translation tasks, ground-truth semantic labels are not easy to obtain since they usually require manual annotations. Instead, inspired by Ma *et al.* [61], we present a feature mask module to compute an approximate estimation of semantic categories without using any ground-truth semantic labels.

Fig. 3 illustrates the feature mask module in the nth layer. We first normalize the feature maps from the guidance image f^n_{guide} into $\hat{f}^n_{\text{guide}}$

$$\hat{f}^n_{\text{guide}} = \frac{f^n_{\text{guide}} - \text{mean}\left(f^n_{\text{guide}}\right)}{\text{std}\left(f^n_{\text{guide}}\right)} \tag{1}$$

where mean(·) and std(·) are the operators to calculate mean and standard deviation of feature maps in each channel. We believe that it is the spatial patterns in feature maps that guide the cross-modality image translation, rather than the norm of the features. The usage of normalization standardizes feature maps and helps them to contain more clear semantic patterns compared with the original feature maps.

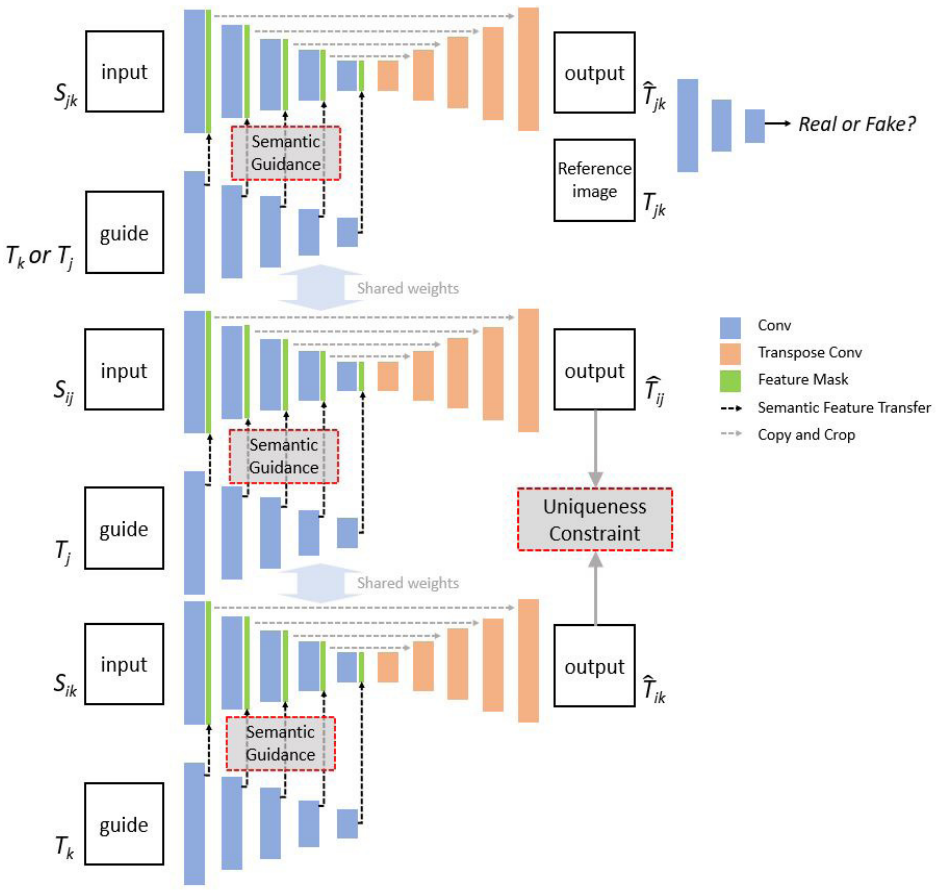

Fig. 2. Illustration of the proposed multistream-guided translation network. T_i, T_j, and T_k denote the images in the target modality acquired at t_i, t_j, and t_k, respectively. S_{ij}, S_{ik}, and S_{jk} are the difference images in the source modality from t_i to t_j, t_i to t_k, and t_j to t_k, respectively. $\hat{T}_{ij}$, $\hat{T}_{ik}$, and $\hat{T}_{jk}$ are the corresponding predicted images from the translation network. T_{ij} is the difference images from t_i to t_j in the target modality.

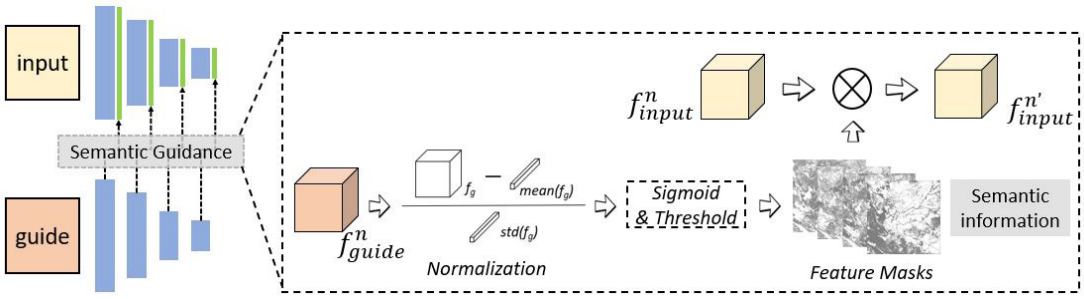

Fig. 3. Illustration of the semantic guidance module in the proposed multistream-guided translation network. f^n_{input} and f^n_{guide} are the nth layer of features of the input and guidance images, respectively. $f^{n'}_{\text{input}}$ is the semantic-guided nth layer of features of the input image.

Then, the feature masks from the given guidance image can be computed as

$$\text{mask}^n_{\text{guide}} = (1-\eta)\cdot\sigma\left(\hat{f}^n_{\text{guide}}\right)+\eta \tag{2}$$

where $\sigma(\cdot)$ denotes a nonlinear activation function, which is usually a sigmoid function in applications, and η is a threshold. In this way, the feature masks could be trained to estimate the semantic categories in an unsupervised way and contain substantial semantic information provided by the guidance image.

Finally, we incorporate such semantic maps into the translation model, in which the feature maps of the input image f^n_{input} are elementwise multiplied with the feature masks

$$f^{n'}_{\text{input}} = \text{mask}^n_{\text{guide}}\cdot f^n_{\text{input}} \tag{3}$$

where $f^{n'}_{\text{input}}$ is the updated feature maps of the input image in the nth layer.

We pointed out that semantic guidance could also be thought of as a kind of attention module that decouples different semantic regions to handle them separately in translation. This module exploits multiscale semantic information in the guidance image while applying it to different layers.

The estimated label maps are then used as additional supervision to help translation in the corresponding semantic regions.

2) Uniqueness Constraint: In the remote sensing modality translation problem, we aim at translating temporal changes from one modality to another while respecting additional guidance. Since multiple image pairs in the source and target modalities are involved as temporal side information, there could be multiple final results in the translation. For example, as shown in Fig. 2, there are two difference images (S_{ij} and S_{ik}) in the source modality with the guidance images (T_j and T_k) to predict their corresponding difference images (T_{ij} and T_{ik}). Two final target images would be produced by the predicted $\hat{T}_{ij}$ and $\hat{T}_{ik}$, respectively. It is obvious that these results, which are derived from different temporal images, should not be far from each other. It is necessary for us to guarantee the uniqueness of the final result.

