Multimodal and Multiresolution Data Fusion for High-Resolution Cloud Removal: A Novel Baseline and Benchmark

Fang Xu, Yilei Shi, *Member, IEEE*, Patrick Ebel, Wen Yang, *Senior Member, IEEE*, and Xiao Xiang Zhu, *Fellow, IEEE*

Abstract— Cloud removal (CR) is a significant and challenging problem in remote sensing, and in recent years, there have been notable advancements in this area. However, two major issues remain hindering the development of CR: the unavailability of high-resolution imagery for existing datasets and the absence of evaluation regarding the semantic meaningfulness of the generated structures. In this article, we introduce M3R-CR, a benchmark dataset for high-resolution CR with multimodal and multiresolution data fusion. M3R-CR is the first public dataset for CR to feature globally sampled high-resolution optical observations, paired with radar measurements and pixel-level land-cover annotations. With this dataset, we consider the problem of CR in high-resolution optical remote-sensing imagery by integrating multimodal and multiresolution information. In this context, we have to take into account the alignment errors caused by the multiresolution nature, along with the more pronounced misalignment issues in high-resolution images due to inherent imaging mechanism differences and other factors. Existing multimodal data fusion-based methods, which assume the image pairs are aligned accurately at the pixel level, are thus not appropriate for this problem. To this end, we design a new baseline named Align-CR to perform the low-resolution synthetic aperture radar (SAR) image-guided high-resolution optical image CR. It gradually warps and fuses the features of the multimodal and multiresolution data during the reconstruction process, effectively mitigating concerns associated with misalignment. In the experiments, we evaluate the performance of CR by analyzing the quality of visually pleasing textures using image reconstruction (IR) metrics and further analyze the generation of semantically meaningful structures using a well-established semantic segmentation task. The proposed Align-CR method is superior to other baseline methods in both areas. The project is available at https://gitlab.lrz.de/ai4eo/M3R-CR.

Index Terms— Cloud removal (CR), data fusion, multimodal, multiresolution.

I. INTRODUCTION

REMOTE-SENSING imagery has been receiving considerable attention as a major promising prospect in various applications such as Earth observation and environmental monitoring [1], [2], [3], [4]. However, haze and clouds in the atmosphere affect the transmission of electromagnetic signals and lead to a deficiency of surface information [5], severely limiting the potential of optical remote-sensing imagery [6]. Cloud removal (CR) aims at reconstructing the cloud-contaminated regions to counteract the degradation caused by clouds and thus is essential for remote-sensing interpretations and applications. To advance the state-of-the-art in CR, many datasets, such as RICE-II [7] and SEN12MS-CR [8], have been proposed in recent years. Most of these datasets are built based on Landsat-8 or Sentinel-2 imagery. However, remote-sensing technology has given rise to the next-generation satellites, which can provide optical imagery with higher spatial resolution [9], [10]. Such data allows geometric analysis on a finer scale while posing new challenges for recovering cloud-covered regions with the corresponding levels of detail. The community is very productive when it comes to proposing novel methods, while *it remains an open question whether and to what extent the CR techniques developed on current datasets with relatively low spatial resolutions can generalize to high-resolution ones*. The unavailability of high-resolution CR datasets severely restricts the development of CR algorithms for recovering clear edges and rich texture details of high-resolution remote-sensing imagery.

To compare different CR algorithms, existing work almost exclusively relies on metrics such as the peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) to evaluate the quality of reconstructed images, which can provide quantitative information about the visual quality of the reconstructed images. However, it is also essential that the reconstructed remote-sensing images are suitable for semantic analysis, in addition to being visually appealing. These visual metrics do not provide a comprehensive evaluation of

Fang Xu is with the School of Electronic Information, Wuhan University, Wuhan 430072, China, and also with the Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: xufang@whu.edu.cn).

Yilei Shi is with the School of Engineering and Design, Technical University of Munich, 80333 Munich, Germany (e-mail: yilei.shi@tum.de).

Patrick Ebel is with the Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: patrick.ebel@tum.de).

Wen Yang is with the School of Electronic Information, Wuhan University, Wuhan 430072, China (e-mail: yangwen@whu.edu.cn).

Xiao Xiang Zhu is with the Data Science in Earth Observation and the Munich Center for Machine Learning, Technical University of Munich, 80333 Munich, Germany (e-mail: xiaoxiang.zhu@tum.de).

Digital Object Identifier 10.1109/TGRS.2023.3337845

the reconstructed image quality, as *it is uncertain whether the resulting visually appealing images perform well in the subsequent semantic tasks*. To identify the usability of the reconstructed images, it is desirable to evaluate them in terms of generating semantically meaningful structures.

In this article, we introduce the *M3R-CR* dataset, illustrated in Fig. 1, which consists of paired cloudy and corresponding cloud-free optical image tiles collected from Planet satellite imagery [11] at a spatial resolution of 3 m. Compared to existing publicly available CR datasets which are mostly built on Landsat-8 data with 30 m resolution or Sentinel-2 data with 10 m resolution, our dataset addresses the current lack of CR datasets with high resolution. Furthermore, our dataset collects corresponding pixel-level land-cover annotations from the WorldCover product [12]. This provides the opportunity to validate the effectiveness of the CR methods in generating globally distributed and semantically meaningful structures through a well-established remote-sensing task.

CR is a highly ill-posed problem due to the loss of surface information. Previous studies have indicated that the ill-posedness can be reduced by resorting to synthetic aperture radar (SAR) data, which is cloud-penetrable and inherently reflects the geometrical characteristics of ground objects [8], [13], [14]. In these studies, SAR data with resolutions comparable to optical data, or downsampled higher-resolution SAR data, is often utilized as a complementary data source for CR. However, compensating for Planet data presents further intricacies, as acquiring high-resolution SAR data with a resolution similar to that of Planet data, such as TerraSAR-X and COSMO-SkyMed satellite data, is typically unaffordable [15], especially on a global scale. Yet, there is a promising alternative: the freely available global Sentinel-1 SAR data from the European Space Agency, albeit of a lower resolution, can be incorporated into our M3R-CR dataset to enhance CR performance. Compared with existing CR datasets like SEN12MS-CR, which mostly explore the fusion of Sentinel-1 SAR data and Sentinel-2 optical data at a consistent resolution of 10 m, our M3R-CR dataset helps to understand a more practical, yet more complex problem: multimodal and multiresolution data fusion-based CR (MMRF-CR), in which the alignment inaccuracies stemming from resolution difference, more pronounced misalignment issues in high-resolution images due to factors like field-of-view mismatch and disparity, have to be accounted for.

To address the problem of MMRF-CR, we propose a novel method called Align-CR, where the low-resolution SAR images guide the reconstruction of high-resolution cloud-free images from cloudy images. Specifically, Align-CR progressively refines the alignment of feature maps from the multimodal and multiresolution data using the deformable convolution (DConv) [16], compensating for their misalignment and thereby enhancing the fusion process. Based on the M3R-CR dataset, we benchmark representative CR methods and analyze their performance in generating visually pleasing textures by image reconstruction (IR) metrics and in generating semantically meaningful structures with a well-established semantic segmentation task. Extensive evaluations demonstrate that the Align-CR method achieves the best performance on

the vast majority of our benchmark tests. In summary, our contributions are as follows.

1) We construct M3R-CR, a dataset collected specially for CR to advance the field. It is the public dataset with the highest spatial resolution to date and comprises a large number of regions all around the world with auxiliary SAR imagery and land-cover information.

2) We benchmark state-of-the-art CR algorithms on the proposed M3R-CR dataset and perform extensive evaluations of the recovered semantic information on a well-established semantic segmentation task.

3) We propose a novel CR algorithm, Align-CR, which can better explore complementary information across the multimodal and multiresolution data to better reconstruct the occluded regions.

The remainder of this article is structured as follows. In Section II, we provide a review of the existing datasets and algorithms for CR. Then, Section III describes the proposed M3R-CR dataset in detail. Our proposed Align-CR method is introduced in Section IV and experimental results are discussed in Section V. Finally, discussions for further work and conclusions are presented in Sections VI and VII, respectively.

II. RELATED WORK

A. Datasets for CR

To promote the progress of deep-learning-based CR, several datasets for CR have been proposed. We provide an overview of publicly available CR datasets, as shown in Table I. RICE-I [7] contains 500 pairs of cloudy and cloud-free images collected on Google Earth by setting whether to display the cloud layer, which only contains filmy, partly transparent clouds. RICE-II [7] contains 450 paired images derived from the Landsat-8 OLI/TIRS dataset, where the acquisition time of cloudy and cloud-free images at the same location is less than 15 days. However, the size of both datasets is relatively small, and the data in them is geographically and topographically homogeneous. STGAN [17] comprises approximately 3 000 images of 256 × 256 pixels, obtained from 945 distinct tiles worldwide, where each tile is a square image from the Sentinel-2 satellite with a size of 10980 × 10980 pixels. The cloud cover for each cloudy image is between 10% and 30%. It excludes images with insufficient visible ground upon manual inspection. Then, SEN12MS-CR [8] collects around 150 000 samples, containing different types of real-life clouds, from 169 nonoverlapping areas of interest (AOI) sampled across all inhabited continents during all meteorological seasons. Each AOI is composed of a pair of orthorectified, geo-referenced Sentinel-2 images, with one image being cloudy and the other being cloud-free. To develop CR models that are robust to extensive cloud coverage conditions, each AOI includes an additional co-registered Sentinel-1 SAR image. The related SEN12MS-CR-TS dataset [18] is structured likewise while featuring multiseason repeated measurements. However, a common limitation of the aforementioned datasets is that they are not stratified by land-cover types, making it impossible to assess the generalizability of CR methods over the individual land-cover types that may be of interest
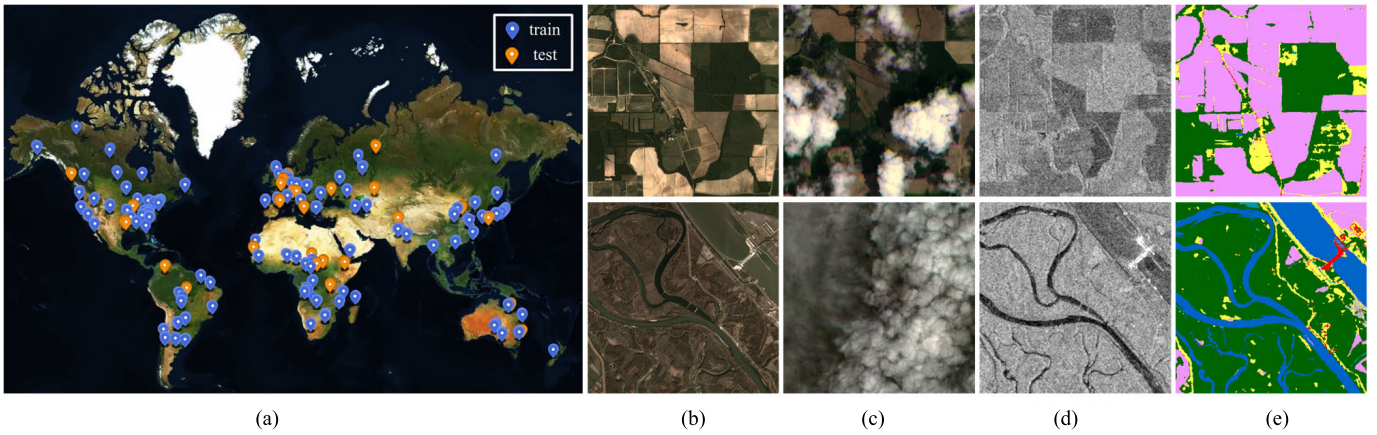
Fig. 1. Visualization of the M3R-CR dataset. (a) Spatial distribution of the globally sampled 780 AOI. (b) Cloud-free optical observations from PlanetScope. (c) Cloudy optical observations from PlanetScope. (d) SAR observations from Sentinel-1 (visualized with the VV polarization mode). (e) Land-cover maps from WorldCover. They are scaled to the same size for better viewing.

