IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 June 2023, accepted 13 July 2023, date of publication 20 July 2023, date of current version 16 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3297513

RESEARCH ARTICLE

Deep Learning-Based Spatiotemporal Fusion of Unmanned Aerial Vehicle and Satellite Reflectance Images for Crop Monitoring

JUAN XIAO[1], ASHWANI KUMAR AGGARWAL[2], (Senior Member, IEEE), UDAY KIRAN RAGE[3], (Senior Member, IEEE), VAIBHAV KATIYAR[4], (Member, IEEE), AND RAM AVTAR[1,5]

[1]Graduate School of Environmental Science, Hokkaido University, Sapporo 060-0810, Japan
[2]Electrical and Instrumentation Engineering Department, Sant Longowal Institute of Engineering and Technology, Longowal, Punjab 148106, India
[3]Division of Information Systems, The University of Aizu, Aizuwakamatsu, Fukushima 965-0006, Japan
[4]Graduate School of Sciences and Technology for Innovation, Yamaguchi University, Ube, Yamaguchi 755-8611, Japan
[5]Faculty of Environmental Earth Science, Hokkaido University, Sapporo 060-0810, Japan

Corresponding author: Ram Avtar (ram@eed.hokudai.ac.jp)

This work was supported in part by the Japan Science and Technology (JST) SPRING under Grant JPMJSP2119, in part by Hirose Grant, and in part by the Asia Pacific Network for Global Change Research (APN-GCR) under Grant CBA2021-02MY-Avtar.

ABSTRACT Spatiotemporal fusion (STF) techniques play important roles in Earth observation analysis as they enable the generation of images with high spatial and temporal resolution. However, existing STF models often fuse images from various satellites, not satisfying the demand for precise crop monitoring. In contrast, unmanned aerial vehicle (UAV) images can deliver detailed data, and deep learning (DL)-based STF models have the potential to automatically extract abstract features. To this end, this study proposed a novel end-to-end DL-based STF model named UAV-Net, which can produce centimeter-scale UAV images. UAV-Net has an encoder-decoder architecture with Modified ResNet (MResNet), Feature Pyramid Network (FPN), and decoder modules. The encoder uses MResNet modules to extract input features, while the FPN module performs a multiscale fusion of these features before reconstructing UAV images using transposed convolution in the decoder module. Through comparative and ablation experiments, this study evaluated the efficacies of MResNet modules with 18, 34, and 50 layers, along with the FPN module of UAV-Net. The experimental results on real-world datasets demonstrated that UAV-Net adequately produces UAV images both visually and quantitatively. Furthermore, a comparison with state-of-the-art STF models highlights the innovation and effectiveness of UAV-Net in producing centimeter-scale images. The predicted centimeter-scale images using UAV-Net have great potential for various environmental monitoring applications.

INDEX TERMS Spatiotemporal fusion, crop monitoring, UAV-Net, ResNet, feature pyramid network (FPN).

I. INTRODUCTION

Crop monitoring is a vital aspect of the agricultural production process, which enables farmers to implement effective management for yield optimization [1]. Conventional field surveys for crop monitoring can be labor-intensive, time-consuming, and potentially destructive. Thus, remote sensing (RS) techniques are an attractive option for high-efficiency and non-destructive crop monitoring. RS technology is widely used for Earth observation and is crucial in agricultural applications. RS platforms can be broadly classified into three categories based on the distance to the target object: spaceborne (e.g., satellites), airborne (e.g., Unmanned Aerial Vehicles [UAVs]), and ground-based (e.g., hand-held devices). Among these platforms, UAVs are highly versatile and can be equipped with a range of sensors, providing

The associate editor coordinating the review of this manuscript and approving it for publication was Gerardo Di Martino.

images with a high spatial resolution that includes precise real-time information regarding crops [2]. On the other hand, satellite images with moderate spatial resolution and fixed revisit frequencies are ideal for long-term and large-scale crop monitoring.

UAV images with dense time series and rich spatial information have a wide range of potential applications [3]. These images are valuable for crop growth monitoring, improved farming efficiency, and yield estimation. However, daily UAV imaging is challenging due to high operational costs, tedious image processing, etc. Spatiotemporal fusion (STF) technology can address these challenges by generating cost-effective, dense time-series UAV images. STF can generate images with high spatial and temporal resolution by combining temporally frequent but spatially coarse images (below ''coarse image'') with spatially fine but less temporally frequent imagers (below ''fine image''). STF poses a challenging and undefined problem: how to reconstruct a fine image by modeling the complex relationship between fine and coarse images [4]. Existing STF models are categorized into four groups based on their methods of linking fine and coarse images: unmixing-based, weight function-based, learning-based, and hybrid-based [5]. Unmixing-based models use linear unmixing theory to map the relationship between fine and coarse images [6], and the multisensor multiresolution technique [44] was the first STF model in this category [7]. Subsequently, multiple STF models have been developed to manage the low spectral accuracy and intraclass spectral variability issues that affect the multisensor multiresolution technique [8], [9]. Weight function-based models reconstruct fine images by empirically weighing the inputs, with the spatial and temporal adaptive reflectance fusion model (STARFM) being widely used [10], [11], [12]. However, both unmixing-based methods and weight function-based have limitations related to unreasonable assumptions [13], [14]. Learning-based STF models have rapidly advanced in recent years and can be further divided into sparse representation, Bayesian, machine learning, and deep learning (DL) models. Sparse representation learning models make key assumptions regarding dictionaries and sparse coding coefficients [15], while existing Bayesian learning models have stringent input requirements or are designed for specific applications [16]. Machine learning algorithms are not effective in high-dimensional RS image prediction. DL-based STF models, on the other hand, exhibit superior performance by establishing complex mappings between input and output images and using a large number of available RS images. Advances in DL network technology have led to a rapid increase in the number of DL-based STF models. Common DL networks used in STF models include the deep convolutional neural network [7], [17], [18], generative adversarial network [19], [20], [21], AutoEncoder [22], [23], Long Short-Term Memory Network [24], and Transformer [25]. The performance of DL-based STF models can be improved by combining various DL strategies to accommodate complex image mapping. Such strategies include residual learning [7], [18], [26], attention mechanisms [23], [27], [28], super-resolution [17], [24], [29], multiscale mechanisms [18], [29], [30] and a compound loss function [21], [22]. Finally, hybrid-based STF models leverage the advantages of the three STF categories to achieve more accurate results. However, this approach increases the computational cost.

STF has advanced significantly with many STF models being developed. Existing STF models, on the other hand, are commonly used to fuse images from various satellites, such as MODIS and Landsat [22], [31], or Landsat and Sentinel [32], [33]. Landsat and MODIS images are widely used in STF considering their similar bandwidths and radiations [34]. However, current fusions with spatial resolutions of 10 m or 30 m are inadequate for precise and small-scale applications. Additionally, existing STF models use fine and coarse images, with scales that differ by less than 16-fold. Therefore, fusions with centimeter-level spatial resolutions and higher magnifications are in great demand for precise Earth observation.

To this end, this study proposed a novel end-to-end STF model named UAV-Net, which can generate UAV images with a centimeter-level spatial resolution by fusing UAV and PlanetScope satellite images with 150-fold magnification. The predicted UAV image provides several advantages for crop monitoring, including the high spatial resolution with multiple spectral bands. This valuable information can be used to generate vegetation indices maps to monitor crop growth performance. As illustrated in Fig. 1, UAV-Net uses a UAV-PlanetScope image pair captured at time t_1 and a PlanetScope image captured at time t_2 as inputs to predict the UAV image at t_2. The PlanetScope and UAV images captured at t_1 and t_2 are designated ${}^{P}I_{t1}$, ${}^{P}I_{t2}$, ${}^{U}I_{t1}$, and ${}^{U}I_{t2}$, respectively. UAV-Net learns the changes in the features of the PlanetScope images at t_1 and t_2 to guide the reconstruction of the UAV image at t_2, with reference to the UAV image at t_1.

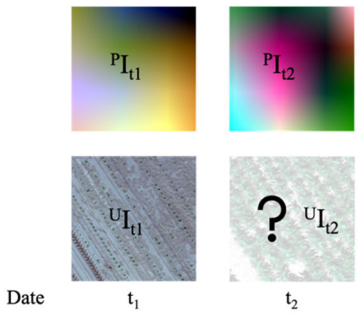

FIGURE 1. Spatiotemporal fusion of UAV and PlanetScope images ($t_1 < t_2$).

UAV-Net is designed to predict centimeter-scale UAV images that can be used for crop monitoring. UAV and PlanetScope images were collected over a corn field and used for the UAV-Net training and validation. UAV-Net exhibits an encoder-decoder architecture with Modified ResNet (MResNet), Feature Pyramid Network (FPN), and decoder modules

that address significant differences in spatial resolution when reconstructing UAV images. The efficacy of the building modules was evaluated through one comparative experiment and one ablation experiment. The key contributions of this study are threefold:

1) Proposed an end-to-end STF model that fuses UAV and PlanetScope images. To our knowledge, this is the first DL-based STF model to predict centimeter-scale images using UAV and PlanetScope image datasets.
2) Investigated how MResNet and FPN modules affect the fusion accuracy of UAV-Net.
3) Provided a cost-effective way to generate time-series centimeter-scale UAV images for crop monitoring and other precise environmental monitoring application.

The remainder of this study is structured as follows: Section II presents an overview of UAV-Net and the detailed architecture of each module. Section III conducts the comparative and ablation experiments of the building modules within UAV-Net, as well as an experiment comparing the model accuracy of UAV-Net with three state-of-the-art models. It also discusses the limitation and future scope. Section IV is the conclusion.

II. METHODOLOGY

This section provided a comprehensive description of the proposed UAV-Net. Firstly, an overview of the entire UAV-Net architecture was presented, followed by a detailed explanation of the design of each module, including the MResNet, FPN, and decoder architectures, as well as the compound loss function. Additionally, the Normalized Difference Vegetation Index (NDVI) and four evaluation metrics used to assess the fusion results were introduced.