Considering this kind of temporal information, we introduce a uniqueness constraint in learning, which is formulated as follows:

$$\begin{aligned}\hat{T}_i &= \hat{T}_{ij} + T_j \\ &= \hat{T}_{ik} + T_k \end{aligned} \tag{4}$$

where we have the final result $\hat{T}_i$ with the predicted difference images ($\hat{T}_{ij}$ and $\hat{T}_{ik}$) and the original data (T_i and T_j). With the uniqueness constraint, the translation network should not be trained independently of each pair of difference images. Instead, the multistream-guided translation network can be jointly learned with multiple pairs of images under the constraint and is encouraged to generate consistent translation results, avoiding the influence from the instability of global optimization algorithms or highly nonlinear deep networks. Consequently, the time-series data information could be further utilized in the learning to produce better results.

3) Network Architecture: In the proposed multistream-guided translation network, we adopt the popular U-Net architecture to build the base translation model. U-Net is a fully convolutional encoder–decoder network with shortcut connections that followed the architecture guidelines in [62] and has been adapted to stabilize the adversarial training process [63]. In the network, the encoding part maps an input image into a higher level feature representation, with several strided convolutional filters followed by a leaky ReLU activation function. Then, the decoder part mirrors the encoding network and invert the downsampling process to generate the output images. The U-Net also contains skip connections that concatenate spatial channels between mirrored layers in the encoder and decoder parts. In this way, low-level information between the input and output images can be transferred by these direct connections, instead of being lost in the bottleneck layer leading to severe degradation in output quality.

Note that, in the proposed network, we have two encoders E_{input} and E_{guidance} for the input and guidance images, respectively. Through the semantic guidance module, features in each layer of E_{input} could be influenced by the corresponding feature layers in E_{guidance}, thus allowing the information flow from the guidance to the output.

4) Network Learning: The main goal of the proposed translation network is to build a function to transform images from the source modality to the target one. To achieve that, we have a training objective that can be decomposed into three components: 1) the *reconstruction* loss, $\mathcal{L}_{\text{rec}}$, which encourages output images in the target modality to be closer to the reference; 2) the consistency loss, $\mathcal{L}_{\text{con}}$, which takes advantage of the time-series data; and 3) the adversarial loss, $\mathcal{L}_{\text{gan}}$, which improves the visual quality of the output images.

We adopt mean squared error to measure the difference between output and reference images in the reconstruction loss

$$\mathcal{L}_{\text{rec}} = \|\hat{T}_{jk} - T_{jk}\|_F \tag{5}$$

where T_{jk} is the available temporal difference image in the target modality; $\hat{T}_{jk}$ is the predicted image from the translation model; and $\|\cdot\|_F$ denotes the Frobenius norm.

Consistency loss helps to encourage an identical final result from multiple temporal images. It exploits the correlations among multiple translation outputs to accommodate the uniqueness constraint. From (4), it could be defined as

$$\mathcal{L}_{\text{con}} = \|\left(\hat{T}_{ij} - \hat{T}_{ik}\right) - T_{jk}\|_F. \tag{6}$$

We also apply the adversarial loss to enhance the visual quality of the output images and make them indistinguishable from the real ones. Instead of manually designed loss functions measuring the similarity, the adversarial loss utilizes a binary classifier, which is the discriminator to enforce a similar distribution between the outputs and references. We adopt a discriminator D (i.e., *Patch*GAN), to classify images with smaller patches, followed by an averaging operation. In this case, the adversarial loss can be written as

$$\begin{aligned}\mathcal{L}_{\text{adv}} = \mathbb{E}_{S_{jk},T_{jk}}\left[\log D\left(S_{jk}, T_{jk}\right)\right] \\ + \mathbb{E}_{S_{jk},\hat{T}_{jk}}\left[1 - \log D\left(S_{jk}, \hat{T}_{jk}\right)\right].\end{aligned} \tag{7}$$

C. Translation and Target Image Reconstruction

Once the training process of the proposed multistream-guided translation network is completed, we move on with the cross-modality image translation and target image reconstruction. With the learned model, we translate input images (S_{ij} and S_{ik}) in the source modality and guidance images (T_j and T_k) in the target modality to the corresponding images ($\hat{T}_{ij}$ and $\hat{T}_{ik}$) in the target modality. In this way, the final target images can be produced with the aid of the temporal images T_j and T_k in the target modality

$$\hat{T}_i^{j\to i} = \hat{T}_{ij} + T_j \tag{8}$$

$$\hat{T}_i^{k\to i} = \hat{T}_{ik} + T_k \tag{9}$$

where $T_i^{j\to i}$ and $T_i^{k\to i}$ are the predicted final target images from T_j and T_k, respectively.

Then, these two predicted target images are combined to generate the final result. We adopt a local weighting strategy for the combination

$$\hat{T}_i = p_j \times \hat{T}_i^{j\to i} + p_k \times \hat{T}_i^{k\to i} \tag{10}$$

where p_j and p_k are the weighting parameters for the two predicted results $\hat{T}^{j\to i}$ and $\hat{T}^{k\to i}$, respectively. To determine the weighting parameters in reconstruction, we believe that

a more similar image in the source modality leads to a more reliable translated image in the target modality [64]. Accordingly, the smoothed absolute difference maps are used to measure the local similarity between multitemporal images in the source modality, and then, a sigmoid function is adopted to calculate the weighting parameters

$$p_j = \frac{1}{1 + e^{-k \cdot \left(|f_s * S_{ik}| - |f_s * S_{ij}|\right)}} \tag{11}$$

where k is the parameter of the sigmoid function, which controls the curve shape, f_s is the average filter of size s, and $*$ denotes the convolution operation. We can also have

$$p_k = 1 - p_j. \tag{12}$$

Note that, although the proposed framework works on the typical case which involves two pairs of images in the source and target modalities (S_j and T_j; S_k and T_k) acquired at different dates, this is not a hard restriction. It can be also extended to exploit multiple pairs of images as a temporal information in the translation and target image reconstruction.

IV. Experimental Results and Analysis

In this section, we first build a multimodal and multitemporal dataset for modality translation in remote sensing time series. Then, extensive experiments are conducted on two cross-modality image translation tasks, i.e., SAR to visible and visible to SWIR image translation, to quantitatively and qualitatively evaluate the performance of the proposed model. Furthermore, we perform ablation studies and report more detailed analyses to illustrate the advantage of the semantic guidance and uniqueness constraint modules in the proposed framework.

A. Dataset

For the task of modality translation in remote sensing time series, multitemporal images with good spatial alignment affect image translation results significantly. However, to the best of our knowledge, there has not been such a multimodality image translation dataset with time-series data available for public usage. To this end, we prepared a dataset that contains multimodal remote sensing image time series, which is shown in Fig. 4, to evaluate the performance of different approaches.

In late March 2017, Tropical Cyclone Debbie crossed the eastern coast of Australia and then moved on to New Zealand. The heavy rainfall resulted in major floods all along its path. The prepared multimodal and multitemporal dataset mainly captured a rapid changing of the ground surface during the flooding water retreating. In the dataset, images were collected over the region of Queensland (23°32′ S, 105°28′ E) located in the northeast of Australia and taken on April 8, 2017, April 28, 2017, and August 31, 2017, by Sentinel-1 and Sentinel-2 satellites. Three modalities of images were involved, including SAR data provided by the C-band SAR instrument on the Sentinel-1 satellite, visible, and SWIR data from the MultiSpectral Instrument (MSI) on the Sentinel-2 satellite. These images were well-coregistered. The size of the images is 2048 × 2048 pixels covering an area of about 20 km × 20 km.