TABLE I
COMPARISON BETWEEN PLANET-CR AND EXISTING PUBLICLY AVAILABLE CR DATASETS. *LC* IS SHORT FOR THE LAND-COVER MAP

Dataset	Optical Image					w/ SAR	w/ LC
	source	resolution	#AOIs	width	#images		
RICE-I [7]	Google Earth	$< 15m$	/	512	500	✗	✗
RICE-II [7]	Landsat-8	$30m$	/	512	450	✗	✗
STGAN [17]	Sentinel-2	$10m$	945	256	3,101	✗	✗
SEN12MS-CR [8]	Sentinel-2	$10m$	169	256	122,218	✓	✗
SEN12MS-CR-TS [18]	Sentinel-2	$10m$	53	256	15,578	✓	✗
WHUS2-CR [19]	Sentinel-2	$10m$	36	256	17,182	✗	✗
Scotland&India* [20]	Sentinel-2	$10m$	445	256	445	✓	✗
M3R-CR	PlanetScope	$3m$	780	300	63,000	✓	✓

* The cloudy observations in the dataset are synthesized.

to specific applications like vegetation monitoring [21], [22], [23] and water resources monitoring [24]. To evaluate the effectiveness of CR methods over different land-cover types, WHU2-CR [19] selects 36 locations from all over the world according to three main land covers: urban, vegetation, and bare land and produces about 20 000 pairs of cloudy and cloud-free Sentinel-2 images. Czerkawski et al. [20] constructed a CR dataset, especially in the context of crop monitoring. The dataset contains paired Sentinel-1 and Sentinel-2 Images for two locations in Scotland and India. Notably, the cloudy observations in the dataset are synthesized by overlaying cloud-free images with artificial cloud masks. To further simulate real-world conditions, plausible cloud masks with a coverage area between 10% and 50%, acquired from real images, are also incorporated. Although simulated cloudy observations provide certain benefits, previous research has indicated that models trained on synthesized data poorly generalize to the scenario of real cloud-covered satellite imagery and their spectral characteristics [8].

In addition to the previously mentioned public datasets, there are also several nonpublic datasets. For example, Enomoto et al. [25] created a dataset by combining clouds simulated by Perlin noise with eight comparatively cloudless WorldView-2 images to generate obscured images for learning. Cresson et al. [26] used Sentinel-2 images acquired over the province of Tuy, Burkina Faso to construct experiments. Gao et al. [13] collected two simulated datasets with around 35% cloud cover based on Gaofen-2 optical imagery and airborne optical imagery, respectively, and a real dataset based on Sentinel-2 optical imagery. These datasets are relatively small in size and contain a limited number of scenarios. While featuring high-resolution imagery, cloudy observations are all simulated.

The majority of datasets containing real-life clouds are based on medium-resolution Sentinel-2 data. It is hard to develop and evaluate the removal of clouds in high-resolution imagery with salient structures and abundant textured features. Our M3R-CR dataset aims to advance the task of CR from high-resolution imagery by releasing the cloudy and cloud-free PlanetScope data in combination with Sentinel-1 SAR data and WorldCover land-cover maps. The inclusion of Sentinel-1 SAR data enables the exploration of multimodal fusion for CR, and the inclusion of the WorldCover land-cover product enables disentangling the performance over different land-cover types and evaluating the quality of recovered semantic information. It is worth noticing that Sarukkai et al. [17] train a baseline land classification model and evaluate its accuracy from cloudy, cloud-free, and predicted cloud-free images, respectively, to demonstrate the power of the predicted cloud-free images for downstream

use. However, CR is a pixel-level reconstruction task. Hence, an image-level classification task cannot adequately reflect the quality of predicted cloud-free images, as the correct class can be inferred even if the image is partially occluded [27]. Within our M3R-CR dataset, a pixel-level classification task, that is, semantic segmentation, is provided—which is more suitable for evaluating the power of CR methods on detailed semantic information recovery.

B. Algorithms for CR

CR in optical remote-sensing imagery is a long-standing research problem. Most early developments to addressing this problem involve utilizing spatial correlation or frequency difference between cloudy and noncloudy areas [28], [29], [30]. For example, Zhu et al. [31] utilized a modified neighborhood similar pixel interpolator approach to remove the thick clouds in Landsat images by predicting the value of cloud-contaminated pixels from neighboring similar pixels. Shen et al. [32] executed a homomorphic filter in the frequency domain to remove thin clouds in the Landsat and GaoFen-1 data by suppressing the low-frequency information while enhancing the high-frequency information. Currently, deep-learning-based methods are gaining considerable attention. They have the potential to solve many of the problems that arise in traditional CR methods and achieve impressive results [33], [34], [35]. For example, multispectral conditional generative adversarial networks (McGANs), leveraging the remarkable generative capabilities of conditional generative adversarial networks (cGANs), remove simulated clouds from Worldview-2 imagery by extending the input channels of cGANs to be compatible with multispectral input [25]. Cloud-GAN learns the mapping between cloudy and cloud-free Sentinel-2 imagery using a cyclic consistent generative adversarial network [36]. RSC-Net estimates the cloud-free output with the contaminated Landsat-8 imagery based on an encoding–decoding framework consisting of multiple residual convolutional layers and residual deconvolutional layers [37].

As the clouds thicken and the cloud-covered region is dominant, the challenge of removing the clouds in optical remote-sensing imagery intensifies. In such cases, it is essential to employ auxiliary images as reference data to ascertain the ground surface information obscured by the clouds. A popular alternative method is to utilize multitemporal images to address this issue, that is, estimating the missing information by integrating cloud-free correspondence images acquired at different time [38], [39], [40]. For example, Sarukkai et al. [17] captured the correlations across multitemporal cloudy images over an area by using a spatiotemporal generator network to approximate a cloud-free Sentinel-2 image. Czerkawski et al. [20] used the historical cloud-free optical data as a source of prior information to inpaint the cloud-affected regions in Sentinel-2 images. However, these methods may be less effective in areas with frequent cloud cover and may not apply to time-critical applications. When encountering continual cloudy days [41], cloud-free reference data from an adjacent period are largely unavailable [14], resulting in the same issue of insufficient ground surface information.

Therefore, a series of works explore the potential of using SAR data, which can penetrate cloud cover to acquire ground information beneath the clouds, as auxiliary data for CR in optical imagery [42], [43], [44]. For example, DSen2-CR [45] uses a deep residual neural network to predict the target cloud-free optical image from the concatenation of the Sentinel-1 SAR image and the Sentinel-2 optical image. Simulation-Fusion GAN [13] fuses SAR and corrupted optical imagery with two generative adversarial networks to acquire the cloud-free results of simulated cloudy Gaofen-2 data and real cloudy Sentinel-2 data. GLF-CR [14] incorporates the contribution of Sentinel-1 SAR image in restoring reliable texture details and maintaining global consistency to reconstruct the occluded region of Sentinel-2 optical image. However, these methods are developed on data with relatively low spatial resolutions. The problem of CR in high-resolution imagery remains largely underexplored.

III. DATA

A. Curation of M3R-CR

To develop CR methods that are equally applicable to heterogeneous Earth observation and evaluate the generalizability, it is desirable to curate highly representative locations. To this end, we sample the geospatial locations of 780 nonoverlapping AOIs that are distributed over all continents and meteorological seasons of the globe, as shown in Fig. 1(a). Each AOI is composed of a quartet of orthorectified, geo-referenced cloudy, and cloud-free optical images, as well as the corresponding SAR image and the land-cover map. As it is not possible to capture both cloudy and cloud-free views of a particular location simultaneously, we can only collect the cloud-free observations that are temporally close to the original cloudy observations as the reference. Although certain factors such as sunlight condition, acquisition geometry, humidity, and change of landscape, inevitably introduce nuisances in the data, these nuisances can be considered largely negligible for a relatively large-scale split that is globally and seasonally sampled without any bias to specific conditions. Since SAR is always capable of obtaining surface information regardless of the presence of clouds, we consistently select the SAR image with the time interval closest to that of the cloudy image.

1) Optical Data: High-resolution remote-sensing imagery with extensive spatial–temporal coverage is not easily available. Some satellites offer very high-resolution imagery only from specific locations, which makes it impractical to construct datasets with a high diversity of globally distributed sites. By trading off the spatial–temporal coverage and the resolution of satellite imagery, we collect data from Planet satellite imagery to investigate high-resolution CR techniques. Planet provides global daily data with a spatial resolution of 3 m. On the one hand, it allows the acquisition of paired cloudy and cloud-free images with a very short temporal offset. It could minimize the surface changes that may appear between the acquisition of cloudy and cloud-free images, thereby reducing the nuisances between cloudy reference images and cloud-free target images. On the other hand, it allows the acquisition of heterogeneous Earth observation

data to encourage general-purpose CR but not along narrowly defined and geo-spatially distinct regions of interest.