A. OVERVIEW OF PROPOSED UAV-NET ARCHITECTURE

The proposed UAV-Net architecture is made up of three components: an encoder, an FPN, and a decoder. The encoder is MResNet modules. As shown in Fig 2, the model inputs are three images including one PlanetScope image at t_1 ($^{P}I_{t1}$) and t2 ($^{P}I_{t2}$,), and a UAV image at t1 ($^{U}I_{t1}$). The encoder is used to compress the inputs to extract and encode important image features, while the FPN is used for multiscale feature fusion. Lastly, the decoder is used to reconstruct the UAV image at t2 ($^{U}I_{t2}$) based on the extracted input features. Considering the approximately 150-fold difference in spatial resolution between UAV and PlanetScope images, it is necessary to interpolate PlanetScope images to match the pixel size of the UAV image before inputting them into UAV-Net. Consequently, the features of the PlanetScope images will become larger. To extract adequate features from PlanetScope images, a larger receptive field is required for the convolution kernel. The downsampling processing in the encoder module helps to increase the receptive field, ensuring that the extracted features include enough information for accurate reconstruction in the decoder module. This encoder and decoder architecture would be highly advantageous for UAV-based STF.

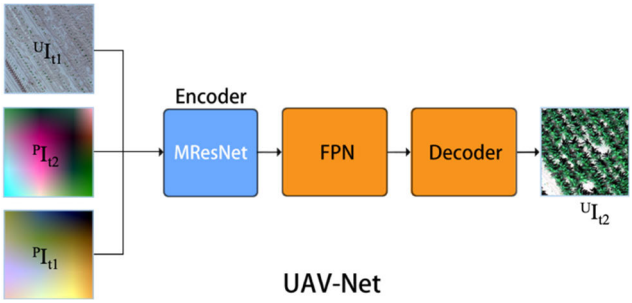

FIGURE 2. The architecture of UAV-Net.

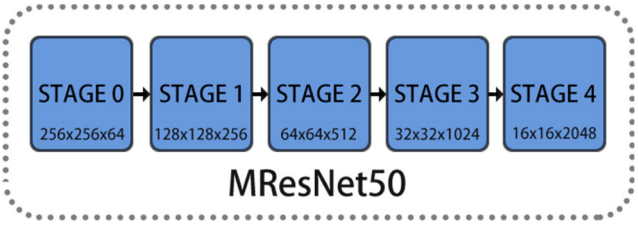

FIGURE 3. The architecture of the modified ResNet module with 50 layers.

B. MRESNET

It has been claimed that the deeper the DL network, the better performance of the network. Because a deeper network can learn more complex and non-linear functions, thereby extracting more abstract features with semantic information. Deeper networks, on the other hand, are more difficult to train due to the issue of vanishing or exploding gradients, as well as overfitting and degradation. Thus, He et al. [35] proposed the deep residual network (ResNet) architecture to alleviate the issues raised by increasing depth in deep neural networks. The most important idea in ResNet is the introduction of shortcut connections, which could facilitate information propagation and enable the networks to operate with fewer parameters.

Although ResNet has gained popularity as a feature extractor network, it has some shortcomings when used for image reconstruction tasks. The maximum pooling layer, which retains only the maximum value of features, increases the risk of detail information loss. Therefore, the MResNet module was used in the UAV-Net by modifying ResNet according to the following operation: 1) the maxpooling layer in Stage 1 was removed, and 2) the stride of the first convolutional layer in Stage 1 was changed to 2 to ensure that the feature map is the same size as the original ResNet architecture. The residual configuration of the MResNet is identical to the original ResNet, as outlined in [35]. This modification enhances the subsequent high-resolution UAV image reconstruction process. Fig. 3 illustrates the architecture of the MResNet module with 50 layers used in the UAV-Net.

C. FEATURE PYRAMID NETWORK

DL architectures typically use multiscale fusion mechanisms to enrich the details of feature maps [36]. Because the deep convolution gradually loses spatial information as the

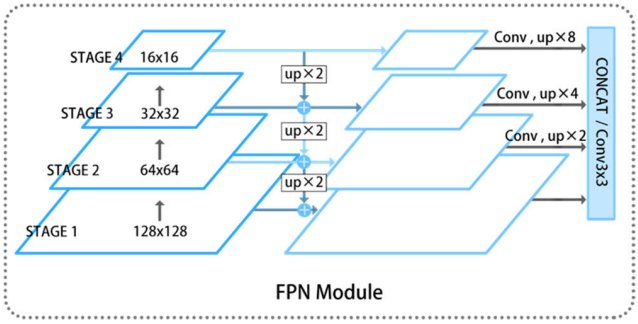

FIGURE 4. The architecture of the FPN module.

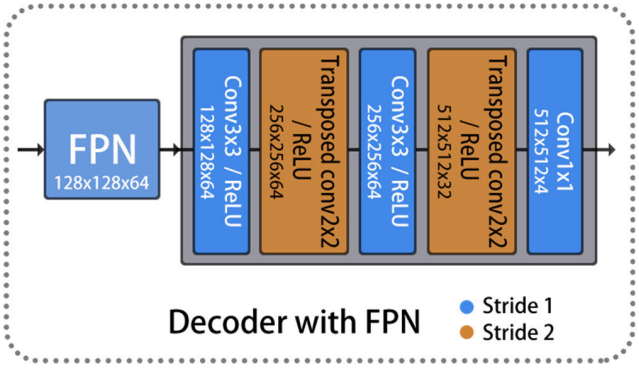

FIGURE 5. The architecture of the decoder module.

convolution layer is continuously downsampled. Therefore, a multiscale mechanism can be used to extract temporal changes and spatial details at various scales from images, resulting in more diverse features and better details preservation. FPN [37] is a multiscale fusion mechanism that uses a top-down architecture with lateral connections to construct high-level semantic feature maps at all scales, thus leveraging the inherent feature hierarchy [38]. The use of top-down and bottom-up pathways to combine high-level semantic information and low-level feature information is particularly useful for multi-scale and small object detection [39], [40]. Thus, the role of FPN in UAV-Net was investigated. The last four-layer feature maps output by the MResNet module are fed into the FPN for multiscale feature fusion, resulting in an output feature map with a dimensional size of 128 × 128 (Fig. 4). In UAV-Net, the FPN module offers several advantages. It enables the features of different scales from the MResNet module to be directly connected into the decoder module, allowing each stage to independently learn features of different size. For example, when the features learned in stage 1 are sufficient for the fusion task, they could be directly used in the decoder module, reducing the learning burden on stages 2 to 4. Conversely, when the FPN module is not used, the increase in receptive field size can potentially introduce more noise and affect the final results. Overall, the FPN module, positioned between the encoder and decoder modules, provides flexibility in optimizing the extracted features for UAV image reconstruction. Note that the batch normalization layer is removed from the FPN module, as previous studies have demonstrated that it causes significant color patches in the predicted image that are difficult to remove [41], [42].

D. DECODER

In image reconstruction tasks, the high-level features extracted from the input must be upsampled to match the scale of the original input images. Several methods exist for this upsampling purpose, including transposed convolution, nearest neighbor interpolation, and bilinear interpolation. However, only transposed convolution incorporates data learning and thus is considered a general technique. Transposed convolution uses learnable parameters to increase the size of the input feature map, making it the most effective method for upsampling abstract representations. Consequently, transposed convolution is commonly used in DL networks that require image reconstruction. However, it is important to note that if the parameters are not properly set, a feature map with checkerboard artifacts may be produced [43]. DL networks are typically reconstructed images with multiple layers of transposed convolution, constructing high-resolution images from low-resolution images through iterative processes [44].

As shown in Fig. 5, the output of the FPN module, which has a feature map size of 128 × 128, is used as input in the decoder module. The decoder module used two transposed convolutions, each with a 2 × 2 filter and a stride size of 2, to upsample the extracted feature map. The final layer is a 1 × 1 convolutional layer that transforms the number of channels to four to reconstruct a UAV image with four spectral bands.

E. LOSS FUNCTION

In this study, the errors between the predicted UAV image and the captured true UAV image were evaluated using a weighted combination of the structural similarity index measure (SSIM) and L1 loss, as shown in Equation (1). The SSIM loss function quantifies differences in luminance, contrast, and structure between the predicted and ground truth images [45]. Meanwhile, the L1 loss measures the mean absolute error between the predicted image and ground truth images [46]. To ensure the predicted high-resolution UAV image retained both the structural similarity and the data distribution and to avoid bias during the optimization process, equal weight was assigned to the two loss functions.

$$\mathcal{L}_{UAVNet} = \mathcal{L}_{SSIM} + \mathcal{L}_1 \tag{1}$$

F. NORMALIZED DIFFERENCE VEGETATION INDEX

The NDVI was used to quantitatively and qualitatively evaluate the spectral band information of the predicted UAV image for crop monitoring. The calculation of NDVI is shown in Equation (2), where NIR represents the near-infrared band and RED represents the red band. The NDVI is a widely used index for determining the richness and health of vegetation, with values ranging from -1 to 1.

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{2}$$

TABLE 1. Evaluation metrics and their Formula for fusion result performance.

Quality metrics	Formula
SSIM↑	$\frac{(2\mu_{\hat{y}_i}\mu_{y_i}+c_1)(2\sigma_{\hat{y}_i y_i}+c_2)}{(\mu_{\hat{y}_i}^2+\mu_{y_i}^2+c_1)(\sigma_{\hat{y}_i}^2+\sigma_{y_i}^2+c_2)}$
CC↑	$\frac{n(\sum \hat{y}_i y_i)-(\sum \hat{y}_i)(\sum y_i)}{\sqrt{[n(\sum \hat{y}_i^2-(\sum \hat{y}_i)^2)]*[n(\sum y_i{}^2-(\sum y_i)^2)]}}$
SAM↓	$cos^{-1}\left(\frac{y_i^T \hat{y}_i}{\sqrt{(\hat{y}_i)^T \hat{y}_i}\sqrt{(y_i)^T y_i}}\right)$
RMSE↓	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$

* ↑ indicates that larger is better, ↓ indicates that smaller is better.

G. EVALUATION METRICS

In addition, the performance of UAV-Net was evaluated using four commonly used quantitative evaluation metrics, as outlined in Table 1. These metrics include the structural similarity index measure (SSIM) [47], correlation coefficient (CC), spectral angle mapper (SAM), and root mean square error (RMSE). Among the equation of these metrics, y represents the value of the ground truth image, ŷ represents predicted image value. The SSIM quantifies the differences in luminance, contrast, and structure between corresponding pixels in fused UAV images and ground truth UAV images [45]. A higher SSIM value indicates higher similarity between the images. The mean intensity, standard deviation, and covariance of the fused images and the ground truth image are represented by $\mu_{\hat{y}}$, μ_y, $\sigma_{\hat{y}}^2$, σ_y^2, $\sigma_{\hat{y}y}$, respectively. The constants c1 and c2 guarantee that the SSIM value is between -1 and 1. The CC metric is used to indicate the linear relationship between pixels in fused and ground truth UAV images. The SAM metric computes the average angle between the spectra of corresponding pixels in two images, with a lower value indicating higher fusion accuracy [48]. Additionally, the RMSE is frequently used to quantify fusion errors, with a lower value indicating higher fusion accuracy [49]. Further information on image quality assessment metrics can be found in [50].