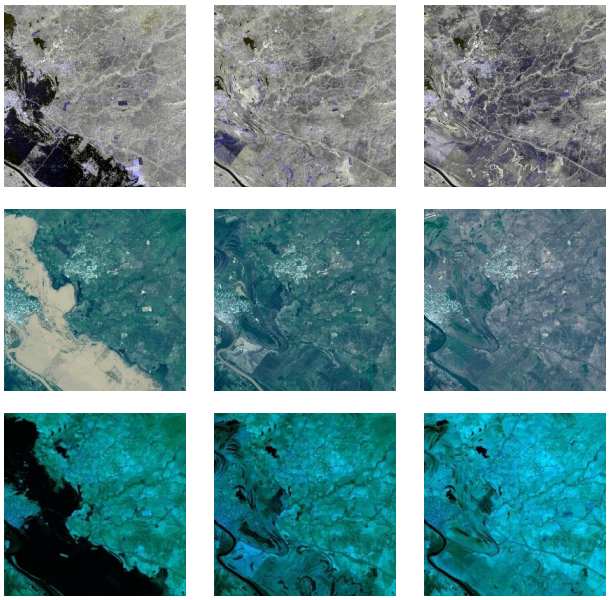

Fig. 4. Multimodal image time series of the same scene from Sentinel-1 and Sentinel-2 satellites. (Top to Bottom) Corresponding SAR, visible, and SWIR images. (Left to Right) Multimodal images taken on April 8, April 28, and August 31, 2017.

Herein, these multimodal and multitemporal images were downloaded from the Copernicus Open Access Hub (https://scihub.copernicus.eu/) and have been preprocessed using the Sentinel Application Platform (SNAP) software provided by ESA. For the SAR image series, we used intensities of dual-polarized (VV and VH) data provided in the ground range detected (GRD) format as SAR images and followed a processing flowchart of calibration, speckle filtering with the Lee sigma filter, and the range-Doppler terrain correction. On the other hand, RGB (bands 4, 3, and 2) and SWIR (bands 11 and 12) channels in Sentinel-2 multispectral data were selected to create the visible and SWIR images. Then, all the images were converted to the uniform projection of latitude and longitude coordinates and resampled at a ground sampling distance (GSD) of 10 m.

B. Experimental Setup

In the experiments, the goal is to predict an image in the target modality given the corresponding images in the source modality, as well as two pairs of temporal images in the source and target modalities as the time-series data. To meet the problem setting, in the dataset, the image T_2 in the target modality acquired at date t_2 is used as the target image, which is predicted from the corresponding image S_2, while the image pairs S_1 and T_1 at t_1 and S_3 and T_3 at t_3 are employed as temporal information.

To illustrate the effectiveness of the proposed model, we first compare our results with temporally neighboring images T_1 and T_3 in the target modality. This test mainly aims at analyzing the benefits of cross-modal information. The proposed models are also compared with a traditional translation algorithm named *pix2pix* in [51], which represents the baseline. We adopt the default parameters given by authors in the experiments to ensure a fair comparison.

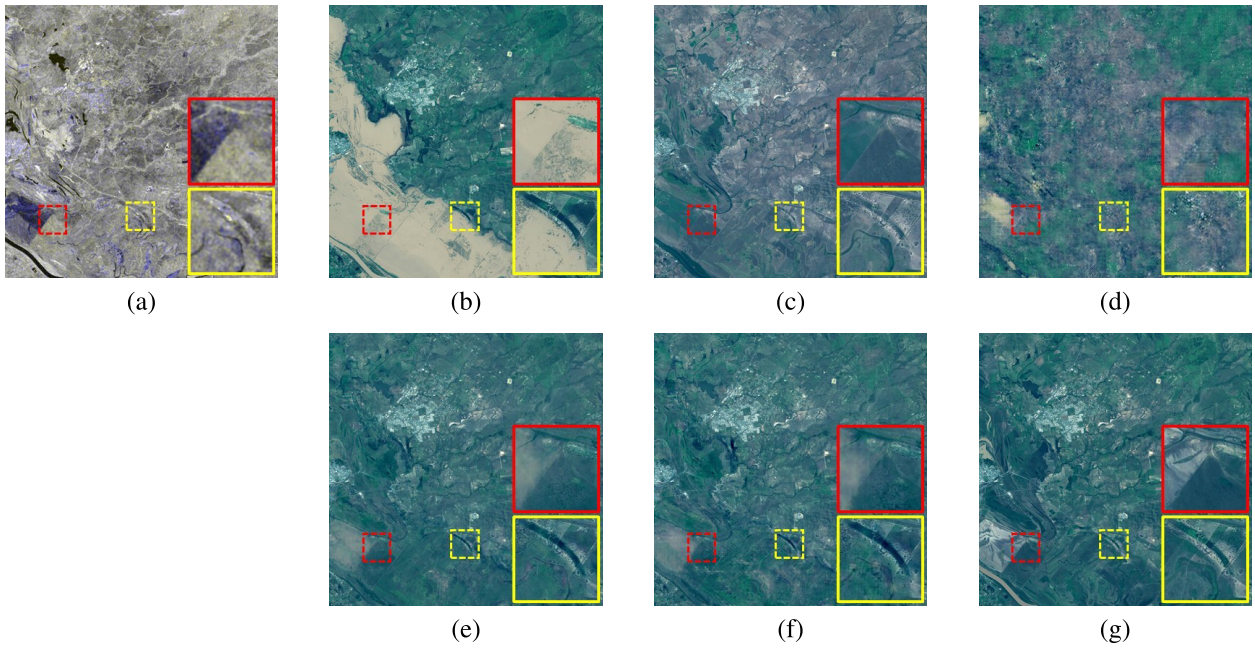

Fig. 5. Qualitative results in the task of the SAR to visible image translation. (a) Input SAR image S_2 taken at t_2. (b) Observed visible image T_1 taken at t_1. (c) Observed visible image T_3 taken at t_3. (d) Predicted result obtained by *pix2pix* model [51]. (e) Predicted result obtained by TIRSIT w/o adversarial loss. (f) Predicted result obtained by TIRSIT. (g) Observed target visible image T_2 taken at t_2.

Regarding quantitative comparison, the availability of the target image T_2 allows us to evaluate the final result in a full referenced manner. We adopt three widely used metrics, root mean square error (RMSE), correlation coefficient (CC), and structural similarity (SSIM) for evaluation.

C. Task 1: SAR to Visible Image Translation

As we know, SAR images have the ability to provide day and night Earth observation data and overcome various undesired weather conditions. Simulating missing visible images from SAR images creates the possibility to have all-time visible observations, also not being dependent on the weather. To achieve this, we attempt to generate a visible image from the corresponding SAR image, as well as the multitemporal SAR-visible image pairs.

1) Implementation Details: In the experiments, we set the weighting parameter of consistency loss $\mathcal{L}_{\text{con}}$ and adversarial loss $\mathcal{L}_{\text{adv}}$ to 2 and 0.001, respectively. The averaging filter size s and the sigmoid function parameter k in target image reconstruction are set to 51 and 10, respectively. For the network training, we crop the input and guidance images into patches of size 50×50 pixels. A five-layer U-Net architecture is adopted as the generator, and the filter numbers in the first layer are set as 32. In the feature mask module, the threshold parameter η is empirically set to 0.5. We train the network for 60 epochs using a learning rate of 0.0005; then, we minimize the learning rate to 0.00025 and train for 20 additional epochs. The discriminator is a three-layer architecture, and the filter numbers in the first layer are set as 32. We train the discriminator for 60 epochs with the same learning rates as the generator. For optimization, we employ the Adam optimizer for both networks with momentum terms β_1 as 0.5 and β_2 as 0.999. The batch size is set to 16 to fit into the GPU memory. The learning model is implemented with the *Pytorch* package and runs on an NVIDIA Titan 1070 GPU with 8 GB of RAM.