Specifically, we gather the cloudy and cloud-free observations from the PlanetScope Level-3B top-of-atmosphere reflectance product, which are radiometrically-, sensor-, and geometrically-corrected. It includes four spectral bands, blue, red, green, and near-infrared. In our dataset, the average time interval between paired cloudy and cloud-free data is 2.8 days. This is shorter than the intramission revisiting period of Landsat-8 and Sentinel-2 satellites and minimizes the amount of land-cover change occurring between our paired data. The cloudy data features a diversity of clouds, ranging from semitransparent to dense, and from light to heavy cloud covering. In addition to the cloudy and reference cloud-free data, we also collect the cloud masks associated with cloudy PlanetScope images according to the unusable data mask (UDM) assets provided by Planet Labs. UDM masks give information about which pixels in the images are clear or cloudy, permitting a statistical evaluation of cloud coverage and enabling cloud mask-guided methods to work.

2) SAR Data: Benefiting from the free and open data policy of the Copernicus program, the SAR data of our dataset originates from the Sentinel-1 mission operated by the European Space Agency. We gather the Sentinel-1 data under ascending/descending orbits with a spatial resolution of 10 m from the Level-1 ground range detected (GRD) product archived by Google Earth Engine [46]. The measurements are acquired in interferometric wide swath (IW) mode with two polarization channels vertical transmit/vertical receive (VV) and vertical transmit/horizontal receive (VH), which are pre-processed with thermal noise removal, radiometric calibration, terrain correction, and are converted to backscatter coefficients ($\sigma°$) in units of decibels (dB). In our dataset, the average time interval between the cloudy data and its corresponding SAR data is 2.2 days.

3) Land-Cover Map: The land-cover maps of our dataset originate from WorldCover, which is an open-access global land-cover product at a 10-m resolution released by the European Space Agency. The product achieves an overall accuracy of about 75% across 11 land-cover classes. We access the data through Google Earth Engine. In this study, we employ a simplification process to reduce complexity and noise within our analysis. We consolidate similar land-cover types that might have minor differences within the original comprehensive classification. The data is reclassified into broader categories of 6 basic land-cover types according to DeepGlobe 2018 [47]: forest land, rangeland, agriculture land, urban land, barren land, and water, which allows us to focus on the fundamental characteristics of each land-cover type without being overwhelmed by finer distinctions.

We partition the dataset into training and testing splits to allow for direct comparison with future works. A total of 780 AOIs are split into 660 scenes for training and 120 for testing following a random global distribution. To train the deep-learning-based methods, we crop each AOI into small patches using slide windows of sensor-specific sizes. Since the spatial resolution of PlanetScope imagery is 3 m, and the spatial resolution of Sentinel-1 imagery as well as WorldCover
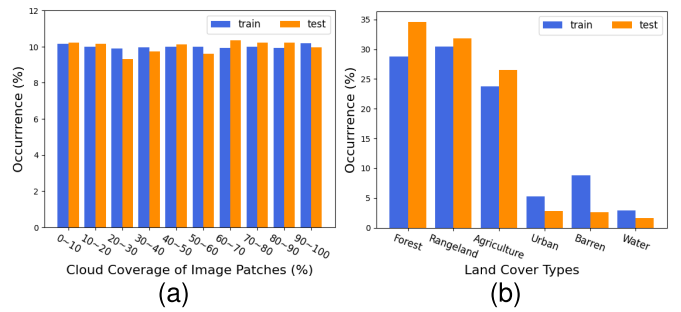

Fig. 2. Statistics of the M3R-CR dataset. (a) Distribution of cloud coverage. (b) Distribution of land-cover types.

land-cover maps is 10 m, we set the corresponding slide window sizes to 300×300 and 90×90 pixels. We manually exclude the samples with inaccurate land-cover annotations and try to distribute the samples uniformly over different cloud coverage levels as well as land-cover types. We finally select 60 000 quartets of training samples and 3 000 quartets of testing samples. The statistics about cloud coverage and land-cover types can be seen in Fig. 2.

B. Properties of M3R-CR

1) Filling High-Resolution Dataset Gap: Unlike existing publicly available CR datasets, which are primarily built on medium-resolution Landsat-8 or Sentinel-2 data, our dataset is constructed using high-resolution PlanetScope data. As a result, it fulfills the requirement of high-resolution CR datasets. It provides the opportunity to push the frontier of current CR models on recovering clear edge and rich texture detail of high-resolution remote-sensing imagery.

2) Integrating Multimodal Multiresolution Data: The inclusion of Sentinel-1 SAR data can provide auxiliary information to promote CR performance. Compared with the existing CR datasets like SEN12MS-CR, which mostly explore the fusion of Sentinel-1 SAR data and Sentinel-2 optical data with the same resolution of 10 m, our M3R-CR dataset helps to understand a more practical yet more complex problem, that is, MMRF-CR.

3) Inclusion of Land-Cover Information: The inclusion of land cover can disentangle the performance of CR methods over different land-cover types, on the one hand, and encourages to design of a pixel-level classification task to evaluate the power of the CR method in generating semantically meaningful structures, on the other hand.

IV. METHODOLOGY

A. Problem Statement

Given a cloudy image $\mathcal{Y}$, the task of CR aims to reconstruct a clear image $\mathcal{F}$ revealing the complete information content of the ground scene so that subsequent analysis can be reliably performed. The basic strategy is to deal with the cloud contamination in a single image without additional information, that is, the problem of single image CR, which can be formulated as follows:

$$\mathcal{F} = \text{CR}(\mathcal{Y}). \tag{1}$$

It is a highly ill-posed problem, usually solved under the assumption that the cloud-contaminated regions have similar spectral/geometrical characteristics to the remaining parts of the image. However, when it comes to the areas with high-frequency textures or different land-cover types, the reconstruction performance cannot be guaranteed. Many studies resort to SAR images that are cloud-penetrable and inherently reflect the geometrical shapes of ground objects as an a priori assumption to reduce the ill-posedness. Thus, the problem of multimodal data fusion-based CR (MMF-CR) is introduced. It restores the clear image from a cloudy image and the corresponding SAR image $\mathcal{S}$, formulated as follows:

$$\mathcal{F} = \text{MMF-CR}(\mathcal{Y}_\Omega, \mathcal{S}_\Omega) \tag{2}$$

where Ω indicates the same spatial domain shared by the cloudy image and the corresponding SAR image. The SAR image guides most MMF-CR methods to restore the cloud-free image from pixel-to-pixel aligned cloudy and SAR images. However, the geo-referenced cloudy images captured by Planet satellites and SAR images captured by Sentinel-1 satellites in the M3R-CR dataset do not meet the assumption of pixel-to-pixel alignment. This lack of alignment primarily stems from two factors. First, the nonalignment issue arises due to the resolution difference between the Sentinel-1 and Planet images. Even when the Sentinel-1 image is upsampled to match the resolution of the Planet image, the process of upsampling introduces uncertainties that contribute to the misalignment. Second, the misalignment issues arising from the inherent differences in imaging mechanisms and other factors are exacerbated when it comes to higher-resolution imagery. In essence, unlike the almost negligible misalignment in the case of Sentinel-1 and Sentinel-2 images with a consistent resolution of 10 m in the SEN12MS-CR dataset, our dataset necessitates addressing the misalignment between the Sentinel-1 and Planet images. The problem of MMRF-CR is thus introduced

$$\mathcal{F} = \text{MMRF-CR}(\mathcal{Y}_{\Omega_y}, \mathcal{S}_{\Omega_s} | \mathcal{M}) \tag{3}$$

where Ω_y and Ω_s, respectively, denote the spatial domain of the cloudy image and the SAR image, and $\mathcal{M} : \Omega_s \rightarrow \Omega_y$ denotes the pixel-level correspondence mapping operator. To solve the problem of MMRF-CR, it is required to precisely align the cloudy and SAR images for the CR process.

B. Align-CR Network

The proposed network, called Align-CR, adopts a two-stream architecture to compensate for the missing information in cloudy regions using ancillary SAR image, as shown in Fig. 3. Rather than directly generating the entire cloud-free image from scratch, Align-CR focuses on learning the residual information, that is, the missing or distorted features caused by the presence of clouds. This is accomplished by incorporating a long additive skip connection for the original cloudy image before the final output, and empirical studies have shown that this approach improves the network's convergence and enhances the quality of the final reconstructed image [48], [49], [50]. Since the SAR image has a lower resolution,

an upsampling operator is first employed to map it to the same resolution as the optical image, facilitating the generation of feature maps with consistent resolutions for subsequent processing. It is worth noting that even though both images are brought to the same resolution, nonstrict alignment issues mentioned earlier, continue to persist. Subsequently, the cloudy optical image and the upsampled SAR image are passed through their respective feature extraction blocks to extract modality-specific features F_{opt}^0 and F_{sar}^0. After that, F_{opt}^0 and F_{sar}^0 are fed into D AlignFuse blocks to obtain knowledgeable features with a holistic understanding of the scene. The AlignFuse block performs alignment and fusion sequentially in the feature space

$$\hat{F}_{\text{sar}}^i = H_{\text{Align}}(F_{\text{opt}}^i, F_{\text{sar}}^i) \tag{4}$$

$$F_{\text{opt}}^{i+1}, F_{\text{sar}}^{i+1} = H_{\text{fuse}}(F_{\text{opt}}^i, \hat{F}_{\text{sar}}^i) \tag{5}$$

where $H_{\text{Align}}(\cdot)$ and $H_{\text{fuse}}(\cdot)$ denote the functions of the alignment block and the fusion block, respectively. The alignment block warps the SAR feature to match the layout of the optical feature, encouraging spatial alignment of corresponding geographic features in both the optical and SAR modalities. The fusion block refines the feature maps from both modalities in an interactive manner. It dynamically transfers complementary information between the aligned feature maps for mutual enhancement. The details of these two blocks are shown in Sections IV-B1 and IV-B2. Finally, all intermediate features $\{F_{\text{opt}}^i\}_{i=1}^D$ are aggregated to reconstruct the high-quality cloud-free image.