III. EXPERIMENTS AND RESULTS

A. STUDY AREA AND DATASETS

A corn field on Yao Farm in Iwamizawa, Hokkaido, Japan (Fig. 6) served as the study area. UAV-Net was trained and evaluated using paired UAV and PlanetScope images that have been captured over the corn field during the growing season of 2021. The duration of the corn growth cycle spanned from the end of May to the beginning of September.

PlanetScope is a satellite constellation operated by Planet (https://www.planet.com). PlanetScope provides high temporal resolution imaging with near-daily coverage, as well as global coverage images that are analysis-ready for real-time analysis. The spatial resolution of PlanetScope is 3m. In this study, the blue, green, red, and NIR bands of PlanetScope images were used for STF. Table 2 summarizes key information on the PlanetScope bands.

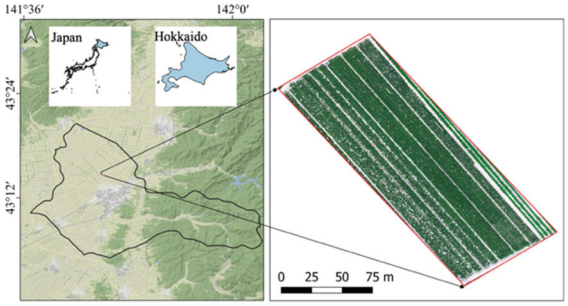

FIGURE 6. Location of the corn field and its corresponding UAV image (RGB composite).

TABLE 2. Bands information of PlanetScopeScope images.

PlanetScope Band	Wavelength (nm)	Spatial Resolution	Temporal Resolution
Blue	455-515	3 m	1 day
Green	500-590		
Red	590-670		
NIR	780-860		

TABLE 3. Acquisition dates (YYYY/MM/DD) of UAV and PlanetScope images for STF.

UAV image		PlanetScope image
DJI Phantom 4 Multispectral	MicaSense RedEdge-MX	
2021/06/26	2021/06/26	2021/06/26
2021/07/10	2021/07/10	2021/07/09
2021/07/17	2021/07/17	2021/07/17
	2021/07/27	2021/07/27
	2021/07/31	2021/07/31
	2021/08/07	2021/08/07

In addition, DJI Phantom 4 Multispectral and MicaSense RedEdge-MX cameras were used to capture high-resolution UAV multispectral images from June to August 2021. The spatial resolution of the UAV was approximately 2 cm. The corresponding blue, green, red, and NIR bands from UAV multispectral images were used for the STF. Table 3 provides the list of UAV and PlanetScope images available for the 2021 corn growing season. In total, 32 sets of cross-paired UAV and PlanetScope images were generated, and each image set with two UAV-PlanetScope image pairs.

B. EXPERIMENTAL SETTINGS

To assess the efficacies of UAV-Net and its building modules, comparative and ablation experiments were performed. The comparative experiment evaluated the accuracy and efficiency of three MResNet configurations with 18, 34, and 50 layers. The ablation experiment examined the effects of the FPN module on UAV-Net performance. To address the reduction in feature map size when the FPN module was ablated, three additional transposed convolution layers were

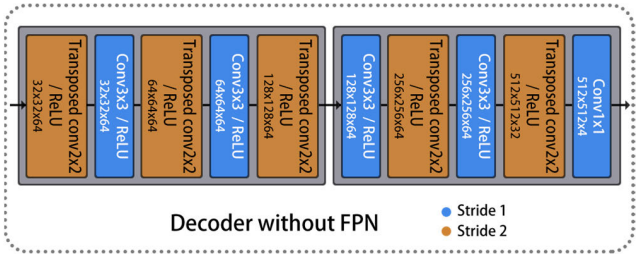

FIGURE 7. Decoder architecture for the experiment without the FPN module.

incorporated into the decoder module (Fig. 7). Furthermore, the fusion accuracy of UAV-Net was compared to three state-of-the-art algorithms. These algorithms include the enhanced Deep Convolutional SpatioTemporal Fusion Network (EDCSTFN) [22], HIgh-resolution SpatioTemporal Image Fusion (HISTIF) [51], and improved HISTIF (IHISTIF) [52] fusion algorithm. The EDCSTFN is a deep learning-based STF model that uses ''encoder-merge-decoder'' architecture. HISTIF is a non-deep learning algorithm that has demonstrated good performance in high-resolution crop monitoring at the sub-field level. IHISTIF aims to enhance the performance of HISTIF.

Among the 32 sets of images, 30 were used for training and validation, while the remaining two were used for testing. This is to ensure that the training dataset contained a diverse range of image pairs that represented various time spans of crop growth stages for training a generalized DL model. Notably, the test datasets were not used in the training process. The two sets of test images were used to produce UAV images on July 31 and August 7, 2021. The corresponding true images captured by the UAV on these two dates served as the ground truths for the evaluation.

To facilitate the training process, the pixels of UAV and PlanetScope images were normalized to a standardized value range of 0 to 1 for all four spectral bands. The PlanetScope images were up-sampled to match the spatial resolution of the UAV image using the bilinear interpolation method. The 30 image sets used for training and validation were then clipped into 512 $\times$ 512 patches to reduce their size before being fed into the UAV-Net. The experiments were conducted using the PyTorch framework on a computer with an Intel(R) Core (TM) i9-12900K central processing unit (CPU) @3.20 GHz, NVIDIA GeForce RTX 3090, and 48 GB RAM configuration. The validation dataset was created by randomly selecting 10% of each of the 30 datasets that were clipped into patches. Validation was performed every 10 epochs using a cross-validation method, and the SSIM metric was used to identify the best model. The following hyperparameters were used in this study: the batch size was set to 32, and the initial learning rate was 0.001, while the WarmupPolyLR learning scheduler was used with a warmup epoch of 3. All training was conducted for 300 epochs using the rectified linear unit (ReLU) activation function and the Adam optimizer.

Note that the UAV image did not cover a square or rectangular area, there are significant areas with no data values. To ensure a fair and accurate evaluation of the fusion results, the predicted images were clipped into small patches, excluding patches with no data value for the evaluation. Four metrics were used to evaluate individual bands, and the mean metrics values of the four bands were also calculated.

TABLE 4. Objective evaluation of mresnet modules with 18, 34, and 50 layers using cross-validation.

		MResNet18	MResNet34	MResNet50
SSIM ↑	Blue	0.9502	0.9503	**0.9504**
	Green	0.9414	0.9417	**0.9418**
	Red	0.9052	0.9059	**0.9066**
	NIR	0.7291	0.7301	**0.7325**
	Mean	0.8815	0.8820	**0.8828**
CC ↑	Blue	0.8399	0.8455	**0.8461**
	Green	0.7998	**0.8062**	0.8039
	Red	0.8566	0.8605	**0.8621**
	NIR	0.7435	0.7470	**0.7523**
	Mean	0.8099	0.8148	**0.8161**
SAM ↓		0.0751	**0.0750**	0.0751
RMSE ↓	Blue	0.0156	0.0156	**0.0155**
	Green	0.0141	**0.0139**	0.0140
	Red	0.0228	0.0224	**0.0223**
	NIR	0.0695	0.0691	**0.0688**
	Mean	0.0305	**0.0302**	**0.0302**

C. COMPARISON AND RESULTS

1) COMPARATIVE STUDY

Table 4 presents a summary of the results on the validation dataset, exhibiting the performance of MResNet 18, 34, and 50 based on four evaluation metrics. The best values are highlighted in bold font. It was observed that the MResNet modules with 50 layers achieved slightly better results compared to the modules with 18 and 34 layers. This is evident from the mean value across most metrics, except for a marginal difference of 0.0001 in SAM compared to MResNet 34. The better performance of the deeper MResNet module can be attributed to the fact that deeper layers can better fit the data and thus exhibit better performance. However, an MResNet module with even more layers was not compared due to the potential risk of overfitting. The MResNet modules with 18, 34, and 50 layers combine with the decoder module underwent a training and validation process that lasted 14h8m1s, 14h26m58s, and 18h32m21s, respectively.

Fig. 8 presents a visual comparison of ground truths (column a), and fusion results (columns b, c, and d) of RGB bands composition on two test image sets. The images in columns b, c, and d were predicted using MResNet 18, 34, and 50, respectively. The images from July 31 and August 7 were presented in the first and second rows, respectively. It can be observed that the fusion results closely resembled the ground truths in terms of shape, texture, color, and brightness,