2) Qualitative Results and Analysis: The qualitative results are presented in Fig. 5, and for a better visual inspection, a closeup view is shown at the right bottom of each picture. We first compare our result with the temporally neighboring images T_1 and T_3 in the target modality. It can be seen that, since the flood water retreats rapidly and the ground surface map changes a lot between the imaging dates, the temporally neighboring images T_1 and T_3 cannot reflect the right reflectance information at T_2 and yield a poor performance. In Fig. 5(d), the image was directly translated from S_2 by utilizing the input image and attempting to recover the visible image from the corresponding SAR image. By only exploiting the cross-modal information for translation, the translation approach will largely suffer from translation ambiguity (as mentioned in Section I-), and the quality of the result will become poor. By fully considering both temporal and cross-modal information, the performance of TIRSIT without adversarial loss is much more superior than that of any other results. For example, the edges of the floodwater and valley can be well presented in the image (see a closeup view in the red and yellow rectangles). Furthermore, by introducing the adversarial loss, TIRSIT can successfully recover clearer details and achieve more realistic results. This indicates that the adversarial loss enforces the output to lie in the real image manifold and achieve a more realistic result to the reference from the visual point of view.

3) Quantitative Results and Analysis: Table I lists the specific quantitative results obtained for different bands. We can see that the proposed framework significantly outperforms other approaches in terms of RMSE, CC, and SSIM.

TABLE I
QUANTITATIVE PERFORMANCE COMPARISON OF DIFFERENT IMAGES IN THE TASK OF SAR TO VISIBLE IMAGE TRANSLATION

	RMSE				SSIM				CC			
	Band1	Band2	Band3	Mean	Band1	Band2	Band3	Mean	Band1	Band2	Band3	Mean
Neighboring image T_1	0.0195	0.0302	0.0560	0.0352	0.9688	0.9418	0.8668	0.9258	0.5051	0.3520	0.2384	0.3651
Neighboring image T_3	0.0162	0.0148	0.0374	0.0228	0.9579	0.9498	0.8487	0.9188	0.7006	0.7419	0.5939	0.6788
Pix2pix	0.0193	0.0236	0.0339	0.0255	0.9334	0.9004	0.8136	0.8824	0.2860	0.1969	0.3269	0.2698
TIRSIT w/o adversarial loss	**0.0082**	**0.0101**	**0.0165**	**0.0116**	**0.9923**	**0.9882**	**0.9686**	**0.9831**	**0.8595**	**0.8507**	**0.8152**	**0.8418**
TIRSIT	0.0088	0.0106	0.0179	0.0124	0.9914	0.9875	0.9640	0.9810	0.8455	0.8382	0.7867	0.8231

TABLE II
QUANTITATIVE RESULTS OF ABLATION STUDIES IN THE TASK OF SAR TO VISIBLE IMAGE TRANSLATION

	RMSE				SSIM				CC			
	Band1	Band2	Band3	Mean	Band1	Band2	Band3	Mean	Band1	Band2	Band3	Mean
w/o UC and SG	0.0091	0.0121	0.0208	0.0140	0.9912	0.9850	0.9504	0.9755	0.8277	0.7896	0.7460	0.7878
w/o SG	0.0083	0.0102	0.0175	0.0120	0.9920	0.9878	0.9639	0.9812	0.8540	0.8481	0.8043	0.8355
w/o UC	0.0089	0.0114	0.0196	0.0133	0.9915	0.9860	0.9552	0.9775	0.8369	0.8119	0.7660	0.8050
TIRSIT w/o adversarial loss	**0.0082**	**0.0101**	**0.0165**	**0.0116**	**0.9923**	**0.9882**	**0.9686**	**0.9831**	**0.8595**	**0.8507**	**0.8152**	**0.8418**

This means that the proposed model can not only achieve the closest result to the ground truth (the smallest RMSE and CC) but also present the most structural details in the actual target image (the largest SSIM). In addition, it can also be observed that TIRSIT without adversarial loss performs better than TIRSIT in terms of quantitative results, which seems inconsistent with the subjective evaluation. This might be explained by the fact that the results that perform better in objective evaluation might be lacking high-frequency details and perceptually unsatisfying. As a consequence, they fail to match the fidelity expected in the target image.

4) Ablation Studies: We further evaluate the necessity of semantic guidance and uniqueness constraint in the proposed framework. To do that, we provide a further evaluation on different building blocks of the model, i.e., TIRSIT without semantic guidance and uniqueness constraint, TIRSIT without semantic guidance, TIRSIT without uniqueness constraint, TIRSIT w/o adversarial loss, and TIRSIT. Taking the translated difference image as an example, Fig. 6 shows the translated images, while Table II lists the corresponding quantitative assessment results. We observe that the adoption of semantic guidance and uniqueness constraint can significantly achieve sharper details and reduce the translation errors compared to other results. Regarding quantitative evaluation, the proposed TIRSIT framework consistently shows superior performance to that of others in terms of the three metrics due to the great contribution of temporal information learning from semantic guidance and uniqueness constraint.

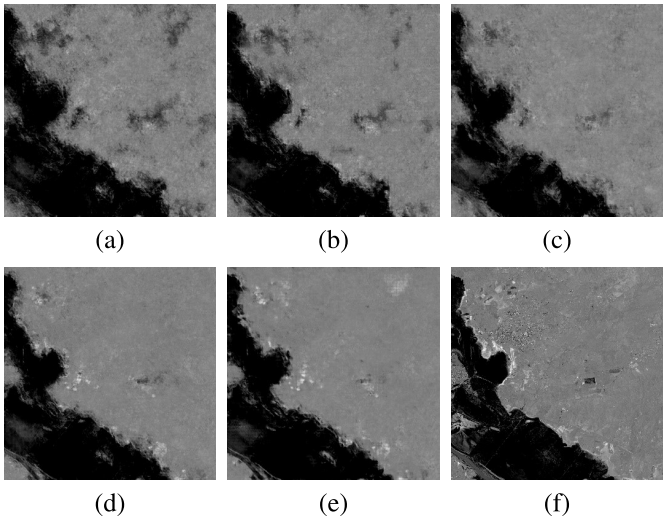

Fig. 6. Qualitative results of ablation studies in the task of SAR to visible image translation. (a) Translated image from TIRSIT w/o UC and SG. (b) Translated image from TIRSIT w/o SG. (c) Translated image from TIRSIT w/o UC. (d) Translated image from TIRSIT w/o adversarial loss. (e) Translated image from TIRSIT. (f) Ground truth.

D. Task 2: Visible to SWIR Image Translation

Visible cameras are the most widely used imaging sensors in remote sensing. Having a million tons of RGB images motivates us to exploit visible images to simulate other modality images in order to enrich observations. Furthermore, visible images usually provide the highest spatial-resolution Earth observation data. Translating visible to other modality images creates the possibility to achieve high-spatial-resolution multi-modality data. To this end, we also perform studies on the visible to SWIR image translation task to demonstrate the effectiveness of the proposed method.