1) Alignment Block: Ideally, the task of MMRF-CR can be considered as a two-step task that aligns the multimodal and multiresolution data first and reconstructs the clear image following 2 later. However, explicit pixel-to-pixel alignment to ensure that the same pixels in the SAR image and the cloudy image reflect the same ground target is very hard to achieve in our case. On the one hand, sharp edges in a high-resolution optical image cannot be exactly aligned with blurry edges in a low-resolution SAR image. On the other hand, due to the inherent differences in the imaging mechanisms of optical and SAR sensors, it is very difficult to resolve the unalignment caused by geometric distortions and the like. Moreover, the occlusions in the cloudy images complicate the pixel-to-pixel alignment. Therefore, we implicitly learn the alignment process for cloud-free IR. In this article, we resort to DConv to align the SAR features to the optical features. Given two features to be aligned as input, that is, F_{opt}^i and F_{sar}^i, $i = 0, \ldots, D-1$, the core to solve feature misalignment is to predict the offset with an offset prediction module

$$\triangle p^i = H_{\text{offset}}(F_{\text{opt}}^i, F_{\text{sar}}^i) \tag{6}$$

where H_{offset} denotes the function of the offset prediction module, which can be implemented by general convolutional layers. With the predicted offset, the SAR feature can be warped to the optical feature using the DConv

$$\hat{F}_{\text{sar}}^i = \text{DConv}(F_{\text{sar}}^i, \triangle p^i). \tag{7}$$

Specifically, we adopt the pyramid, cascading, and deformable (PCD) convolutions [51] for the alignment process. It performs
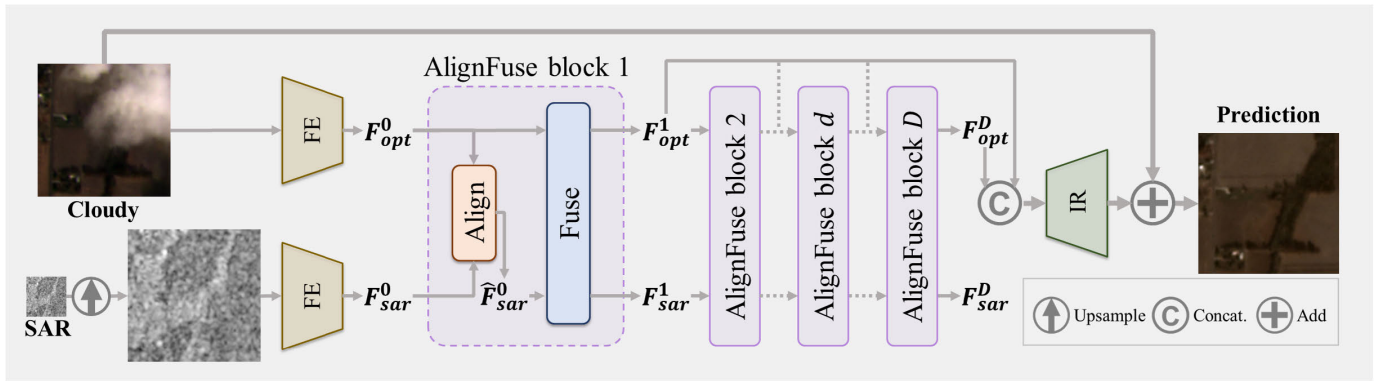
Fig. 3. Overview of the proposed Align-CR method. An upsampling operator is first employed to map the SAR image to the same resolution as the optical image. Then, the cloudy optical image and the upsampled SAR image are passed through their respective feature extraction (FE) blocks to extract modality-specific features. After that, the features are fed into D AlignFuse blocks to obtain knowledgeable features with comprehensive information. The AlignFuse block performs alignment and fusion sequentially in the feature space. Finally, the outputs of all AlignFuse blocks are concatenated and fed to the IR block to restore the high-quality cloud-free image.

the alignment in a pyramid structure, that is, aligning the features in lower scales with coarse estimations first and propagating the aligned features and learned offsets to higher scales to refine the estimations later. Embedding it into the network, the network's ability to model transformations can be enhanced.

2) Fusion Block: The aligned features, that is, F_{opt}^i and $\hat{F}_{\text{sar}}^i$, are then fed into the fusion block for the transfer of complementary information. Similar to our prior work [14], we exploit the power of SAR information from two aspects: global fusion, to guide the global interactions among all local optical windows; local fusion, to transfer the SAR feature corresponding to cloudy areas to compensate for the missing information. Specifically, each fusion block contains an adapted SAR-guided global context interaction (SGCI) block followed by a SAR-based local feature compensation (SLFC) block. To reduce the complexity of the SGCI block, instead of adding a Swin transformer layer (STL) [52] after each convolutional layer in the densely connected layers, we only add an STL after the convolutional layer in the local feature fusion of the residual dense block (RDB) [49] for cross-window feature interaction.

3) Loss Function: We train the Align-CR network by minimizing the difference between the output of network $\mathbf{y}$ and the cloud-free image $\hat{\mathbf{y}}$ temporally close to the input cloudy image. Though the M3R-CR has avoided the surface changes that may appear between the acquisitions of cloudy and cloud-free images as much as possible through a short time lag for data collection, there are some inevitable nuisances as described in Section III-A. In this article, we use the Charbonnier loss [53] that can better handle outliers for training, and extra constraints on cloudy regions are applied

$$\mathcal{L} = (1 + w\mathbf{M}) \odot ((\mathbf{y} - \hat{\mathbf{y}})^2 + \epsilon^2)^\alpha \qquad (8)$$

where $\odot$ denotes the Hadamard product operator, w denotes the extra weight for cloudy regions, $\mathbf{M}$ denotes the cloudy mask, and ϵ and α are constants.

V. EVALUATIONS

A. Experimental Settings

1) Preprocessing: Before the PlanetScope and Sentinel-1 data are fed into the neural networks, we apply value clipping to eliminate a small number of anomalous pixels and data scaling to improve the stability of the neural networks. We clip the values of all bands of the PlanetScope data to [0, 10 000] and divide by 10 000 for all bands. We clip the VV and VH polarizations of the Sentinel-1 data to values $[-25, 0]$ and $[-32.5, 0]$, respectively, and rescale them to the range [0, 1]. All experiments in this article use these preprocessing steps, following previous best practices [8], [45].

2) Implementation Details: The proposed Align-CR network is implemented using PyTorch and trained on 2 NVIDIA Geforce RTX 3090 GPUs with a batch size of 12. During training, we randomly crop the samples into 160×160 patches. The Adam optimizer is used and the maximum epoch of training iterations is set to 30. The learning rate is set to 10^{-4} for the whole network except for the Alignment blocks where the learning rate is set to a smaller value of 10^{-5}. The learning rates decay by 50% every five epochs after the first ten epochs. For the network architecture, the upsampling operator adopts the nearest neighbor interpolation and the number of the AlignFuse blocks D is set to 6. For the loss function, w, ϵ, and α are set to 5, 10^{-3}, and 0.45, respectively.

3) Baselines: In this article, we compare the proposed method with five baseline methods on the M3R-CR dataset with the proposed data splits, including the single image CR methods, McGAN [25] and SpA GAN [54], and the MMF-CR methods, SAR-Opt-cGAN [55], DSen2-CR [45], and GLF-CR [14]. Since existing MMF-CR methods require the input SAR images to be of the same spatial resolution as the input optical images, upsampling the SAR images in the M3R-CR dataset is necessary for these algorithms to work properly. Here, all MMF-CR methods utilize the SAR images upsampled by the nearest neighbor interpolation as input. Additionally, to determine the benefits of including auxiliary low-resolution SAR images, we train the Align-CR network without the use of the SAR images, denoted as

w/o SAR. Moreover, to validate the superiority of Align-CR in integrating multimodal and multiresolution information, we train the Align-CR network by removing the alignment blocks in the AlignFuse blocks, denoted as w/o Align.

B. Evaluation of Visual Recovery Quality

To evaluate the quality of the reconstructed images in terms of generating visually pleasing textures, we start by reporting the mean absolute error (MAE) and PSNR as quantitative measures to assess the reconstruction error. These metrics offer valuable insights into the fidelity of the reconstructed images when compared to their original counterparts. Additionally, we utilize the spectral angle mapper (SAM) and the SSIM to further evaluate the quality of the reconstructed images from the perspectives of spectral and structural similarity, which are more closely aligned with human visual perception. Notably, with the benefit of the land-cover annotations, we disentangle the performance of CR methods over different land-cover types by pixel-wise MAE and SAM metrics, as shown in Table II. The proposed Align-CR method outperforms the selected state-of-the-art CR methods across all evaluated metrics, demonstrating its superiority in addressing the problem of MMRF-CR. It achieves a gain of about 1.68 dB compared to DSen2-CR and a gain of about 0.54 dB compared to GLF-CR in terms of PSNR. The McGAN, SpA GAN, and SAR-Opt-cGAN methods, which were developed on relatively small datasets with geographically and topographically homogeneous data, perform poorly in terms of generalizability on our M3R-CR dataset. Their results have relatively limited spectral fidelity, as suggested by the values of the SAM metric in Table II. It indicates the need to take the point of global distribution into consideration when creating a practical dataset. In Fig. 4, we choose eight scenes to present a qualitative analysis of the reconstructed images. Our Align-CR method can handle various types of clouds, restoring images with more details and fewer artifacts. Moreover, we compare the results over different land-cover types. We can find that removing clouds over urban land that has highly complex geometrical structures is more challenging than others. Urban land in SAR images is characterized by bright-pixel clusters with indistinct boundaries and structures, making it difficult to reconstruct corresponding pixels based on the spectral and texture characteristics of optical images. Besides, we can find that the results over water do not perform well on SAM but relatively well on MAE. It indicates the challenge of maintaining spectral fidelity over water.

To determine what contributes to the superior performance of the proposed method, we analyze the effectiveness of each component by comparing the proposed method and its variants, that is, w/o SAR and w/o Align. As shown in Table II, we can observe that the performance of our methods incorporating SAR images is superior to that of w/o SAR over all types of land cover, highlighting the advantages of utilizing SAR images in CR tasks. Furthermore, as shown in Fig. 4, we can find that w/o SAR tends to generate undesirable artifacts for cloud-covered regions due to the lack of ground information. While Align-CR exploits the geometrical information embedded in the SAR images, which can reconstruct

the ground object. We further compare the CR performance of w/o Align to Align-CR. When the Alignment blocks are removed, the gain of integrating low-resolution SAR information is reduced. In regards to the misalignment, such as the junction of different land-cover types in the first sample and the river border in the fifth sample in Fig. 4, w/o Align tends to generate blurring artifacts, while the reconstructed results of Align-CR have sharper edges. It demonstrates the superiority of our proposed Align-CR method in integrating multimodal and multiresolution information.