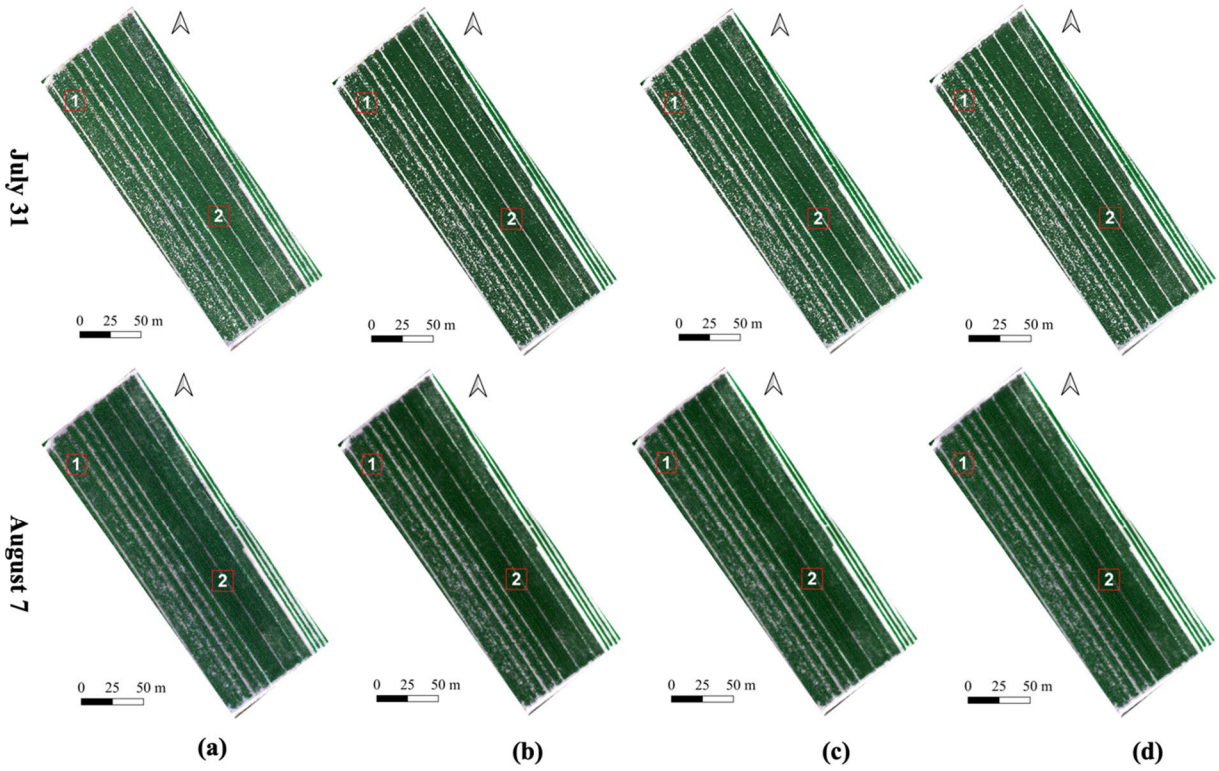

FIGURE 8. Red-green-blue composition of the fusion results for July 31, 2021 (first row), and August 7, 2021 (second row), with different MResNet layers. (a) Ground truth, (b) MResNet 18, (c) MResNet 34, (d) MResNet 50.

highlighting the UAV image fusion capacities of the MResNet with the decoder architecture. Additionally, two randomly selected red-boxed areas are magnified regions, allowing a more detailed comparison of the fusion results.

Fig. 9 displays two magnified areas, with the top two rows showing images from July 31, and the bottom two rows showing images from August 7. By the visual inspection, it is evident that the predicted images from August 7 exhibit better performance than those from July 31 compare to their respective ground truths. Notably, the highlighted red circle area in the topmost row indicates that MResNet 50 (column d) optimally predicted UAV images that were more accurate compared to images predicted using other MResNet modules (columns b, c). Furthermore, the MResNet module with more layers produced sharper images, as observed in the fusion images displayed in columns b, c, and d.

The feasibility of the fusion results for crop monitoring was further validated by visually examining NDVI maps and corresponding error maps. The NDVI maps and their error maps for July 31 (Fig. 10) and August 7 (Fig. 11) based on MResNet 18, 34, and 50 are shown in columns b, c, and d, respectively, and ground truth NDVI maps are shown in column a. The NDVI error maps in the first rows of Figs. 10 and 11 were obtained by subtracting the ground truth NDVI values from the fusion results NDVI values, and the error map values were stretched to the range of 0 and 0.2 to emphasize errors. This error map stretching strategy has also been used in previous studies [49], [23]. An examination of the overall distributions of NDVI values reveals that NDVI maps produced using the MResNet with the decoder modules exhibited a closer resemblance to the ground truths. Additionally, the NDVI error maps were dominated by light blue color, indicating minimal errors and thus confirming the reliabilities of the predicted red and NIR bands for crop monitoring. However, the error maps for July 31 exhibit more error noise representation compared with the error maps for August 7.

The visual comparison of the fusion results obtained from MResNet with 18, 34, and 50 layers on two test image sets demonstrated that each combination of MResNet with the decoder module produced visually satisfactory images. To further evaluate and compare their performance, quantitative metrics were calculated for two test image sets, as shown in Table 5. The results indicate that the MResNet 50 outperformed both MResNet 18 and 34, with the highlighted values across most metrics for the two test image sets. However, the SAM values were higher for MResNet 50 compare to MResNet 18 and 34. Moreover, the inferior performance of the NIR band, as indicated by poorer SSIM and RMSE values compared to RGB bands, could be attributed to differences in the distribution of spectral band data. The distribution of NIR band data significantly differs from that of the RGB bands. Additionally, the values of four metrics for the fusion image on August 7 are consistently better than those for the fusion image on July 31, which is consistent with the visual observation results.

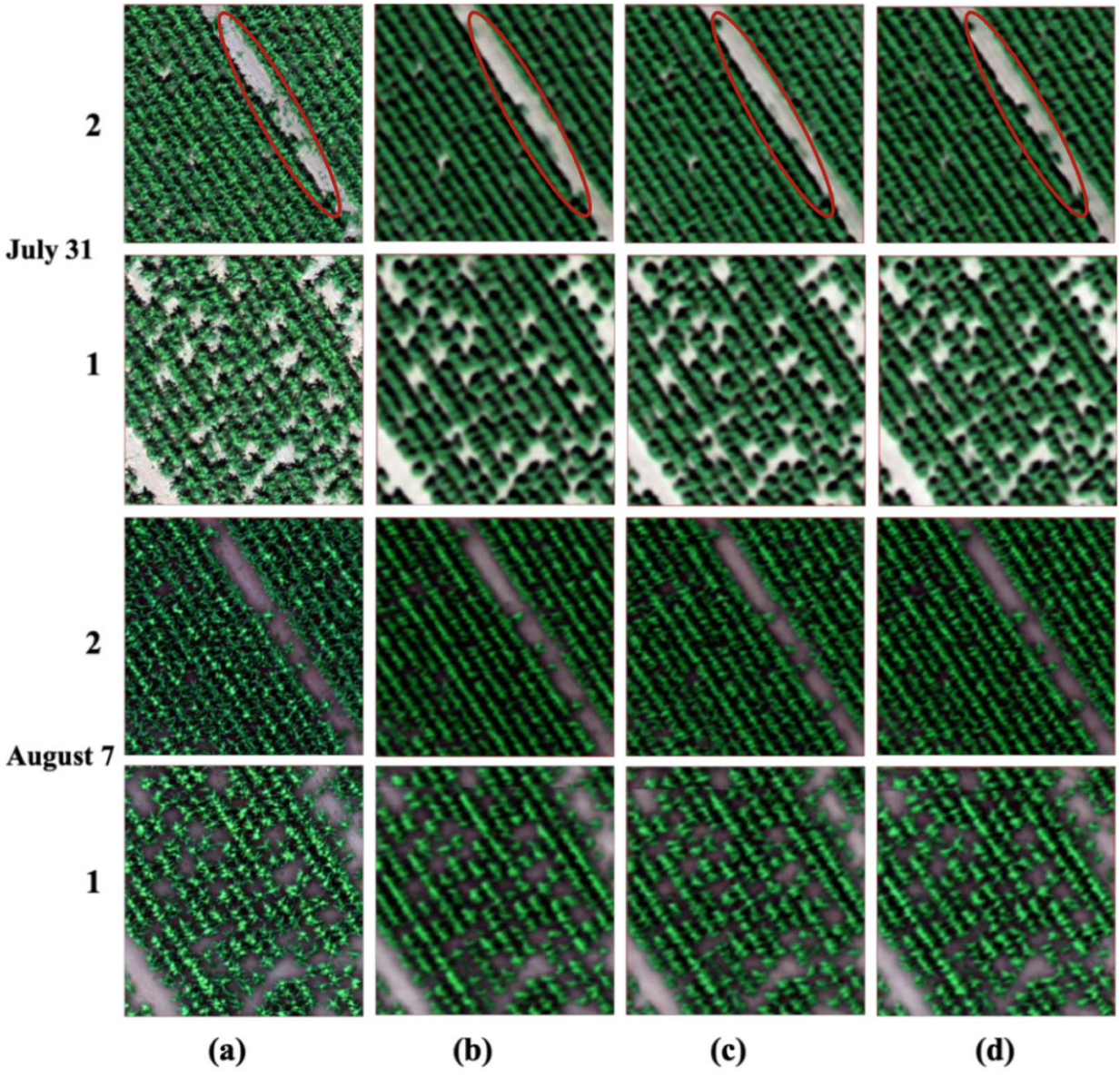

FIGURE 9. Images of two magnified areas from July 31 (top two rows) and August 7 (bottom two rows) using different MResNet modules. (a) Ground truth, (b) MResNet 18, (c) MResNet 34, (d) MResNet 50.

2) ABLATION EXPERIMENT

The comparative study revealed that MResNet 50 was optimal as the MResNet module for UAV-Net. Therefore, MResNet 50 was used as the backbone for the ablation experiment that aimed to investigate the effect of FPN on the fusion performance. An FPN module is used before the decoder module to enable multiscale fusion.

Table 6 summarizes the results for the four metrics on the validation dataset using MResNet 50, and MResNet 50 with an FPN module. A comparison between MResNet 50 alone and MResNet 50 with the FPN module revealed that the FPN module improved all of the metrics values. Notably, the improvements achieved by incorporating the FPN module surpassed the improvements observed by using deeper layers in MResNet. This finding highlights the beneficial effect of multiscale fusion on UAV-Net performance. Furthermore, the training and validation processing time of UAV-Net is 20h40m20s.

Fig. 12 provides a visual comparison of the ground truth (column a) and the fusion images obtained from MResNet 50 (column b), and MResNet 50 with FPN (column c) in the RGB bands composition on two test image sets. The images display in the first and second rows from July 31 and August 7, respectively. The fusion images resulting from the ablation experiment exhibit a remarkable resemblance to the ground truth images in terms of shape, texture, brightness, and color. This observation demonstrated the effectiveness of FPN modules in UAV image prediction. The two red-boxed areas are magnified regions used for closer inspection, as shown in Fig. 13.

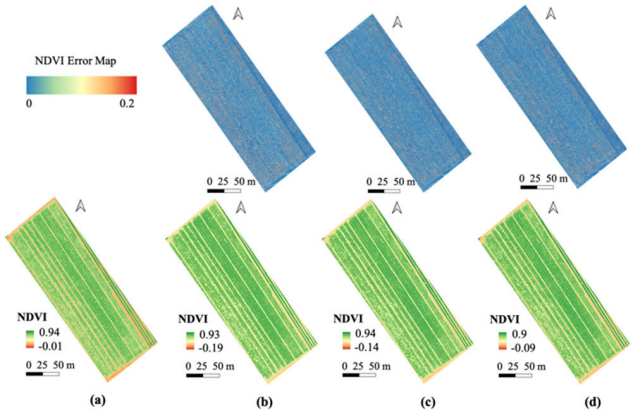

FIGURE 10. The NDVI maps of fusion result on July 31 and their error maps based on (a) ground truth. (b) MResNet 18. (c) MResNet 34. (d) MResNet 50.