1) Implementation Details: For the visible to SWIR image translation task, the weighting parameters of the consistency loss $\mathcal{L}_{\text{con}}$ and the adversarial loss $\mathcal{L}_{\text{adv}}$ are set to 5 and 0.001, respectively. The averaging filter size is set to 11, and the parameter k in the sigmoid function is set to 10 to reconstruct the target image. The threshold parameter η is set to 0.5 in

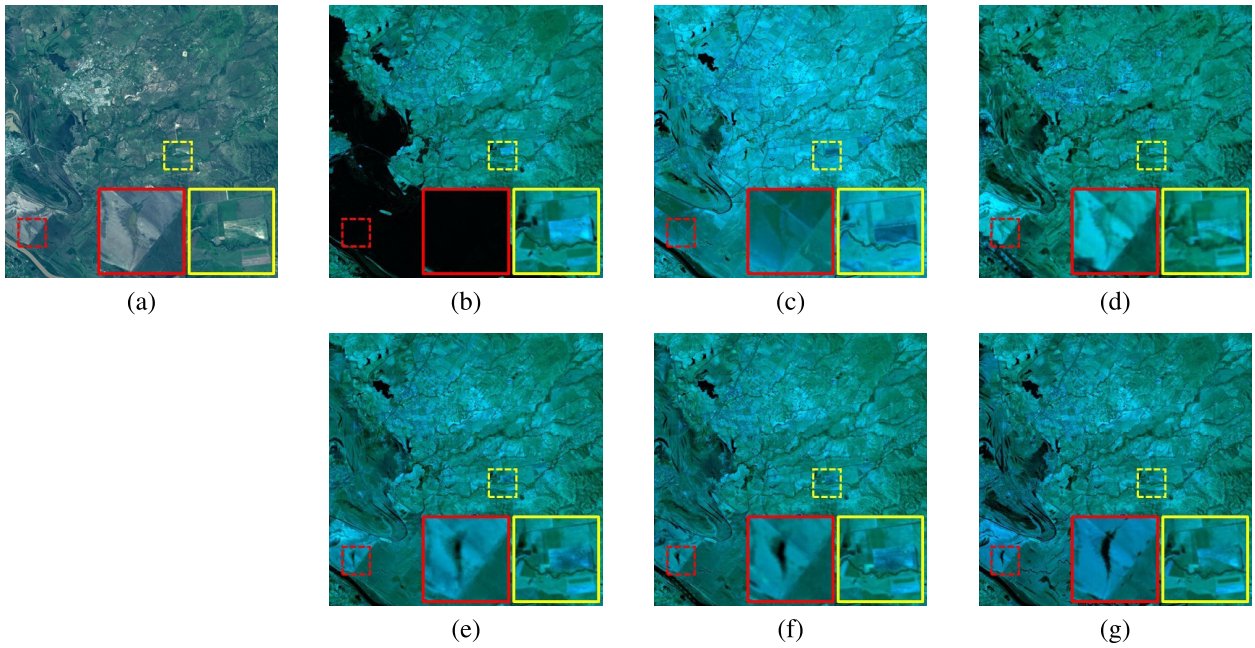

Fig. 7. Qualitative results for the task of the visible to SWIR image translation. (a) Input visible image S_2 taken at t_2. (b) Observed SWIR image T_1 taken at t_1. (c) Observed SWIR image T_3 taken at t_3. (d) Predicted image obtained by *pix2pix* model proposed in [51]. (e) Predicted image obtained by TIRSIT w/o adversarial loss. (f) Predicted image obtained by TIRSIT. (g) Observed target SWIR image T_2 taken at t_2.

TABLE III
QUANTITATIVE PERFORMANCE COMPARISON OF DIFFERENT IMAGES FOR THE TASK OF VISIBLE TO SWIR IMAGE TRANSLATION

	RMSE			SSIM			CC		
	Band1	Band2	Mean	Band1	Band2	Mean	Band1	Band2	Mean
Neighboring image T_1	0.0917	0.0638	0.0777	0.7252	0.7324	0.7288	0.7300	0.6591	0.6945
Neighboring image T_3	0.0762	0.0658	0.0709	0.8283	0.8193	0.8237	0.8040	0.7491	0.7765
Pix2pix	0.0434	0.0299	0.0366	0.8222	0.8592	0.8406	0.8050	0.8437	0.8244
TIRSIT w/o adversarial loss	0.0304	**0.0194**	**0.0249**	**0.9528**	**0.9618**	**0.9573**	0.9056	**0.9298**	**0.9177**
TIRSIT	**0.0301**	0.0213	0.0257	0.9504	0.9582	0.9542	**0.9083**	0.9151	**0.9177**

the feature mask module. The input and guidance images are cropped into patches of size 50×50 pixels for training. For the generator and discriminator, we adopt the same architectures as used for the task of SAR to visible image translation. The generator and discriminator are trained for 60 epochs using a learning rate of 0.002 and then trained for 20 additional epochs with a learning rate of 0.001. Similar to optimization strategies used for the experiments in Section IV-C, the Adam optimizer is used for both networks with momentum terms β_1 as 0.5 and β_2 as 0.999. The batch size is also set to 16 to fit into the GPU memory.

2) Qualitative Results and Analysis: We first visually assess the results for the task of visible to SWIR image translation, as shown in Fig. 7. Overall, there is a consistent trend for the results compared with that of the task of SAR to visible image translation. That is, for the temporally neighboring images T_1 and T_3 in the target modality, they fail to recover the rapidly changing areas in the ground surface due to the presence of acquired information at t_2. The original translation approach yields a relatively superior performance since it utilizes the input image in the source modality acquired at t_2 to generate the corresponding image in the target modality. However, as observed in the closeup view in Fig. 7(d), it is still limited by the disparities between multimodality data and tends to missing notable structure details that are originally present in the target image. In contrast, by introducing temporal side information into the process of cross-modal image translation, the proposed framework can effectively improve the visual quality of the predicted image, as shown in Fig. 7(e) and (f). This demonstrates that the proposed models can successfully benefit from both cross-modal information and temporal information and are able to have the translation result more similar to the reference.

3) Quantitative Results and Analysis: We also show the quantitative results for different bands in Table III to obtain an objective performance comparison. Similar to the qualitative evaluation, it can be observed that the original translation approach that considers the cross-modality data in the source modality always performs better than those approaches that only adopt temporally neighboring images. Not unexpectedly, the proposed framework, which integrates rich cross-modal and temporal information, achieves a superior performance,

TABLE IV
QUANTITATIVE RESULTS OF THE ABLATION STUDIES FOR THE TASK OF SAR TO VISIBLE IMAGE TRANSLATION

	RMSE			SSIM			CC		
	Band1	Band2	Mean	Band1	Band2	Mean	Band1	Band2	Mean
w/o UC and SG	0.0349	0.0222	0.0286	0.9499	0.9595	0.9547	0.9085	0.9148	0.9117
w/o SG	0.0321	0.0212	0.0267	0.9503	0.9597	0.9550	0.8992	0.9158	0.9075
w/o UC	**0.0301**	0.0207	0.0254	0.9514	0.9603	0.9559	**0.9132**	0.9209	0.9171
TIRSIT w/o adversarial loss	0.0304	**0.0194**	**0.0249**	**0.9528**	**0.9618**	**0.9573**	0.9056	**0.9298**	**0.9177**

TABLE V
QUANTITATIVE PERFORMANCE COMPARISON OF DIFFERENT IMAGES IN THE TASK OF SAR TO VISIBLE IMAGE TRANSLATION (EXCHANGING DATA OF DIFFERENT TIMES FOR TRAINING)