We further assess the reconstruction performance on different cloud cover levels, as shown in Fig. 5. The performance of all methods, except McGAN, decreases roughly as the percentage of cloud cover increases. Among them, the performance of the methods with the benefit of SAR images degrades more slowly than the one without, since the utilization of SAR images can alleviate the decline to some extent. Align-CR performs favorably when compared with all baseline methods. When the Alignment blocks are removed, w/o Align behaves very similar to GLF-CR, while GLF-CR contains more Transformer layers for global fusion and thus performs slightly better in terms of SSIM when more prior information from cloud-free regions is available. Our method Align-CR, which aligns the multimodal and multiresolution data during the reconstruction process, can better exploit the power of SAR information. It steadily outperforms w/o Align on all cloud cover levels.

C. Evaluation of Semantic Recovery Quality

In addition to evaluating the quality of reconstructed images in generating visually pleasing textures, we further evaluate their ability to generate semantically meaningful structures, as semantic information is crucial for future analytical applications. In this article, we assess the quality of the recovered semantic information using a well-established land-cover semantic segmentation model. We train a land-cover semantic segmentation model using the cloud-free images and associated land-cover annotations, using the same data splits adopted for the experiments on CR. The model is based on DeepLabv3plus [56] with ResNet-50 [57] as the backbone network. We evaluate the performance of the trained land-cover semantic segmentation model in predicting the correct class of each pixel by using cloud-free images, cloudy images, and predicted cloud-free images from baseline models as inputs, respectively. Ideally, the results with the predicted images as input should be as consistent as possible with the results with the cloud-free images as input, that is, the closer they are to the results with the cloud-free images as input, the better the corresponding CR method performs in terms of semantic recovery. We report the mean intersection over union (mIoU) and pixel accuracy (PA) over varying levels of cloud cover in Table III, where mIoU can better deal with the class imbalance issue. Additionally, we show the discrepancy between the results obtained using cloud-free images as input and the results obtained using predicted cloud-free images or cloudy images as input in Fig. 6.

Clearly, the existence of clouds deteriorates the semantic analysis, and the degradation is more severe when the cloud

TABLE II
QUANTITATIVE RESULTS OF VISUAL RECOVERY QUALITY OVER DIFFERENT LAND-COVER TYPES

(a)

	per class MAE (10^{-2}) ↓						MAE (10^{-2}) ↓	PSNR (dB) ↑
	Forest	Rangeland	Agriculture	Urban	Barren	Water		
McGAN [25]	3.31±0.10	4.14±0.03	4.91±0.06	4.87±0.05	4.98±0.09	4.37±0.05	4.25±0.00	26.06±0.36
SpA GAN [54]	3.63±0.29	3.72±0.09	3.95±0.07	4.40±0.07	4.28±0.05	4.74±0.62	3.83±0.15	26.22±0.27
SAR-Opt-cGAN [55]	3.46±0.46	3.31±0.34	3.49±0.18	3.58±0.05	3.58±0.15	4.61±0.47	3.62±0.39	27.47±0.74
DSen2-CR [45]	2.85±0.05	2.89±0.03	3.10±0.01	3.57±0.04	3.42±0.01	3.53±0.01	2.93±0.04	28.86±0.08
GLF-CR [14]	2.30±0.01	2.56±0.01	2.93±0.00	3.30±0.01	3.22±0.00	2.76±0.02	2.55±0.01	30.00±0.01
Ours (wo/ SAR)	2.51±0.10	2.57±0.02	2.92±0.04	3.23±0.01	3.16±0.01	3.48±0.21	2.66±0.05	29.73±0.04
Ours (wo/ Align)	2.33±0.05	2.53±0.02	2.82±0.00	3.19±0.00	3.10±0.02	2.90±0.21	2.53±0.02	30.04±0.03
Ours (Align-CR)	**2.18±0.01**	**2.42±0.01**	**2.73±0.03**	**3.16±0.05**	**3.06±0.04**	**2.62±0.02**	**2.37±0.01**	**30.54±0.03**

(b)

	per class SAM (°) ↓						SAM (°) ↓	SSIM ↑
	Forest	Rangeland	Agriculture	Urban	Barren	Water		
McGAN [25]	9.91±0.40	11.8±0.28	13.1±0.35	14.0±0.67	13.8±0.68	15.7±0.01	12.3±0.05	0.783±0.015
SpA GAN [54]	8.80±0.27	8.37±0.24	8.19±0.30	9.82±0.39	9.19±0.43	15.4±1.97	8.77±0.35	0.826±0.006
SAR-Opt-cGAN [55]	8.26±0.80	6.94±0.49	6.86±0.08	7.82±0.35	7.00±0.15	10.9±0.23	7.66±0.44	0.839±0.021
DSen2-CR [45]	7.10±0.15	6.55±0.07	6.42±0.08	8.04±0.22	7.42±0.17	9.98±0.22	6.68±0.08	0.882±0.002
GLF-CR [14]	5.76±0.03	5.71±0.00	5.78±0.04	7.85±0.02	7.07±0.02	9.90±0.11	5.65±0.03	0.906±0.000
Ours (wo/ SAR)	6.59±0.44	6.07±0.19	6.24±0.11	7.75±0.08	7.09±0.04	10.2±0.11	6.27±0.27	0.898±0.005
Ours (wo/ Align)	5.93±0.12	5.71±0.03	5.70±0.06	**7.52±0.01**	**6.88±0.03**	9.67±0.02	5.68±0.11	0.905±0.003
Ours (Align-CR)	**5.52±0.04**	**5.46±0.09**	**5.55±0.08**	7.56±0.16	6.92±0.17	**9.60±0.30**	**5.35±0.05**	**0.914±0.001**

TABLE III
PERFORMANCE OF SEMANTIC RECOVERY QUALITY OVER DIFFERENT CLOUD COVER LEVELS

	0~20%		20~40%		40~60%		60~80%		80~100%		Overall	
	mIoU	PA	mIoU	PA	mIoU	PA	mIoU	PA	mIoU	PA	mIoU	PA
Cloud-Free	44.63	62.68	43.66	65.44	41.85	64.40	40.58	64.16	42.98	66.70	43.24	64.66
Cloudy	18.89	34.08	13.26	24.84	10.24	19.27	8.62	16.21	7.89	12.99	11.72	21.46
McGAN [25]	14.24	33.95	14.79	36.89	12.68	32.08	12.30	31.29	10.22	32.48	13.42	33.29
SpA GAN [54]	21.57	39.07	19.85	38.36	16.54	35.74	13.94	33.30	11.46	32.02	16.01	35.66
SAR-Opt-cGAN [55]	12.59	30.70	10.95	29.38	9.96	29.36	8.81	28.56	8.30	26.52	10.19	28.90
DSen2-CR [45]	32.96	43.49	29.75	38.90	25.96	36.99	22.91	36.05	22.91	41.49	27.97	39.40
GLF-CR [14]	34.80	48.32	32.38	44.96	27.27	40.68	23.00	39.33	23.64	44.79	29.34	43.69
Ours (w/o SAR)	33.92	47.76	27.68	41.49	23.94	39.42	19.11	38.11	13.50	41.94	23.96	41.75
Ours (w/o Align)	35.32	48.94	31.70	43.34	28.36	41.75	23.80	40.86	**24.59**	45.87	29.78	44.17
Ours (Align-CR)	**37.41**	**51.80**	**32.66**	**45.32**	**29.18**	**43.73**	**24.92**	**42.89**	23.29	**46.28**	**30.63**	**46.02**

cover level is higher. The benchmarked CR models can, to some extent, counteract the degradation, in which Align-CR generally performs the best. Notably, we can observe that the trained land-cover semantic segmentation model performs better with the predicted images of McGAN and SpA GAN as input than with the predicted images of SAR-Opt-cGAN as input. SAR-Opt-cGAN performs better than McGAN and SpA GAN in terms of all visual quality metrics, as shown in Fig. 5. It indicates that the metrics for measuring the visual recovery quality cannot adequately reflect the performance of CR methods in terms of semantic recovery. We can also find this by comparing the performance of w/o Align and Align-CR. Align-CR steadily outperforms w/o Align on all cloud-cover levels in terms of all visual quality metrics. However, it does not perform as well as w/o Align on the images with cloud cover 80% to 100% in terms of mIoU. What is more, we can find that its performance in terms of mIoU is superior to that of w/o Align when more prior

information from cloud-free regions is available, since there is more information available for alignment.