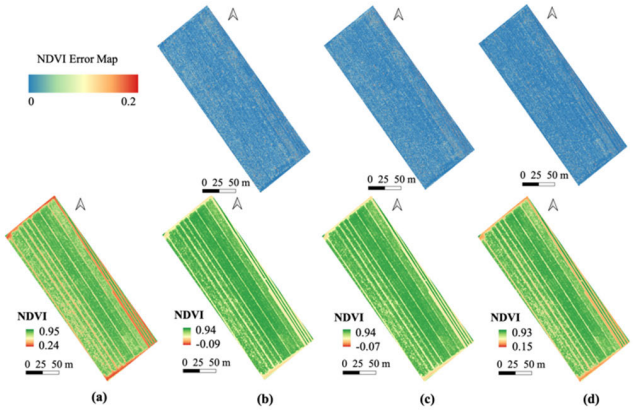

FIGURE 11. NDVI maps of the fusion result for August 07 and their error maps based on (a) ground truth, (b) MResNet 18, (c) MResNet 34, (d) MResNet 50.

TABLE 5. Objective evaluations of mresnet modules 18, 34, and 50 based on Test datasets.

		August 7			July 31		
		MResNet18	MResNet34	MResNet50	MResNet1	MResNet34	MResNet50
SSIM↑	Blue	**0.9708**	0.9705	0.9702	0.8935	0.8933	**0.8944**
	Green	**0.9491**	0.9477	0.9482	0.8150	0.8139	**0.8159**
	Red	**0.9682**	0.9680	0.9680	0.8296	0.8292	**0.8311**
	NIR	0.6323	0.6267	**0.6355**	0.4954	0.4922	**0.4972**
	Mean	0.8801	0.8782	**0.8805**	0.7584	0.7572	**0.7597**
CC↑	Blue	0.7191	0.7188	**0.7254**	0.7352	0.7385	**0.7423**
	Green	0.6543	0.6451	**0.6589**	0.6668	0.6659	**0.6756**
	Red	0.8271	0.8280	**0.8320**	0.7822	0.7872	**0.7899**
	NIR	0.7941	0.7904	**0.8020**	0.6555	0.6545	**0.6636**
	Mean	0.7487	0.7456	**0.7546**	0.7099	0.7115	**0.7179**
SAM↓		**0.0310**	**0.0310**	0.0318	**0.0589**	**0.0589**	0.0590
RMSE↓	Blue	**0.0070**	**0.0070**	**0.0070**	0.0162	0.0160	**0.0159**
	Green	0.0147	0.0148	**0.0146**	0.0269	0.0266	**0.0263**
	Red	0.0111	0.0110	**0.0109**	0.0296	0.0291	**0.0288**
	NIR	0.1037	0.1035	**0.1018**	0.0987	**0.0971**	0.0974
	Mean	0.0341	0.0341	**0.0336**	0.0429	0.0422	**0.0421**

The images in the top two rows of Fig. 13 were from July 31, while the images in the bottom two rows were from August 7. Visual examination demonstrated that the fusion results obtained using MResNet 50 alone and MResNet 50 with FPN had a high similarity. All of the models predicted UAV images with good color preservation and smooth backgrounds.

TABLE 6. Objective evaluation of mresnet 50, and mresnet 50 with FPN model based on cross-validation.

		MResNet50	MResNet50+FPN
SSIM ↑	Blue	0.9504	**0.9512**
	Green	0.9418	**0.9423**
	Red	0.9066	**0.9076**
	NIR	0.7325	**0.7347**
	Mean	0.8828	**0.8840**
CC ↑	Blue	0.8461	**0.8560**
	Green	0.8039	**0.8175**
	Red	0.8621	**0.8696**
	NIR	0.7523	**0.7613**
	Mean	0.8161	**0.8261**
SAM ↓		0.0751	**0.0739**
RMSE ↓	Blue	0.0155	**0.0154**
	Green	0.0140	**0.0137**
	Red	**0.0223**	**0.0223**
	NIR	0.0688	**0.0679**
	Mean	0.0302	**0.0298**

The fusion results by MResNet 50 with FPN (column c) were also validated by their NDVI maps and error maps. A comparison was made between the NDVI map of the ablation experiment fusion results, the NDVIs map of the ground truth image (column a), and the NDVI maps of MResNet 50 (column b). Figs. 14 and 15 present the NDVI maps of July 31 and August 7, respectively. The NDVI value range for the fusion results on July 31 was close to the ground truths, but the NDVI error maps revealed a significant visual noise. In contrast, although the NDVI range for the fusion results on August 7 was less similar to the ground truths, the error maps exhibited significantly less noise. The difference in error noise between the error maps of July 31 and August 7 suggested the potential influence of input qualities. Nevertheless, it was observed that the NDVI values of the soil background (without crops) in the fused images of August 7 were much lower than the ground truths compared to those of July 31.

The fusion results obtained from the ablation experiment on the two test image sets were objectively evaluated using four metrics (Table 7). Incorporating the FPN module into UAV-Net resulted in better mean values across all four metrics, indicating the effectiveness of multiscale feature fusion in UAV image reconstruction. The values of SSIM and RMSE for the NIR band were consistently poorer on both test sets than those for the RGB bands, indicating the need for further attention and investigation to address this issue in the future. Additionally, the mean values of four metrics were better for the fusion results on August 7 than those on July 31, which is consistent with the results of the comparative experiment.

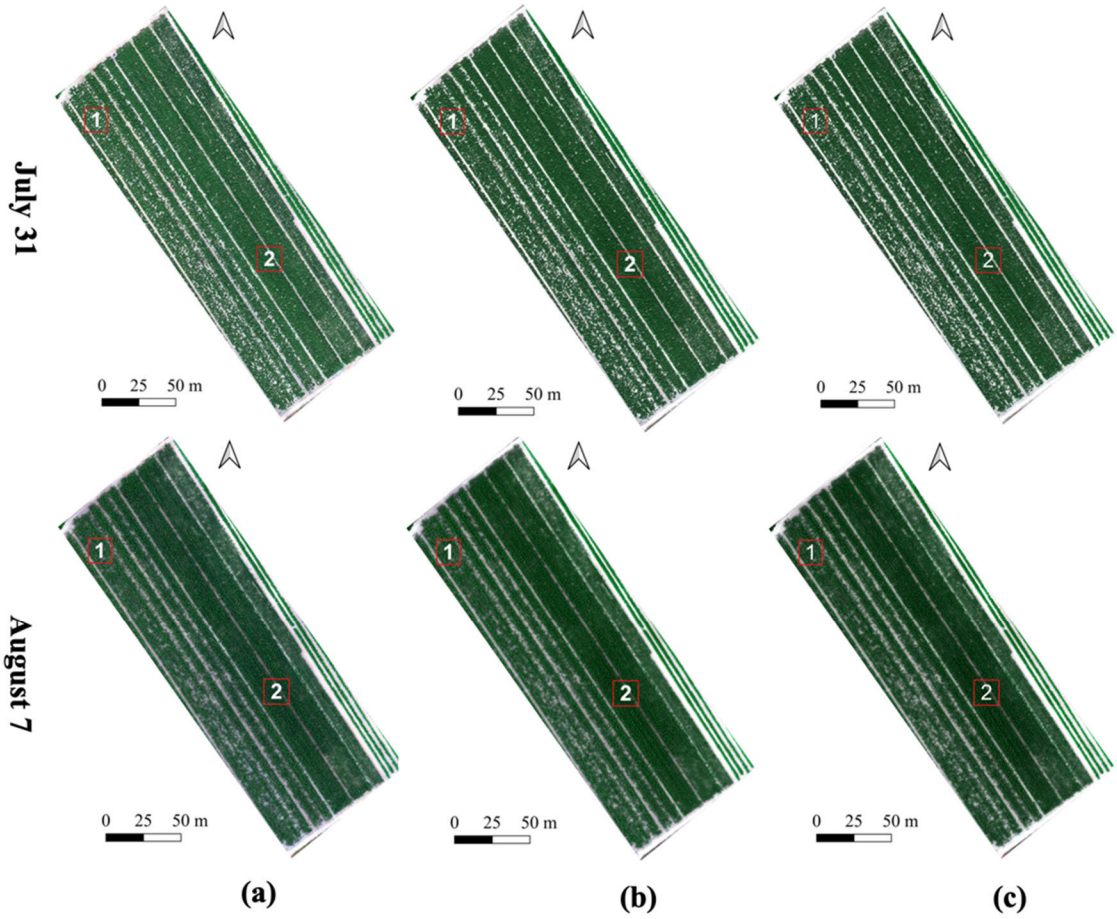

FIGURE 12. Red-green-blue composition of the fusion results on July 31 (first row) and August 7 (second row) using different MResNet alone or augmented. (a) Ground truth, (b) MResNet 50, and (c) MResNet 50+FPN.

Therefore, it is indicated that input image quality significantly affects the quality of fusion images.

3) MODEL ACCURACY COMPARISON

The two test image sets were used to assess the accuracy of the EDCSTFN, HISTIF, and IHISTIF models. Fig. 16 illustrates the visual comparison results of the EDCSTFN model (column c) in comparison to the ground truth images (column a) and the results from UAV-Net (column b). The fusion results produced by EDCSTFN exhibit noticeable dissimilarities when compared to the ground truth images. Fig. 17 displays two magnified areas to compare the fusion results in more detail, revealing that EDCSTFN did not effectively learn accurate information for the UAV and Planet image fusion.

Table 8 compares four metrics between UAV-Net and EDCSTFN on two test sets. In terms of SSIM value, EDCSTFN exhibits relatively high values in RGB bands, possibly due to the datasets being from a crop field where changes are less in structures but rather in crop growth. However, the CC values in RGB bands indicate a lower correlation between the ground truth and fused image obtained by EDCSTFN. Moreover, the significant difference in the SAM values between UAV-Net and EDCSTFN, suggests that EDCSTFN has limited capability in capturing the spectral information changes from UAV and Planet images. Furthermore, in terms of processing time, CPU-based predictions for a 512×512 UAV image take 0.38s with UAV-Net and 1.89s with EDCSTFN. When using the GPU, the times decrease to 0.02s for UAV-Net and 1.00s for EDCSTFN. UAV-Net has 23,740,308 parameters, while EDCSTFN has 447,972 parameters. Despite having a higher parameter count, UAV-Net demonstrates greater computational efficiency due to the downsampling process in the encoder module, which consistently reduces the feature map size. In contrast, EDCSTFN consists only of stacked convolutional layers without a downsampling process.