	RMSE				SSIM				CC			
	Band1	Band2	Band3	Mean	Band1	Band2	Band3	Mean	Band1	Band2	Band3	Mean
Neighboring image T_1	0.0207	0.0281	0.0588	0.0359	0.9549	0.9334	0.7991	0.8958	0.3042	0.2892	0.0686	0.1749
Neighboring image T_2	0.0162	0.0148	0.0374	0.0228	0.9579	0.9498	0.8487	0.9188	0.7006	0.7419	0.5939	0.6788
Pix2pix	0.0172	0.0214	0.0343	0.0243	0.0901	0.8712	0.7632	0.8482	0.2821	0.2420	0.3440	0.2894
TIRSIT	**0.0129**	**0.0130**	**0.0255**	**0.0171**	**0.9646**	**0.9530**	**0.8816**	**0.9330**	**0.7645**	**0.7587**	**0.6832**	**0.7354**

which demonstrates that incorporating temporal side information into cross-modal translation is applicable to generate a more accurate result. By comparing the quantitative results of the two translation tasks (visible to SWIR and SAR to visible), we can notice that the performance difference between the original translation approach and the proposed model is much smaller in the former task than that in the latter one. This is due to the considerable cross-modal similarities between visible and SWIR images, which are both categorized as optical data and have almost similar imaging mechanisms.

4) Ablation Studies: As shown in Fig. 8, we present multiple translation results to demonstrate the effectiveness of different module settings in the proposed framework, including TIRSIT without semantic guidance and uniqueness constraint, TIRSIT without semantic guidance, TIRSIT without uniqueness constraint, TIRSIT without adversarial loss, and TIRSIT. Correspondingly, Table IV lists the quantitative assessment results in terms of three metrics. As expected, the visual quality of the results obtained by TIRSIT without adversarial loss and TIRSIT models is significantly superior to that of the others due to the inclusion of the temporal information. In this way, the semantic guidance and uniqueness constraint modules help the models to transfer knowledge from temporal images to the target one in the learning phase to some extent. This leads to more accurate external textures that are not given in the input image of the source modality (see the left bottom part in the results). For quantitative evaluation, we observe it again that the proposed models can achieve significant improvement compared to other setups of the framework. This phenomenon further demonstrates the advantage of introducing temporal side information, i.e., the semantic guidance and uniqueness constraint, in network learning.

E. How About Exchanging Data of Different Times for Training?

In some remote sensing applications, only historical images are available and can be used as the prior temporal information in cross-modality image translation, especially for some time-sensitive tasks. This could be more challenging for the translation since we have to predict the patterns that might never exist in the historical images. To further show the effectiveness of the proposed model, we consider exchanging data from different times in the training phase for the SAR to visible image translation task and predicting cross-modality images only with cross-modal and historical data. In the experiment, the goal is to predict the unknown visible image T_3 acquired at date t_3 from the corresponding SAR image S_3, as well as the temporal SAR and visible image pairs S_1 and T_1 at t_1, and S_2 and T_2 at t_2 as the temporal side information.

The qualitative results have been shown in Fig. 9. We can see that the result produced by the TIRSIT model generally recovers the patterns in image T_3 by simultaneously considering temporal and cross-modal information and shows more similarities with the ground truth compared to the neighboring images and the result produced by the *pix2pix* model. We also present the quantitative results of different bands in Table V. Three indices show that there is a consistent trend for the qualitative evaluation, and the proposed model yields the most accurate translation results than any other images. These results further prove the validity of the proposed model in the translation task of visible to SAR images.

However, it should be noted that compared with the neighboring images, the result produced by the proposed model has a smaller performance improvement than the experimental results without exchanging data of different times. This indicates that the proposed model shows an inferior performance

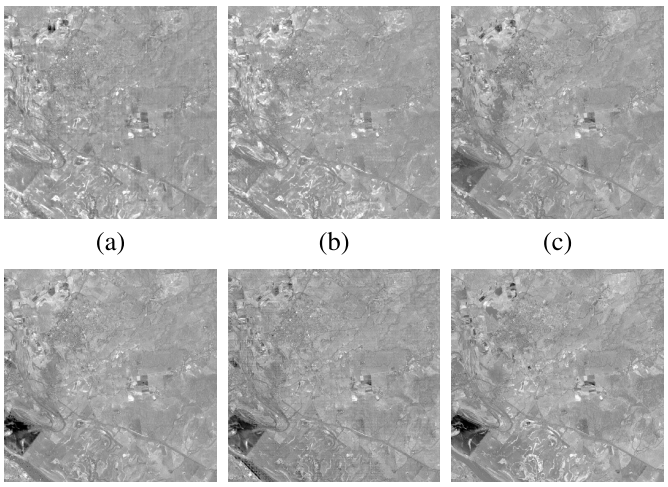

Fig. 8. Qualitative results of the ablation studies for the task of visible to SWIR image translation. (a) Translated image from TIRSIT w/o UC and SG. (b) Translated image from TIRSIT w/o SG. (c) Translated image from TIRSIT w/o UC. (d) Translated image from TIRSIT w/o adversarial loss. (e) Translated image from TIRSIT. (f) Ground truth.

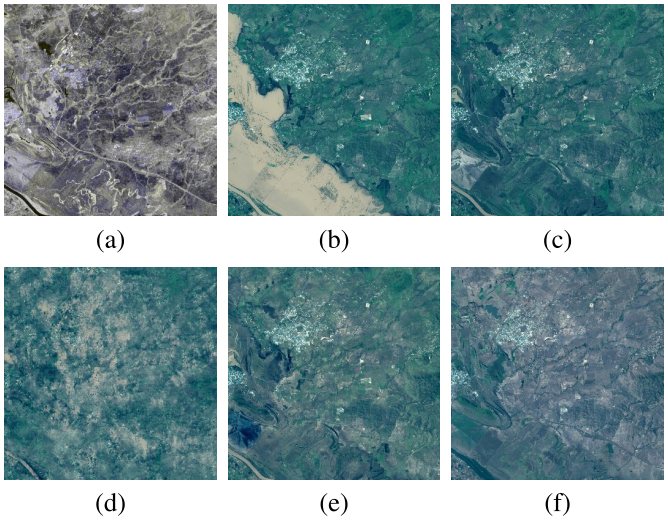

Fig. 9. Qualitative results of exchanging different times data for training in SAR to visible image translation. (a) Input SAR image S_3 taken at t_3. (b) Observed visible image T_1 taken at t_1. (c) Observed visible image T_2 taken at t_2. (d) Translated image from *pix2pix* model. (e) Translated image from TIRSIT. (f) Observed target visible image T_3 taken at t_3.

when we exchange data of different times for training. This is considerable since, when we predict the unknown image T_3, some components in the image do not exist in the training data. Consequently, the training and test data cannot be very similar in the task, and the trained model might not work well in the prediction phase, especially when some abrupt type changes exist.

V. Conclusion

In this work, we presented a novel TIRSIT framework. Unlike the other translation approaches that only rely on input images to generate the cross-modality outputs, the proposed model attempts to involve time-series data to resolve the inherent ambiguities in translation. In particular, we introduced a guidance image from the temporal images in translation and exploited semantic information in the guidance image with a feature mask module. Furthermore, while incorporating multiple pairs of images in time series, we formulated a temporal constraint during the learning process to guarantee the uniqueness of the prediction result. Extensive experiments were conducted on a new dataset that contains multimodal and multitemporal remote sensing images, and experimental results on two translation tasks demonstrated the superiority and effectiveness of the proposed framework.