In Fig. 7, the land-cover mapping results are visualized for comparison. We can find that the land-cover mapping result obtained using a cloud-free image as input shows the highest consistency with the land-cover map from WorldCover. When the image is disturbed by clouds, the corresponding land-cover mapping result is significantly negatively impacted. We apply our Align-CR method to reconstruct the image obscured by clouds, as illustrated in Fig. 7(c). The method effectively removes cloud cover and enhances the overall visual quality of the resulting image. By using the reconstructed image as input for predicting land-cover maps, we can effectively mitigate the performance degradation issue caused by cloud interference and significantly improve the accuracy of the mapping result. It shows the effectiveness of our method in both generating visually pleasing textures and generating semantically meaningful structures.
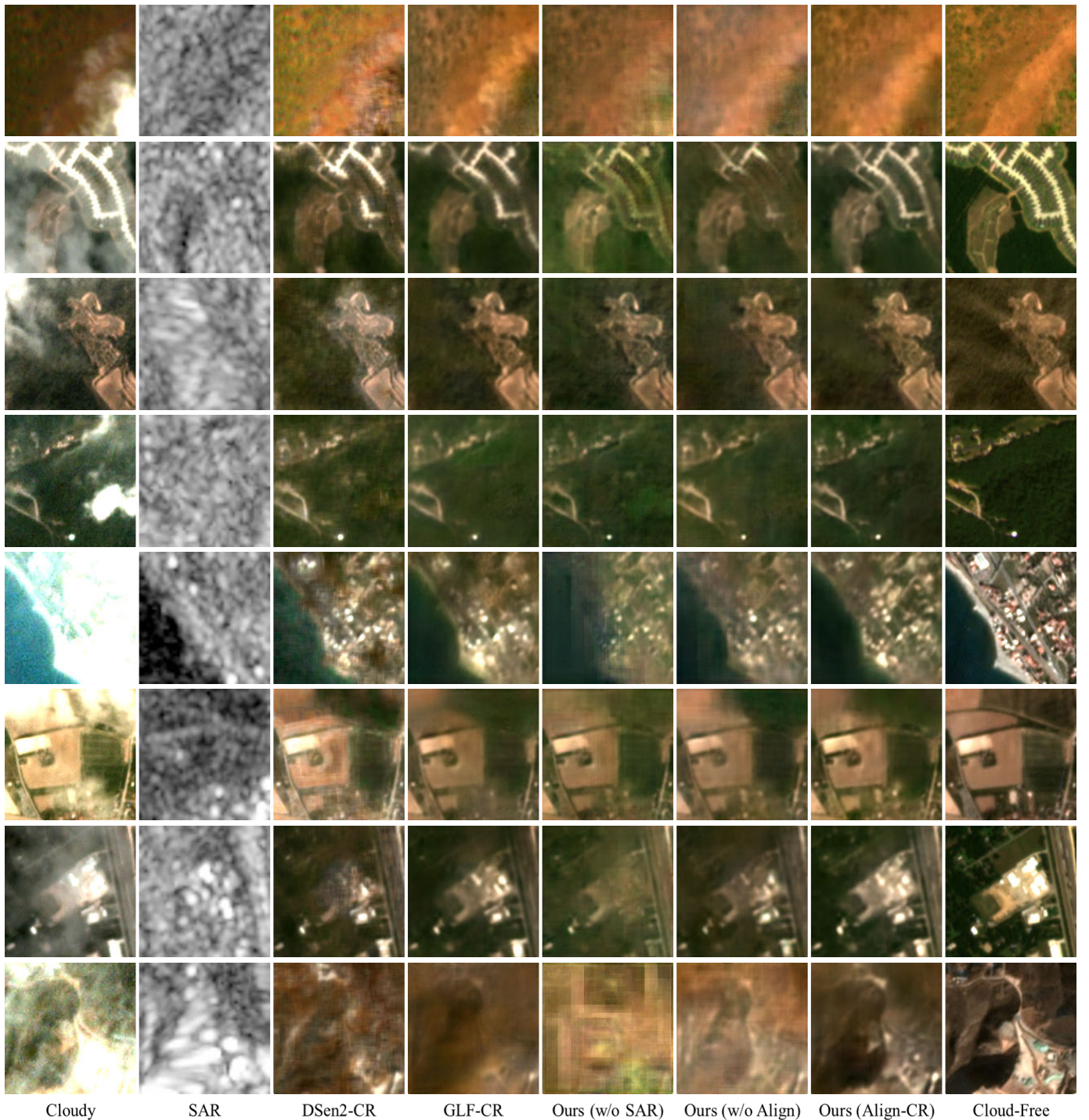
Fig. 4. Qualitative results of visual recovery quality for eight different samples. The first row shows the cloudy images, the second row shows the SAR images, the third to seventh rows show the results from the DSen2-CR, GLF-CR, w/o SAR, w/o Align and Align-CR models, and the eighth row shows the cloud-free images.

VI. DISCUSSION

A. Semantic Considerations in CR Beyond Visual Evaluation

Due to the limited availability of pixel-wise semantic annotations provided along with the CR datasets, most existing studies rely solely on metrics that evaluate the visual similarity between two images when quantifying the effectiveness of CR methods. Toward performance gains in visual metrics, it is common to use loss functions that are constructed based on these metrics to guide the training of CR models, for example, the L1 loss function computes the MAE [14]. It will motivate the predicted cloud-free image to move toward oversmoothness and lead to the potential loss of semantic information. As validated in Section V-C, the current visual metrics cannot adequately evaluate the quality of recovered semantic information. Although our method achieves the best results in terms of semantic recovery, there remains a significant discrepancy between the predicted cloud-free images and
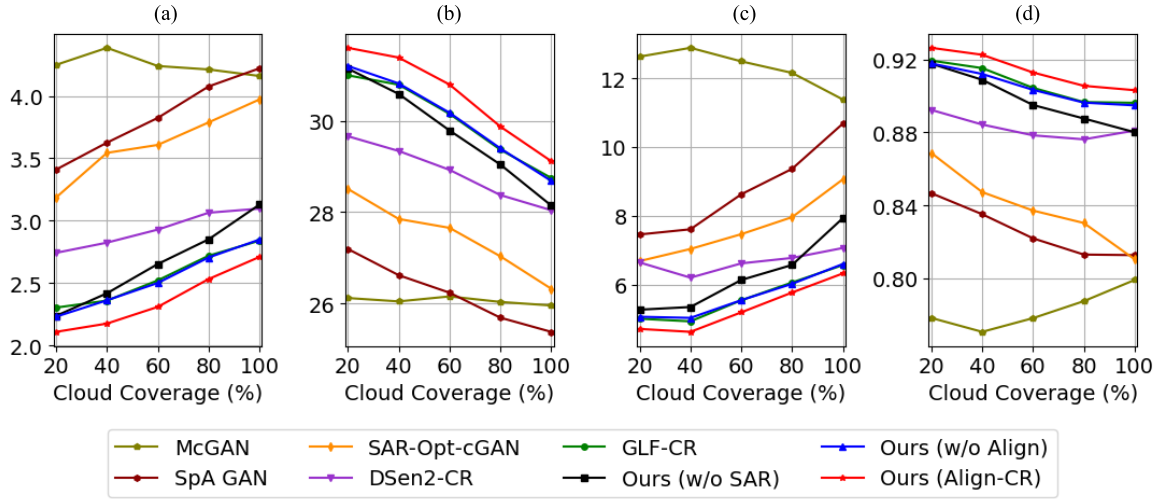
Fig. 5.   Quantitative results of visual recovery quality over different cloud-cover levels in terms of the MAE, PSNR, SAM, and SSIM quality metrics. (a) MAE (10^{-2}) ↓. (b) PSNR (dB) ↑. (c) SAM (°) ↓. (d) SSIM ↑.
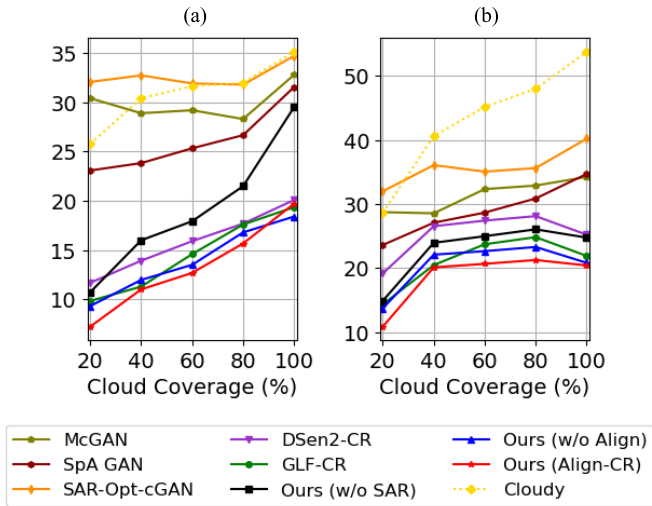

Fig. 6.   Discrepancy between the results obtained using cloud-free images as input and the results obtained using predicted cloud-free images or cloudy images as input. (a) mIoU difference (↑). (b) PA difference (↓).

TABLE IV
QUANTITATIVE COMPARISONS OF THE NDVI VALUES DERIVED FROM CLOUD-FREE IMAGES WITH THOSE FROM RECONSTRUCTED IMAGES

	RMSE ↓	mIoU ↑
McGAN [25]	0.1878	16.20
SpA GAN [54]	0.2417	23.28
SAR-Opt-cGAN [55]	0.1126	22.37
DSen2-CR [45]	0.1015	30.78
GLF-CR [14]	0.0845	32.06
Ours (w/o SAR)	0.1003	28.17
Ours (w/o Align)	0.0844	32.14
Ours (Align-CR)	**0.0802**	**33.77**

their corresponding real cloud-free images. Consequently, it is crucial to develop loss functions that can effectively guide the recovery of semantic information. Future research could focus on the coupling between CR and downstream tasks and leverage this relationship to design a loss function that balances both the semantic context and image details, thus providing more meaningful and practically useful guidance for the reconstruction process.