Fig.18 presents visual comparison results of the HISTIF (column b) and IHISTIF (column c) models. Due to the requirement of square-shaped input images for HISTIF and IHISTIF models, the fusion results only represent a portion of the corn field. The two non-deep learning-based STF models do not require a training process and directly input two PlanetScope images and one UAV image into their model. However, they inadequately capture the changes between reference to prediction dates. Notably, the fusion results keep

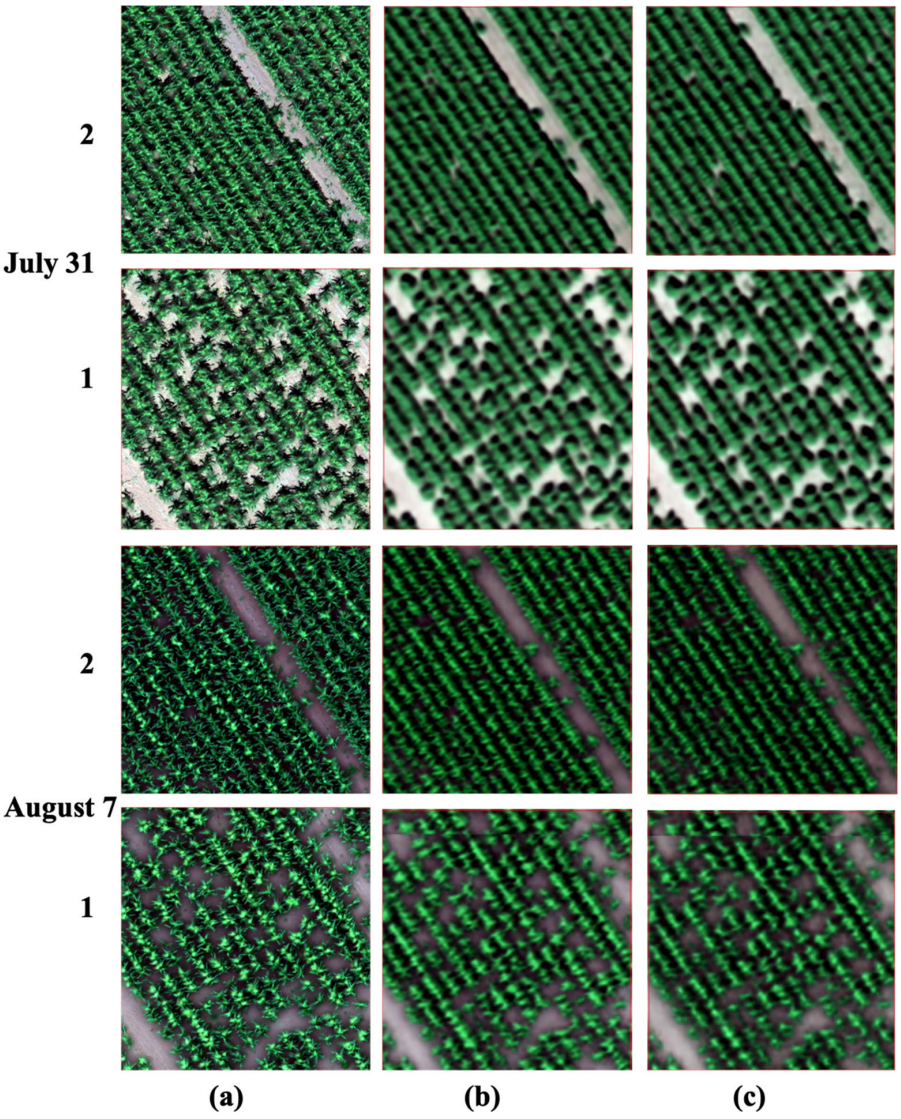

FIGURE 13. Magnified area images from August 7 (bottom two rows) and July 31 (top two rows). (a) Ground truth, (b) MResNet 50, and (c) MResNet 50+FPN.

the same as the input reference UAV images (column d) rather than resemble the ground truth (column a), indicating the limitation of the HISTIF and IHISTIF in accurately fusing UAV and satellite images for crop monitoring.

Note that existing proposed STF models, such as EDC-STFN, HISTIF, and IHISTIF, were designed for fusing satellite images with meter-scale fusion results. Hence, using these STF models directly for centimeter-scale image prediction presents challenges. This highlights the innovation and effectiveness of UAV-Net for centimeter-scale UAV image fusion.

D. LIMITATIONS AND FUTURE SCOPE

Currently, there are no benchmark datasets available that pair UAV and satellite images for STF, thereby hindering comparative evaluations of the performance of UAV-Net in crop monitoring. The lack of benchmark datasets and the fact that existing DL-based STF models are typically designed for fusing satellite images with scales differing by less than 16-fold, pose a challenge when attempting to compare the performance of UAV-Net (handles 150-fold scale difference)

TABLE 7. Objective evaluation of the performance of mresnet 50 alone, and mresnet 50 with fpn using two Test datasets.

		August 7		July 31	
		MResNet50	MResNet50+FPN	MResNet50	MResNet50+FPN
SSIM↑	Blue	0.9702	**0.9718**	0.8944	**0.8961**
	Green	0.9482	**0.9500**	0.8159	**0.8181**
	Red	0.9680	**0.9697**	0.8311	**0.8333**
	NIR	0.6355	**0.6382**	0.4972	**0.5017**
	Mean	0.8805	**0.8824**	0.7597	**0.7623**
CC↑	Blue	0.7254	**0.7321**	0.7423	**0.7520**
	Green	0.6589	**0.6638**	0.6756	**0.6843**
	Red	0.8320	**0.8388**	0.7899	**0.7995**
	NIR	0.8020	**0.8032**	0.6636	**0.6694**
	Mean	0.7546	**0.7595**	0.7179	**0.7263**
SAM↓		0.0318	**0.0301**	0.0590	**0.0571**
RMSE↓	Blue	0.0070	**0.0068**	0.0159	**0.0157**
	Green	0.0146	**0.0144**	0.0263	**0.0261**
	Red	0.0109	**0.0106**	0.0288	**0.0284**
	NIR	0.1018	**0.1017**	0.0974	**0.0972**
	Mean	0.0336	**0.0334**	0.0421	**0.0419**

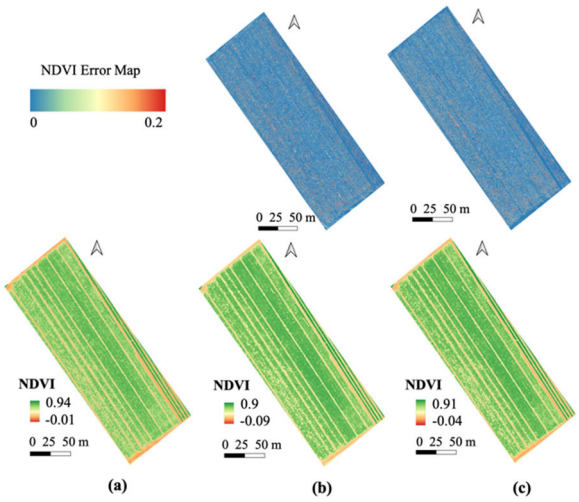

FIGURE 14. NDVI maps of fusion results for July 31 and their error maps. (a) Ground truth, (b) MResNet 50, and (c) MResNet 50+FPN.

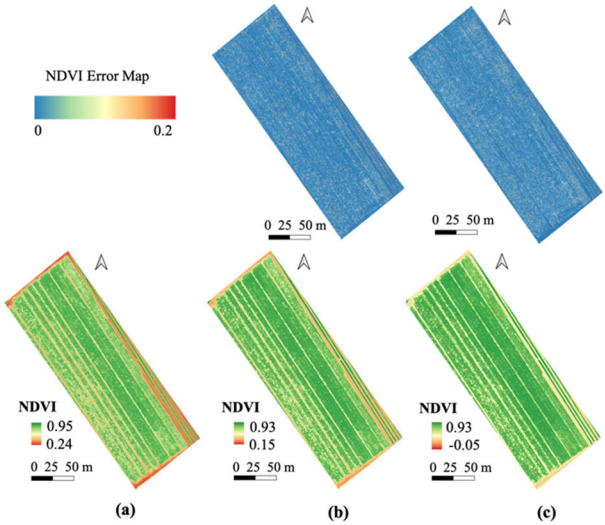

FIGURE 15. NDVI maps of fusion results for August 07 and their error maps. (a) Ground truth, (b) MResNet 50, (c) MResNet 50+FPN.

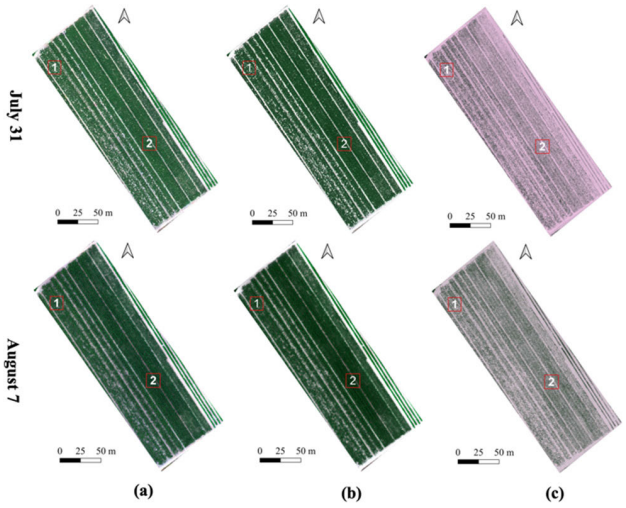

FIGURE 16. Red-green-blue compositions of the fusion results on July 31 (first row) and August 7 (second row). (a) Ground truth, (b) UAV-Net, and (c) EDCSTFN.