We also observed many interesting avenues of future work, including extending the guided image translation framework into many specific remote sensing image restoration tasks, such as cloud removal, denoising, and super-resolution.

References

[1] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[2] D. Hong, N. Yokoya, N. Ge, J. Chanussot, and X. X. Zhu, "Learnable manifold alignment (LeMA): A semi-supervised cross-modality learning framework for land cover and land use classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 193–205, Jan. 2019.

[3] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, "Deep learning-based image segmentation on multimodal medical imaging," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 3, no. 2, pp. 162–169, Mar. 2019.

[4] P. Song, X. Deng, J. F. C. Mota, N. Deligiannis, P. L. Dragotti, and M. R. D. Rodrigues, "Multimodal image super-resolution via joint sparse representations induced by coupled dictionaries," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 57–72, 2020.

[5] C. Deng, X. Liu, J. Chanussot, Y. Xu, and B. Zhao, "Towards perceptual image fusion: A novel two-layer framework," *Inf. Fusion*, vol. 57, pp. 102–114, May 2020.

[6] D. Hong *et al.*, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[7] X. Wu, K. Xu, and P. Hall, "A survey of image synthesis and editing with generative adversarial networks," *Tsinghua Sci. Technol.*, vol. 22, no. 6, pp. 660–674, Dec. 2017.

[8] A. Gokaslan, V. Ramanujan, D. Ritchie, K. I. Kim, and J. Tompkin, "Improving shape deformation in unsupervised image-to-image translation," in *Proc. ECCV*, 2018, pp. 649–665.

[9] J. Wang, J. Zhang, Z. Lu, and S. Shan, "DFT-Net: Disentanglement of face deformation and texture synthesis for expression editing," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3881–3885.

[10] L. Gao, D. Hong, J. Yao, B. Zhang, P. Gamba, and J. Chanussot, "Spectral superresolution of multispectral imagery with joint sparse and low-rank learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2269–2280, Mar. 2021.

[11] J. Li, Z. Ling, L. Niu, and L. Zhang, "Bi-directional domain translation for zero-shot sketch-based image retrieval," 2019, *arXiv:1911.13251*. [Online]. Available: http://arxiv.org/abs/1911.13251

[12] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "CoSpace: Common subspace learning from hyperspectral-multispectral correspondences," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4349–4359, Jul. 2019.

[13] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big data for remote sensing: Challenges and opportunities," *Proc. IEEE*, vol. 104, no. 11, pp. 2207–2219, Nov. 2016.

[14] X. Deng *et al.*, "Geospatial big data: New paradigm of remote sensing applications," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 10, pp. 3841–3851, Oct. 2019.

[15] L. Gomez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, Sep. 2015.

[16] H. Peng, X. Chen, and J. Zhao, "Residual pixel attention network for spectral reconstruction from RGB images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 486–487.

[17] J. D. Bermudez, P. N. Happ, D. A. B. Oliveira, and R. Q. Feitosa, "Sar to optical image synthesis for cloud removal with generative adversarial networks," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. IV-1, pp. 5–11, Sep. 2018.

[18] L. Liu and B. Lei, "Can SAR images and optical images transfer with each other?" in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 7019–7022.

[19] L. Wang *et al.*, "SAR-to-optical image translation using supervised cycle-consistent adversarial networks," *IEEE Access*, vol. 7, pp. 129136–129149, 2019.

[20] M. Fuentes Reyes, S. Auer, N. Merkle, C. Henry, and M. Schmitt, "SAR-to-optical image translation based on conditional generative adversarial networks—Optimization, opportunities and limits," *Remote Sens.*, vol. 11, no. 17, p. 2067, Sep. 2019.

[21] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[22] J. C. Curlander and R. N. McDonough, *Synthetic Aperture Radar*, vol. 11. New York, NY, USA: Wiley, 1991.

[23] C. Oliver and S. Quegan, *Understanding Synthetic Aperture Radar Images*. SciTech, 2004.

[24] O. Tasar, S. L Happy, Y. Tarabalka, and P. Alliez, "SemI2I: Semantically consistent image-to-image translation for domain adaptation of remote sensing data," 2020, *arXiv:2002.05925*. [Online]. Available: http://arxiv.org/abs/2002.05925

[25] O. Tasar, Y. Tarabalka, A. Giros, P. Alliez, and S. Clerc, "StandardGAN: Multi-source domain adaptation for semantic segmentation of very high resolution satellite images by data standardization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 192–193.

[26] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, "ColorMapGAN: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7178–7193, Oct. 2020.

[27] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, early access, Aug. 18, 2020, doi: 10.1109/TGRS.2020.3015157.

[28] L. T. Luppino *et al.*, "Deep image translation with an affinity-based change prior for unsupervised multimodal change detection," 2020, *arXiv:2001.04271*. [Online]. Available: http://arxiv.org/abs/2001.04271

[29] Y. Li, S. Tang, R. Zhang, Y. Zhang, J. Li, and S. Yan, "Asymmetric GAN for unpaired image-to-image translation," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5881–5896, Dec. 2019.

[30] H.-Y. Lee *et al.*, "DRIT++: Diverse image-to-image translation via disentangled representations," *Int. J. Comput. Vis.*, vol. 128. no. 10, pp. 1–16, 2020.

[31] A. Royer *et al.*, "XGAN: Unsupervised image-to-image translation for many-to-many mappings," in *Domain Adaptation for Visual Understanding*. Springer, 2020, pp. 33–49.

[32] D. Hong, X. Wu, P. Ghamisi, J. Chanussot, N. Yokoya, and X. X. Zhu, "Invariant attribute profiles: A spatial-frequency joint feature extractor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 3791–3808, Jun. 2020.

[33] J.-Y. Zhu *et al.*, "Toward multimodal image-to-image translation," in *Proc. NIPS*, 2017, pp. 465–476.

[34] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[35] Z. Zhu, "Change detection using landsat time series: A review of frequencies, preprocessing, algorithms, and applications," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 370–384, Aug. 2017.

[36] 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.

[37] X. Liu, C. Deng, B. Zhao, and J. Chanussot, "Multimodal-temporal fusion: Blending multimodal remote sensing images to generate image series with high temporal resolution," in *Proc. IGARSS*, Jul. 2019, pp. 10083–10086.

[38] B. Rasti *et al.*, "Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 4, pp. 60–88, Dec. 2020.

[39] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[40] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2794–2802.

[41] L. Yu, W. Zhang, J. Wang, and Y. Yu, "SeqGAN: Sequence generative adversarial nets with policy gradient," in *Proc. AAAI*, 2017, pp. 2852–2858.

[42] D. Hong, J. Yao, D. Meng, Z. Xu, and J. Chanussot, "Multimodal GANs: Toward crossmodal hyperspectral-multispectral image segmentation," *IEEE Trans. Geosci. Remote Sens.*, Sep. 16, 2020, doi: 10.1109/TGRS.2020.3020823.

[43] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," 2017, *arXiv:1701.00160*. [Online]. Available: http://arxiv.org/abs/1701.00160

[44] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784

[45] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," 2016, *arXiv:1605.05396*. [Online]. Available: http://arxiv.org/abs/1605.05396

[46] C.-H. Wan, S.-P. Chuang, and H.-Y. Lee, "Towards audio to scene image synthesis using generative adversarial network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 496–500.