B. NDVI-Based Evaluation

In addition to the evaluations on visual and semantic recovery quality detailed in Sections V-B and V-C, we also harness the normalized difference vegetation index (NDVI), a widely used remote-sensing index that offers invaluable insights into vegetation land-cover detection and interpretation, to enrich our analysis. The NDVI calculation hinges on the spectral reflectance differences between the near-infrared and red bands, with values spanning from -1 to 1. A higher value correlates with increased vegetation density. Typically, the values can be categorized into five levels: values less than -0.1 depict areas like ground cloud, water, and snow cover; values between -0.1 and 0.1 indicate areas of rock and soil; values between 0.1 and 0.4 signify sparsely vegetated areas; values between 0.4 and 0.8 denote densely vegetated areas; and values greater than or equal to 0.8 characterize super-densely vegetated areas [58]. By comparing the NDVI values extracted from the recovered images to those from the cloud-free images, we are able to evaluate the methods' capability to preserve the distinctive land-cover characteristics within the recovered images. In this study, we utilize the root mean square error (RMSE) to quantify the discrepancies between NDVI values extracted from the recovered images and those from the cloud-free images. Additionally, by taking the vegetable classification results derived from the cloud-free NDVI values as the reference, we employ mIoU to evaluate the precision of the classification based on the declouded NDVI values. The corresponding results are presented in Table IV. Remarkably, the Align-CR method exhibits a higher overall accuracy when compared to other approaches. To offer a more intuitive representation, we employ scatter plots that depict the NDVI values. Specifically, in Fig. 8, we showcase the NDVI values for two distinct samples from Fig. 4: the third sample covering both densely and sparsely vegetated
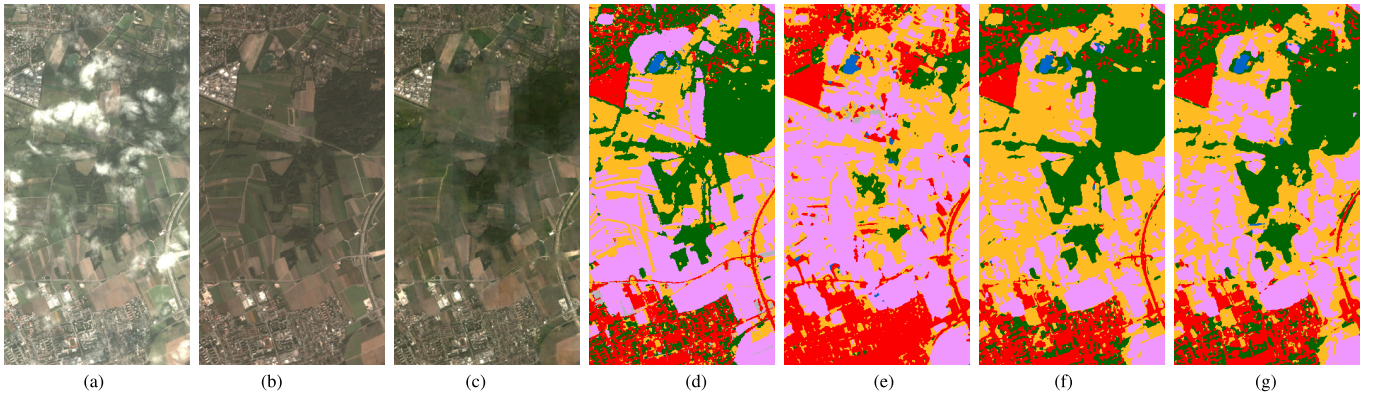
Fig. 7. Visualization of the land-cover mapping results. (a) Cloudy image. (b) Cloud-free image. (c) Reconstruction results of Align-CR. (d) Land-cover map from WorldCover. (e) Land-cover mapping results obtained using a cloudy image as input. (f) Land-cover mapping results obtained using a cloud-free image as input. (g) Land-cover mapping results obtained using the reconstruction result of Align-CR as input. (a) Cloudy. (b) Cloud-Free. (c) Align-CR. (d) *LC*. (e) *LC* from Cloudy. (f) *LC* from Cloud-Free. (g) *LC* from Align-CR.
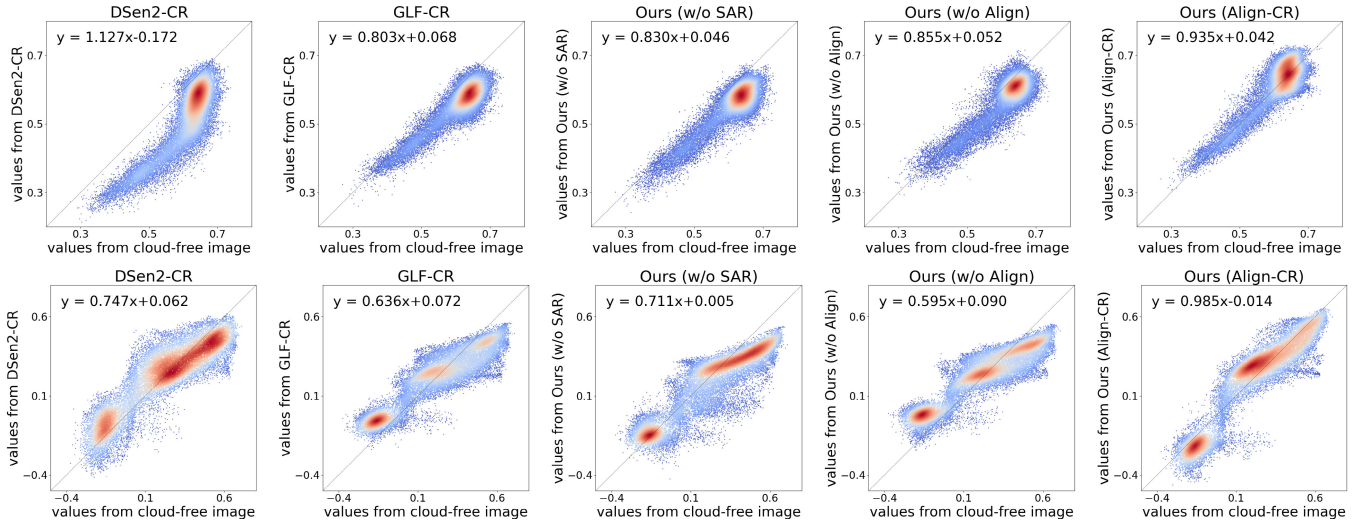

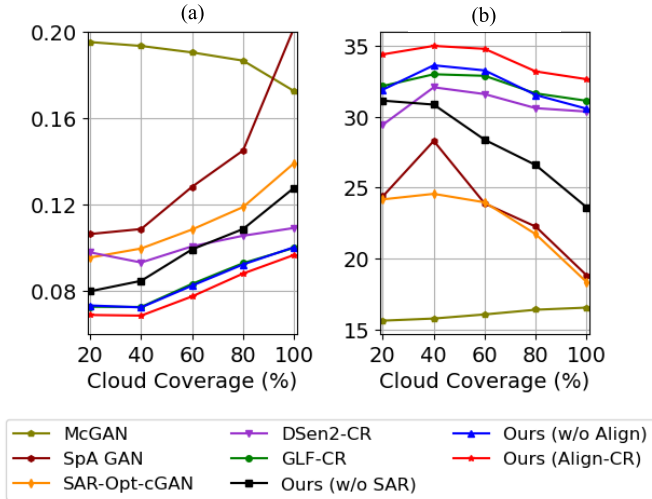Fig. 8. Scatter plots comparing NDVI values derived from cloud-free images with those from reconstructed images.

Fig. 9. Comparison of the NDVI values over different cloud-cover levels. (a) RMSE ($\downarrow$). (b) mIoU ($\uparrow$).

cloud-free images, indicating the superior capability of our method in preserving land-cover characteristics. Furthermore, the points representing water bodies, urban areas, and vegetated zones exhibit heightened concentration, reinforcing the effectiveness of Align-CR in capturing distinct NDVI values across various land-cover types. Delving further, we illustrate the comparison of NDVI values over different cloud cover levels, as shown in Fig. 9. The Align-CR method consistently outperforms competitors across all cloud cover levels in terms of both RMSE and mIoU metrics. It is worth noting that the trend in RMSE correlates with cloud-cover levels in a manner analogous to that observed for SAM. This parallelism arises because the NDVI is computed based on the spectral correlation of the reconstructed images, which closely aligns with the SAM. As for mIoU, it leans more toward a semantic understanding of the scene, offering a perspective distinct from the aforementioned visual quality metrics.

C. Model and Computational Complexity

In Table V, we present a comparison detailing the number of model parameters, floating-point operations (FLOPs), and the inference time when evaluated on an NVIDIA GeForce

regions, and the fifth sample highlighting water bodies and urban terrains. We can find that the NDVI values derived using the Align-CR method more closely align with those from

	#Params	#FLOPs	Inference Time
DSen2-CR [45]	18.91M	483.77G	228.7ms
GLF-CR [14]	14.68M	95.13G	318.9ms
Align-CR	43.53M	174.94G	309.4ms

RTX 3090 GPU with a batch size of 1. Notably, while Align-CR exhibits higher model complexity, it significantly reduces computational complexity compared to DSen2-CR. Furthermore, Align-CR achieves a slightly faster inference speed than that of GLF-CR. This improvement can be attributed to our strategic reduction in the complexity of the SGCI block, as explained in Section IV-B2.

In future research, we will integrate the fusion of multitemporal and multimodal multiresolution data for high-resolution CR, aiming to develop more advanced methods that leverage the complementary information provided by different data sources at different times to enhance the effectiveness of remote-sensing analysis under cloudy conditions.

VII. CONCLUSION

We propose M3R-CR—an open-source multimodal and multiresolution dataset designed specifically for CR in high-resolution optical remote-sensing imagery. With this effort, we aim to encourage the development of innovative CR approaches that leverage the integration of multimodal and multiresolution information. To address the challenge of CR in such datasets, we propose a novel method called Align-CR. This method utilizes an implicit alignment of feature maps during the reconstruction process to compensate for misalignment between multimodal and multiresolution data. Our experimental results demonstrate the effectiveness and superiority of the Align-CR method compared to existing representative CR methods, both in terms of visual recovery quality and semantic recovery quality.

ACKNOWLEDGMENT

REFERENCES

[1] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

[2] Q. Yuan et al., "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, May 2020, Art. no. 111716.

[3] C. Requena-Mesa, V. Benson, M. Reichstein, J. Runge, and J. Denzler, "EarthNet2021: A large-scale dataset and challenge for Earth surface forecasting as a guided video prediction task," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1132–1142.

[4] Z. Xiong, F. Zhang, Y. Wang, Y. Shi, and X. X. Zhu, "EarthNets: Empowering AI in Earth observation," 2022, *arXiv:2210.04936*.

[5] C. Duan and R. Li, "Multi-head linear attention generative adversarial network for thin cloud removal," 2020, *arXiv:2012.10898*.

[6] J. Gawlikowski, P. Ebel, M. Schmitt, and X. X. Zhu, "Explaining the effects of clouds on remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9976–9986, 2022.

[7] D. Lin, G. Xu, X. Wang, Y. Wang, X. Sun, and K. Fu, "A remote sensing image dataset for cloud removal," 2019, *arXiv:1901.00600*.

[8] P. Ebel, A. Meraner, M. Schmitt, and X. X. Zhu, "Multisensor data fusion for cloud removal in global and all-season Sentinel-2 imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5866–5878, Jul. 2021.

[9] L. Kondmann et al., "DENETHOR: The DynamicEarthNET dataset for Harmonized, inter-Operable, analysis-Ready, daily crop monitoring from space," in *Proc. 35th Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2021, pp. 1–13.

[10] J. Cornebise, I. Oršolić, and F. Kalaitzis, "Open high-resolution satellite imagery: The WorldStrat dataset—With application to super-resolution," in *Proc. 36th Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2022, pp. 1–13.

[11] Planet Team PBC. (2018). *Planet Application Program Interface: In Space for Life on Earth*. [Online]. Available: https://api.planet.com

[12] D. Zanaga et al., Oct. 2021, "ESA WorldCover 10 m 2020 v100," *Zenodo*, doi: 10.5281/zenodo.5571936.

[13] J. Gao, Q. Yuan, J. Li, H. Zhang, and X. Su, "Cloud removal with fusion of high resolution optical and SAR images using generative adversarial networks," *Remote Sens.*, vol. 12, no. 1, p. 191, Jan. 2020.

[14] F. Xu et al., "GLF-CR: SAR-enhanced cloud removal with global–local fusion," *ISPRS J. Photogramm. Remote Sens.*, vol. 192, pp. 268–278, Oct. 2022.