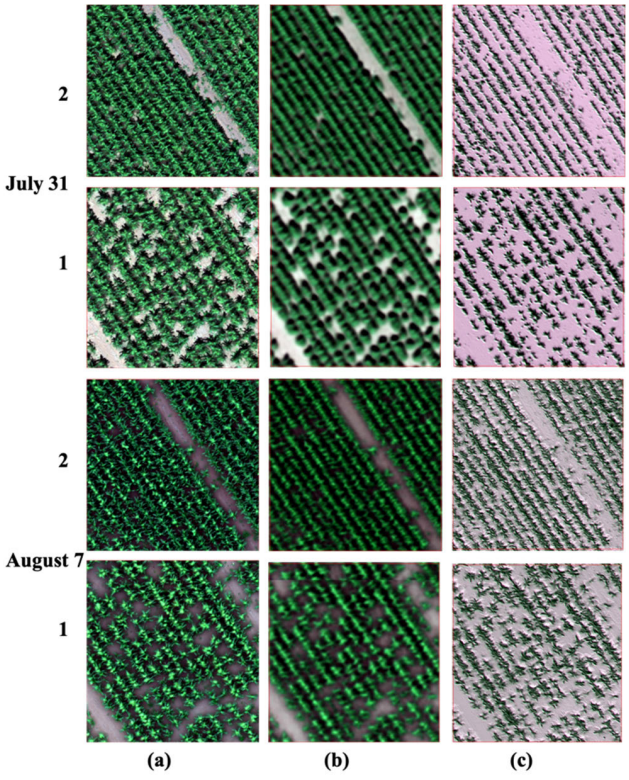

FIGURE 17. Magnified area images from August 7 (bottom two rows) and July 31 (top two rows). (a) Ground truth, (b) UAV-Net, and (c) EDCSTFN.

with other STF models. Moreover, the interval between the reference and prediction dates of the training datasets used in this study do not all lie within the difference span of the test dataset derivation, which may reduce fusion accuracy. Thus, it is suggested that the interval between the reference and the prediction dates of the future UAV-satellite STF benchmark datasets should be varied to improve STF model generalizability. Additionally, as discussed previously [53], [54], high-quality satellite image datasets are important to ensure STF performance. This study found that different test images yielded different fusion results, and high-quality input UAV images improved UAV image prediction. Some UAV images in the training datasets in this study exhibited mosaic distortion (Fig. 17), which was caused during three-dimensional photogrammetry reconstruction. Therefore, the repair of both distortion and noise during UAV image reconstruction is important to achieve accurate UAV image prediction. Furthermore, the results showed that the performance of the NIR band in the predicted UAV images is comparatively lower than that of RGB bands, which requires further investigation.

Note that the proposed UAV-Net adequately predicted UAV images for crop monitoring because the training dataset focused on crop phenological changes. This is due to deep neural networks performing well in learning data distribution patterns. Therefore, UAV-Net may predict land cover changes

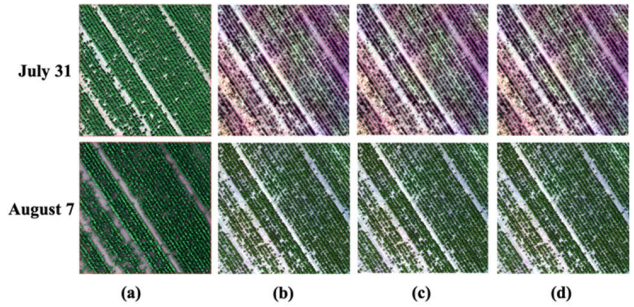

FIGURE 18. Red-green-blue compositions of the fusion results on July 31 (first row) and August 7 (second row). (a) Ground truth, (b) HISTIF, (c) IHISTIF, and (d) input UAV images.

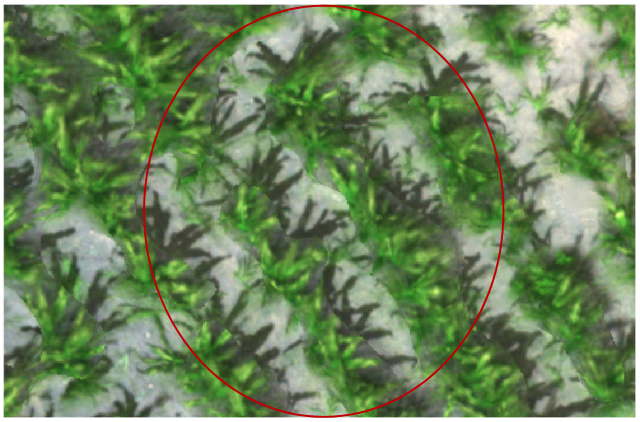

FIGURE 19. Distortion of a UAV image of the training dataset.

TABLE 8. Objective evaluation of the performance of UAV-Net and EDCSTFN using two Test datasets.

		August 7		July 31	
		UAV-Net	EDCSTFN	UAV-Net	EDCSTFN
SSIM↑	Blue	**0.9718**	0.9182	**0.8961**	0.8379
	Green	**0.9500**	0.9210	**0.8181**	0.7691
	Red	**0.9697**	0.8381	**0.8333**	0.7484
	NIR	**0.6382**	0.5219	**0.5017**	0.4197
	Mean	**0.8824**	0.7998	**0.7623**	0.6938
CC↑	Blue	**0.7321**	0.1318	**0.7520**	0.1706
	Green	**0.6638**	-0.0917	**0.6843**	0.0699
	Red	**0.8388**	0.2917	**0.7995**	0.2315
	NIR	**0.8032**	0.4979	**0.6694**	0.3628
	Mean	**0.7595**	0.2074	**0.7263**	0.2087
SAM↓		**0.0301**	0.1478	**0.0571**	0.1365
RMSE↓	Blue	**0.0068**	0.0144	**0.0157**	0.0268
	Green	**0.0144**	0.0220	**0.0261**	0.0392
	Red	**0.0106**	0.0384	**0.0284**	0.0507
	NIR	**0.1017**	0.2262	**0.0972**	0.1531
	Mean	**0.0334**	0.0753	**0.0419**	0.0674

if training dataset images show such changes, but prediction accuracy may be low if the training dataset does not teach the model to learn specific data distribution patterns. Moreover, crop randomness and external environmental factors can create significant uncertainty in the growth of plant branches and leaves. Therefore, DL networks cannot accurately predict the growth of specific branches and leaves of crops but can detect overall growth trends.

Considering the computational constraints and the high resolution of the UAV image over the corn field that is larger than 2 GB, it is not feasible to predict the entire UAV image during the fusion process. Instead, fused UAV images in small patches were merged to form the entire scene. As a result, the merged images are bordered by lines. Although the presence of such lines can be overlooked if the entire UAV image is examined, expanding the computation resource would resolve this issue.

In the future, further investigation into the fusion of other spectral bands, such as the red edge band, could be explored to generate other vegetation indices for crop monitoring. In addition to vegetation indices, the produced centimeter-scale UAV images have a high potential to enhance crop monitoring in various aspects. They can improve crop classification accuracy, enable effective detection of crop diseases, and facilitate time-series monitoring of crop growth. As a result, they can further enhance management practices and yield prediction accuracy. Furthermore, to improve image fusion accuracy, multimodal DL techniques can be employed that incorporate various information, such as the interval between the reference date and date of prediction, and the details of precipitation, temperature, and fertilizer. Additionally, attention mechanisms could also be used to focus on the spatial and spectral information of images, further enhancing UAV image fusion.

IV. CONCLUSION

State-of-the-art STF models can fuse images from various satellites to predict images with meter-level spatial resolutions. However, precise environmental monitoring such as crop monitoring requires images with centimeter-scale spatial resolution. Thus, the UAV-Net STF model was proposed in this study to predict centimeter-scale UAV images for crop monitoring. The comparative experiment revealed that the MResNet module with 50 layers outperformed those with 18 or 34 layers in producing UAV images. Additionally, the ablation experiment demonstrated that incorporating the FPN module further improved the fusion performance. The comparison with three STF models confirmed the superior performance of UAV-Net. Visual and quantitative assessments indicated that UAV-Net adequately fused UAV and PlanetScope images. Nevertheless, some issues remain, such as the impact of input image quality, blurry corn plant leaves in the predicted images, and bordered lines in merged images. To improve the accuracy and generalizability of STF models to predict centimeter-scale resolution images, high-quality UAV benchmark datasets are urgently needed. The results of this study show the potential of UAV-Net for precise crop monitoring, as well as broader environmental monitoring applications.

ACKNOWLEDGMENT

The authors would like to thank Yao Farm, Iwamizawa for allowing them to conduct this research.

REFERENCES

[1] M. Ji-Hua, W. Bing-Fang, and L. Qiang-Zi, "A global crop growth monitoring system based on remote sensing," in *Proc. Int. Geosci. Remote Sens. Symp.*, Jul./Aug. 2006, pp. 2277–2280.

[2] S. B. Tirado, C. N. Hirsch, and N. M. Springer, "UAV-based imaging platform for monitoring maize growth throughout development," *Amer. Soc. Plant Biol.*, vol. 4, no. 6, Jun. 2020, Art. no. e00230.

[3] H. Yao, R. Qin, and X. Chen, "Unmanned aerial vehicle for remote sensing applications–A review," *Remote Sens.*, vol. 11, no. 12, p. 1443, Jun. 2019.

[4] P. Ghamisi, B. Rasti, N. Yokoya, Q. Wang, B. Hofle, L. Bruzzone, F. Bovolo, M. Chi, K. Anders, R. Gloaguen, P. M. Atkinson, and J. A. Benediktsson, "Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 1, pp. 6–39, Mar. 2019.

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

[6] 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, Sep. 2015.

[7] Y. Chen, K. Shi, Y. Ge, and Y. Zhou, "Spatiotemporal remote sensing image fusion using multiscale two-stream convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4402112.

[8] Z. Niu, "Use of MODIS and Landsat time series data to generate high-resolution temporal synthetic Landsat data using a spatial and temporal reflectance fusion model," *J. Appl. Remote Sens.*, vol. 6, no. 1, Mar. 2012, Art. no. 063507.

[9] W. Liu, Y. Zeng, S. Li, and W. Huang, "Spectral unmixing based spatiotemporal downscaling fusion approach," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 88, Jun. 2020, Art. no. 102054.

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

[11] T. Hilker, M. A. Wulder, N. C. Coops, J. Linke, G. McDermid, J. G. Masek, F. Gao, and J. C. White, "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, Aug. 2009.

[12] Q. Wang, Y. Zhang, A. O. Onojeghuo, X. Zhu, and P. M. Atkinson, "Enhancing spatio-temporal fusion of MODIS and Landsat data by incorporating 250 m MODIS data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 9, pp. 4116–4123, Sep. 2017.

[13] Y. Rao, X. Zhu, J. Chen, and J. Wang, "An improved method for producing high spatial-resolution NDVI time series datasets with multi-temporal MODIS NDVI data and Landsat TM/ETM+ images," *Remote Sens.*, vol. 7, no. 6, pp. 7865–7891, Jun. 2015.