[47] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[48] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 613–621.

[49] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, "Image to image translation for domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4500–4509.

[50] M. Sela, E. Richardson, and R. Kimmel, "Unrestricted facial geometry reconstruction using image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1576–1585.

[51] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[52] Y. Han, C. Deng, B. Zhao, and D. Tao, "State-aware anti-drift object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4075–4086, Aug. 2019.

[53] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[54] A. Gonzalez-Garcia, J. Van De Weijer, and Y. Bengio, "Image-to-image translation for cross-domain disentanglement," in *Proc. NIPS*, 2018, pp. 1287–1298.

[55] M.-Y. Liu *et al.*, "Few-shot unsupervised image-to-image translation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10551–10560.

[56] P. Ghamisi and N. Yokoya, "IMG2DSM: Height simulation from single imagery using conditional generative adversarial net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 794–798, May 2018.

[57] M. E. Paoletti, J. M. Haut, P. Ghamisi, N. Yokoya, J. Plaza, and A. Plaza, "U-IMG2DSM: Unpaired simulation of digital surface models with generative adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, early access, Jun. 4, 2020, doi: 10.1109/LGRS.2020.2997295.

[58] D. Ao, C. O. Dumitru, G. Schwarz, and M. Datcu, "Dialectical GAN for SAR image translation: From sentinel-1 to TerraSAR-X," *Remote Sens.*, vol. 10, no. 10, p. 1597, Oct. 2018.

[59] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proc. AAAI*, 2018, pp. 3942–3951.

[60] B. Albahar and J.-B. Huang, "Guided image-to-image translation with Bi-directional feature transformation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9016–9025.

[61] L. Ma, X. Jia, S. Georgoulis, T. Tuytelaars, and L. Van Gool, "Exemplar guided unsupervised image-to-image translation with semantic consistency," 2018, *arXiv:1805.11145*. [Online]. Available: http://arxiv.org/abs/1805.11145

[62] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2015, pp. 234–241.

[63] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434

[64] 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.

[65] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," 2016, *arXiv:1610.07629*. [Online]. Available: https://arxiv.org/abs/1610.07629

Xun Liu (Student Member, IEEE) received the B.Eng. and Ph.D. degrees from the Beijing Institute of Technology, Beijing, China, in 2014 and 2021, respectively.

From 2017 to 2018, he was a Visiting Ph.D. Student with GIPSA-lab, Centre national de la recherche scientifique (CNRS), Grenoble Institute of Technology (Grenoble INP), Université Grenoble Alpes, Grenoble, France. His research interests include remote sensing image fusion, super-resolution, and machine learning.

Dr. Liu was a recipient of the IEEE Mikio Takagi Student Prize for winning the Student Paper Competition at the IEEE International Geoscience and Remote Sensing Symposium (IGARSS) in 2019.

Danfeng Hong (Senior Member, IEEE) received the M.Sc. degree *(summa cum laude)* in computer vision from the College of Information Engineering, Qingdao University, Qingdao, China, in 2015, and the Dr.Ing degree *(summa cum laude)* in signal processing in Earth observation (SiPEO) from the Technical University of Munich (TUM), Munich, Germany, in 2019.

Since 2015, he has been a Research Associate with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Weßling, Germany. He is a Research Scientist and leads the Spectral Vision Working Group, IMF, DLR, and also an Adjunct Scientist with GIPSA-lab, Centre national de la recherche scientifique (CNRS), Grenoble Institute of Technology (Grenoble INP), Université Grenoble Alpes, Grenoble, France. His research interests include signal/image processing and analysis, hyperspectral remote sensing, machine/deep learning, and artificial intelligence and their applications in Earth Vision.

Dr. Hong is an Editorial Board Member of *Remote Sensing* and a Topical Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He was a recipient of the Best Reviewer Award of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING in 2020 and the Jose Bioucas Dias Award for recognizing the outstanding paper at the Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS) in 2021. He is a Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING and *Remote Sensing*.

Jocelyn Chanussot (Fellow, IEEE) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995, and the Ph.D. degree from the Université de Savoie, Annecy, France, in 1998.

Since 1999, he has been with Grenoble INP, where he is a Professor of signal and image processing. He has been a Visiting Scholar with Stanford University, Stanford, CA, USA, the KTH Royal Institute of Technology, Stockholm, Sweden, and the National University of Singapore (NUS), Singapore. Since 2013, he has been an Adjunct Professor with the University of Iceland, Reykjavik, Iceland. From 2015 to 2017, he was a Visiting Professor with the University of California at Los Angeles (UCLA), Los Angeles, CA, USA. He holds the AXA Chair in remote sensing and is an Adjunct Professor with the Chinese Academy of Sciences, Aerospace Information Research Institute, Beijing, China. His research interests include image analysis, hyperspectral remote sensing, data fusion, machine learning, and artificial intelligence.

Dr. Chanussot was a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society from 2006 to 2008. He was a member of the Institut Universitaire de France from 2012 to 2017. He was the Founding President of the IEEE Geoscience and Remote Sensing French Chapter from 2007 to 2010, which received the 2010 IEEE GRS-S Chapter Excellence Award. He received multiple outstanding paper awards. He was the Vice-President of the IEEE Geoscience and Remote Sensing Society, in charge of meetings and symposia, from 2017 to 2019. He was the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing in 2009. He was the General Chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote sensing (WHISPERS).

He was the Chair and the Co-Chair of the GRS Data Fusion Technical Committee from 2009 to 2011 and 2005 to 2008, respectively. In 2014, he has served as a Guest Editor for the *IEEE Signal Processing Magazine*. He was the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING from 2011 to 2015. He is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the PROCEEDINGS OF THE IEEE. He has been a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters) since 2018.

Baojun Zhao (Member, IEEE) received the Ph.D. degree in electromagnetic measurement technology and equipment from the Harbin Institute of Technology (HIT), Harbin, China, in 1996.

From 1996 to 1998, he was a Post-Doctoral Fellow with the Beijing Institute of Technology (BIT), Beijing, China, where he is a Full Professor. He is also the Vice-Director of the Laboratory and Equipment Management, Beijing, and the Director of the National Signal Acquisition and Processing Professional Laboratory, Beijing. He has authored or coauthored more than 100 publications. His research interests include image/video coding, image recognition, infrared/laser signal processing, and parallel signal processing.

Dr. Zhao received five provincial-/ministerial-level scientific and technological progress awards in his research fields.

Pedram Ghamisi (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Iceland, Reykjavik, Iceland, in 2015.

He is the Head of the Machine Learning Group, Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Freiberg, Germany, the CTO and a Co-Founder of VasoGnosis Inc., Milwaukee, WI, USA, and a Visiting Professor with the Institute of Advanced Research in Artificial Intelligence (IARAI), Vienna, Austria. His research interests include interdisciplinary research on machine (deep) learning, image and signal processing, and multisensor data fusion.

Dr. Ghamisi was a recipient of the IEEE Mikio Takagi Prize for winning the Student Paper Competition at the IEEE International Geoscience and Remote Sensing Symposium (IGARSS) in 2013, the First Prize of the Data Fusion Contest organized by the IEEE Image Analysis and Data Fusion Committee (IADF) in 2017, the Best Reviewer Prize of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS in 2017, and the IEEE Geoscience and Remote Sensing Society 2020 Highest-Impact Paper Award. He is also the Co-Chair of IEEE IADF. For detailed information, please see http://pedram-ghamisi.com/