[14] X. Zhang, F. Gao, J. Wang, and Y. Ye, "Evaluating a spatiotemporal shape-matching model for the generation of synthetic high spatiotemporal resolution time series of multiple satellite data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 104, Dec. 2021, Art. no. 102545.

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

[16] M. Liu, Y. Ke, Q. Yin, X. Chen, and J. Im, "Comparison of five spatiotemporal satellite image fusion models over landscapes with various spatial heterogeneity and temporal variation," *Remote Sens.*, vol. 11, no. 22, p. 2612, Nov. 2019.

[17] J. Cai, B. Huang, and T. Fung, "Progressive spatiotemporal image fusion with deep neural networks," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 108, Apr. 2022, Art. no. 102745.

[18] W. Li, C. Yang, Y. Peng, and J. Du, "A pseudo-Siamese deep convolutional neural network for spatiotemporal satellite image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1205–1220, 2022.

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

[20] Y. Ma, J. Wei, W. Tang, and R. Tang, "Explicit and stepwise models for spatiotemporal fusion of remote sensing images with deep neural networks," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 105, Dec. 2021, Art. no. 102611.

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

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

[23] B. Song, P. Liu, J. Li, L. Wang, L. Zhang, G. He, L. Chen, and J. Liu, "MLFF-GAN: A multilevel feature fusion with GAN for spatiotemporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4410816.

[24] Z. Yang, C. Diao, and B. Li, "A robust hybrid deep learning model for spatiotemporal image fusion," *Remote Sens.*, vol. 13, no. 24, p. 5005, Dec. 2021.

[25] W. Li, D. Cao, Y. Peng, and C. Yang, "MSNet: A multi-stream fusion network for remote sensing spatiotemporal fusion based on transformer and convolution," *Remote Sens.*, vol. 13, no. 18, p. 3724, Sep. 2021.

[26] J. Wu, L. Lin, T. Li, Q. Cheng, C. Zhang, and H. Shen, "Fusing Landsat 8 and Sentinel-2 data for 10-m dense time-series imagery using a degradation-term constrained deep network," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 108, Apr. 2022, Art. no. 102738.

[27] D. Jia, C. Cheng, S. Shen, and L. Ning, "Multitask deep learning framework for spatiotemporal fusion of NDVI," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5616313.

[28] Z. Ao, Y. Sun, X. Pan, and Q. Xin, "Deep learning-based spatiotemporal data fusion using a patch-to-pixel mapping strategy and model comparisons," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5407718.

[29] W. Li, C. Yang, Y. Peng, and X. Zhang, "A multi-cooperative deep convolutional neural network for spatiotemporal satellite image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10174–10188, 2021.

[30] D. Lei, G. Ran, L. Zhang, and W. Li, "A spatiotemporal fusion method based on multiscale feature extraction and spatial channel attention mechanism," *Remote Sens.*, vol. 14, no. 3, p. 461, Jan. 2022.

[31] 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.

[32] Z. Ao, Y. Sun, and Q. Xin, "Constructing 10-m NDVI time series from Landsat 8 and Sentinel 2 images using convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 8, pp. 1461–1465, Aug. 2021.

[33] A. Htitiou, A. Boudhar, and T. Benabdelouahab, "Deep learning-based spatiotemporal fusion approach for producing high-resolution NDVI time-series datasets," *Can. J. Remote Sens.*, vol. 47, no. 2, pp. 182–197, Mar. 2021.

[34] D. Jia, C. Cheng, C. Song, S. Shen, L. Ning, and T. Zhang, "A hybrid deep learning-based spatiotemporal fusion method for combining satellite images with different resolutions," *Remote Sens.*, vol. 13, no. 4, p. 645, Feb. 2021.

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

[36] W. Li, X. Zhang, Y. Peng, and M. Dong, "Spatiotemporal fusion of remote sensing images using a convolutional neural network with attention and multiscale mechanisms," *Int. J. Remote Sens.*, vol. 42, no. 6, pp. 1973–1993, Mar. 2021.

[37] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[38] R. Li, L. Wang, C. Zhang, C. Duan, and S. Zheng, "A^2-FPN for semantic segmentation of fine-resolution remotely sensed images," *Int. J. Remote Sens.*, vol. 43, no. 3, pp. 1131–1155, Feb. 2022.

[39] X. Ying, Q. Wang, X. Li, M. Yu, H. Jiang, J. Gao, Z. Liu, and R. Yu, "Multi-attention object detection model in remote sensing images based on multi-scale," *IEEE Access*, vol. 7, pp. 94508–94519, 2019.

[40] Y. Sun, W. Liu, Y. Gao, X. Hou, and F. Bi, "A dense feature pyramid network for remote sensing object detection," *Appl. Sci.*, vol. 12, no. 10, p. 4997, May 2022.

[41] D. Gaur, J. Folz, and A. Dengel, "Training deep neural networks without batch normalization," 2020, *arXiv:2008.07970*.

[42] H. Wang, A. Zhang, S. Zheng, X. Shi, M. Li, and Z. Wang, "Removing batch normalization boosts adversarial training," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022, pp. 23433–23445.

[43] A. Aitken, C. Ledig, L. Theis, J. Caballero, Z. Wang, and W. Shi, "Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize," 2017, *arXiv:1707.02937*.
[44] Z. Tan, P. Yue, L. Di, and J. Tang, "Deriving high spatiotemporal remote sensing images using deep convolutional network," *Remote Sens.*, vol. 10, no. 7, p. 1066, Jul. 2018.
[45] Y. Xie, W. Wu, H. Yang, N. Wu, and Y. Shen, "Detail information prior net for remote sensing image pansharpening," *Remote Sens.*, vol. 13, no. 14, pp. 1–22, Jul. 2021.
[46] W. Li, F. Wu, and D. Cao, "Dual-branch remote sensing spatiotemporal fusion network based on selection kernel mechanism," *Remote Sens.*, vol. 14, no. 17, p. 4282, Aug. 2022.
[47] 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.
[48] R. Yuhas, A. F. H. Goetz, and J. W. Boardman, "Descrimination among semi-arid landscape endmembers using the Spectral Angle Mapper (SAM) algorithm," in *Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, vol. 1, 1992, pp. 147–149.
[49] Z. Tan, M. Gao, X. Li, and L. Jiang, "A flexible reference-insensitive spatiotemporal fusion model for remote sensing images using conditional generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5601413.
[50] P. Jagalingam and A. V. Hegde, "A review of quality metrics for fused image," *Aquatic Proc.*, vol. 4, pp. 133–142, Mar. 2015.
[51] J. Jiang, Q. Zhang, X. Yao, Y. Tian, Y. Zhu, W. Cao, and T. Cheng, "HISTIF: A new spatiotemporal image fusion method for high-resolution monitoring of crops at the subfield level," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4607–4626, 2020.
[52] V. Chhabra, R. U. Kiran, J. Xiao, P. K. Reddy, and R. Avtar, "A spatiotemporal image fusion method for predicting high-resolution satellite images," in *Advances and Trends in Artificial Intelligence. Theory and Practices in Artificial Intelligence*. Cham, Switzerland: Springer, 2022, pp. 470–481.
[53] Y. Qiu, J. Zhou, J. Chen, and X. Chen, "Spatiotemporal fusion method to simultaneously generate full-length normalized difference vegetation index time series (SSFIT)," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 100, Aug. 2021, Art. no. 102333.
[54] D. Xie, F. Gao, L. Sun, and M. Anderson, "Improving spatial–temporal data fusion by choosing optimal input image pairs," *Remote Sens.*, vol. 10, no. 7, p. 1142, Jul. 2018.

JUAN XIAO received the master's degree in environmental science from Hokkaido University, Sapporo, Japan, in 2020, where she is currently pursuing the Ph.D. degree in environmental science. Her research interests include precision agriculture, deep learning, and remote sensing application.

ASHWANI KUMAR AGGARWAL (Senior Member, IEEE) received the Ph.D. degree in computer vision from the Information and Communication Engineering Department, The University of Tokyo, Japan. He is currently working as a Professor with the Department of Electrical and Instrumentation Engineering, Sant Longowal Institute of Engineering and Technology, India. He has more than 20 years of research and teaching experience in the field of computer vision and image processing.

UDAY KIRAN RAGE (Senior Member, IEEE) received the Ph.D. degree in computer science from the International Institute of Information Technology, Hyderabad, Telangana, India. He is currently an Associate Professor with the Database Systems Laboratory, Division of Information Systems, The University of Aizu, Aizuwakamatsu, Fukushima, Japan. He also holds the visiting researcher position with The University of Tokyo and the National Institute of Information and Communications Technology, Tokyo, Japan. He has published over 80 papers in top-tier conferences, such as ICDE, EDBT, SSDBM, PAKDD, IEEE FUZZY, IEEE BIGDATA, DSAA, DASFAA, and DEXA. He has also published articles in refereed journals, such as IEEE Access, *KBS*, *Information Science*, *JIIS*, and *JSS*. His research interests include big data analytics, artificial intelligence, cloud computation, air pollution analytics, traffic congestion data analytics, recommender systems, and ICTs for agriculture.

VAIBHAV KATIYAR (Member, IEEE) received the master's degree in geoinformatics from the Asian Institute of Technology, Thailand, and the Ph.D. degree from Yamaguchi University, Japan. He is currently working as an Associate Professor at Yamaguchi University. He is also managing the Research and Development Division (RDD) at New Space Intelligence (NSI), Japan. His area of expertise is developing the deep learning-based solution for SAR and optical satellite images with a focus on disaster and agriculture applications.

RAM AVTAR received the master's degree in environmental science from Jawaharlal Nehru University, New Delhi, India, and the Ph.D. degree in civil engineering from The University of Tokyo, Japan. He is currently an Associate Professor with the Faculty of Environmental Earth Science, Hokkaido University, Japan, and the Director of the Global Land Program (GLP), Japan Nodal Office. He is actively contributing to study land systems and co-design solutions for global sustainability as a part of GLP Program. He was with the Institute for the Advanced Study of Sustainability (UNU-IAS), United Nations University, as a Research Fellow for four years (2012–2016). He has developed methods for mapping natural resources using multi-sensor remote sensing techniques and scenario analysis for sustainable management of these resources. He is also working on the synergistic use of remote sensing and unmanned aerial vehicles (UAVs) techniques to monitor the environment more precisely to solve environmental issues from a global to local scale.