[15] Q. Luo, J. Li, and Y. Zhang, "Monitoring subsidence over the planned Jakarta–Bandung (Indonesia) high-speed railway using Sentinel-1 multi-temporal InSAR data," *Remote Sens.*, vol. 14, no. 17, p. 4138, Aug. 2022.

[16] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[17] V. Sarukkai, A. Jain, B. Uzkent, and S. Ermon, "Cloud removal in satellite images using spatiotemporal generative networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1785–1794.

[18] P. Ebel, Y. Xu, M. Schmitt, and X. X. Zhu, "SEN12MS-CR-TS: A remote-sensing data set for multimodal multitemporal cloud removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5222414.

[19] J. Li, Z. Wu, Z. Hu, Z. Li, Y. Wang, and M. Molinier, "Deep learning based thin cloud removal fusing vegetation red edge and short wave infrared spectral information for Sentinel-2A imagery," *Remote Sens.*, vol. 13, no. 1, p. 157, Jan. 2021.

[20] M. Czerkawski et al., "Deep internal learning for inpainting of cloud-affected regions in satellite imagery," *Remote Sens.*, vol. 14, no. 6, p. 1342, Mar. 2022.

[21] J. Rogan, J. Franklin, and D. A. Roberts, "A comparison of methods for monitoring multitemporal vegetation change using thematic mapper imagery," *Remote Sens. Environ.*, vol. 80, no. 1, pp. 143–156, Apr. 2002.

[22] T. A. Lima, R. Beuchle, A. Langner, R. C. Grecchi, V. C. Griess, and F. Achard, "Comparing Sentinel-2 MSI and Landsat 8 OLI imagery for monitoring selective logging in the Brazilian Amazon," *Remote Sens.*, vol. 11, no. 8, p. 961, Apr. 2019.

[23] F. Ngadze, K. S. Mpakairi, B. Kavhu, H. Ndaimani, and M. S. Maremba, "Exploring the utility of Sentinel-2 MSI and Landsat 8 OLI in burned area mapping for a heterogenous savannah landscape," *PLoS ONE*, vol. 15, no. 5, May 2020, Art. no. e0232962.

[24] A. Fisher, N. Flood, and T. Danaher, "Comparing Landsat water index methods for automated water classification in Eastern Australia," *Remote Sens. Environ.*, vol. 175, pp. 167–182, Mar. 2016.

[25] K. Enomoto et al., "Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1533–1541.

[26] R. Cresson, D. Ienco, R. Gaetano, K. Ose, and D. H. T. Minh, "Optical image gap filling using deep convolutional autoencoder from optical and radar images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 218–221.

[27] Z. Gu, P. Ebel, Q. Yuan, M. Schmitt, and X. X. Zhu, "Explicit haze & cloud removal for global land cover classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2022, pp. 1–6.

[28] Z. K. Liu and B. R. Hunt, "A new approach to removing cloud cover from satellite imagery," *Comput. Vis., Graph., Image Process.*, vol. 25, no. 2, pp. 252–256, Feb. 1984.

[29] C. Zhang, W. Li, and D. Travis, "Gaps-fill of SLC-off Landsat ETM+ satellite image using a geostatistical approach," *Int. J. Remote Sens.*, vol. 28, no. 22, pp. 5103–5122, Nov. 2007.

[30] E. H. Helmer and B. Ruefenacht, "Cloud-free satellite image mosaics with regression trees and histogram matching," *Photogrammetric Eng. Remote Sens.*, vol. 71, no. 9, pp. 1079–1089, Sep. 2005.

[31] X. Zhu, F. Gao, D. Liu, and J. Chen, "A modified neighborhood similar pixel interpolator approach for removing thick clouds in Landsat images," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 3, pp. 521–525, May 2012.

[32] H. Shen, H. Li, Y. Qian, L. Zhang, and Q. Yuan, "An effective thin cloud removal procedure for visible remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 96, pp. 224–235, Oct. 2014.

[33] L. Sun, Y. Zhang, X. Chang, Y. Wang, and J. Xu, "Cloud-aware generative network: Removing cloud from optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 691–695, Apr. 2020.

[34] J. Li et al., "Thin cloud removal in optical remote sensing images based on generative adversarial networks and physical model of cloud distortion," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 373–389, Aug. 2020.

[35] L. Liu and S. Hu, "SACTNet: Spatial attention context transformation network for cloud removal," *Wireless Commun. Mobile Comput.*, vol. 2021, pp. 1–8, Oct. 2021.

[36] P. Singh and N. Komodakis, "Cloud-GAN: Cloud removal for Sentinel-2 imagery using a cyclic consistent generative adversarial networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 1772–1775.

[37] W. Li, Y. Li, D. Chen, and J. C.-W. Chan, "Thin cloud removal with residual symmetrical concatenation network," *ISPRS J. Photogramm. Remote Sens.*, vol. 153, pp. 137–150, Jul. 2019.

[38] F. N. Darbaghshahi, M. R. Mohammadi, and M. Soryani, "Cloud removal in remote sensing images using generative adversarial networks and SAR-to-optical image translation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4105309.

[39] A. Sebastianelli et al., "PLFM: Pixel-level merging of intermediate feature maps by disentangling and fusing spatial and temporal data for cloud removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5412216.

[40] Y. Chen, L. Tang, X. Yang, R. Fan, M. Bilal, and Q. Li, "Thick clouds removal from multitemporal ZY-3 satellite images using deep learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 143–153, 2020.

[41] M. D. King, S. Platnick, W. P. Menzel, S. A. Ackerman, and P. A. Hubanks, "Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 7, pp. 3826–3852, Jul. 2013.

[42] Y. Li, F. Wei, Y. Zhang, W. Chen, and J. Ma, "HS^{2}P: Hierarchical spectral and structure-preserving fusion network for multimodal remote sensing image cloud and shadow removal," *Inf. Fusion*, vol. 94, pp. 215–228, Jun. 2023.

[43] S. Han, J. Wang, and S. Zhang, "Former-CR: A transformer-based thick cloud removal method with optical and SAR imagery," *Remote Sens.*, vol. 15, no. 5, p. 1196, Feb. 2023.

[44] W. Li, Y. Li, and J. C. Chan, "Thick cloud removal with optical and SAR imagery via convolutional-mapping-deconvolutional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2865–2879, Apr. 2020.

[45] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 333–346, Aug. 2020.

[46] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sens. Environ.*, vol. 202, pp. 18–27, Dec. 2017.

[47] I. Demir et al., "DeepGlobe 2018: A challenge to parse the Earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 172–17209.

[48] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1132–1140.

[49] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2480–2495, Jul. 2021.

[50] X. Yang, Y. Zhao, and R. R. Vatsavai, "Deep residual network with multi-image attention for imputing under clouds in satellite imagery," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 643–649.

[51] X. Wang, K. C. K. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1954–1963.

[52] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

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

[54] H. Pan, "Cloud removal for remote sensing imagery via spatial attention generative adversarial network," 2020, *arXiv:2009.13015*.

[55] C. Grohnfeldt, M. Schmitt, and X. Zhu, "A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from Sentinel-2 images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 1726–1729.

[56] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

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

[58] Y. Zhang et al., "Two-step ResUp&Down generative adversarial network to reconstruct multispectral image from aerial RGB image," *Comput. Electron. Agricult.*, vol. 192, Jan. 2022, Art. no. 106617.
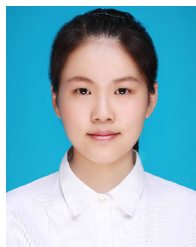
Fang Xu received the B.S. degree in electronic and information engineering from Wuhan University, Wuhan, China, in 2018, where she is currently pursuing the Ph.D. degree in communication and information systems.

Her research involves remote-sensing image processing, including multimodal data matching and fusion.

Yilei Shi (Member, IEEE) received the Dipl.-Ing. degree in mechanical engineering and the Dr.-Ing. degree in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2010 and 2019, respectively.

He is a Senior Scientist with the Chair of Remote Sensing Technology, TUM. His research interests include fast solver and parallel computing for large-scale problems, high-performance computing, and computational intelligence, advanced methods on synthetic aperture radar (SAR) and InSAR processing, machine learning and deep learning for a variety of data sources, such as SAR, optical images, and medical images, and partial differential equation (PDE)-related numerical modeling and computing.

Patrick Ebel received the B.Sc. degree in cognitive science from the University of Osnabrück, Osnabrück, Germany, in 2015, and the M.Sc. degree in cognitive neuroscience and the M.Sc. degree in artificial intelligence from Radboud University Nijmegen, Nijmegen, The Netherlands, in 2018. He is currently pursuing the Ph.D. degree in data science with the Earth Observation Laboratory, Department of Aerospace and Geodesy, Technical University of Munich (TUM), Munich, Germany.

His research interests include machine learning as well as its applications in computer vision and remote sensing. Specifically, he is working on multimodal and multitemporal data fusion and automated IR methods.

Wen Yang (Senior Member, IEEE) received the B.S. degree in electronic apparatus and surveying technology, the M.S. degree in computer application technology, and the Ph.D. degree in communication and information system from Wuhan University, Wuhan, China, in 1998, 2001, and 2004, respectively.

From 2008 to 2009, he worked as a Visiting Scholar with the Apprentissage et Interfaces (AI) Team, Laboratoire Jean Kuntzmann, Grenoble, France. From 2010 to 2013, he worked as a Post-Doctoral Researcher with the State Key Laboratory of Information Engineering, Surveying, Mapping and Remote Sensing, Wuhan University. Since then, he has been a Full Professor with the School of Electronic Information, Wuhan University. His research interests include object detection and recognition, multisensor information fusion, and remote-sensing image analysis.

Dr. Yang received the U.V. Helava Award for the Best Paper in the *ISPRS Journal of Photogrammetry and Remote Sensing* in 2021.

Xiao Xiang Zhu (Fellow, IEEE) received the M.Sc., Dr.-Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She is the Chair Professor of Data Science in Earth Observation at Technical University of Munich (TUM) and was the founding Head of the Department "EO Data Science" at the Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. From 2019 to 2022, she was a co-coordinator of the Munich Data Science Research School (www.mu-ds.de) and the Head of the Helmholtz Artificial Intelligence—Research Field "Aeronautics, Space and Transport." Since May 2020, she has been the PI and the Director of the International Future AI Lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich, Germany. Since October 2020, she has been serving as the Director of the Munich Data Science Institute (MDSI), TUM. She was a Guest Scientist or a Visiting Professor at the Italian National Research Council (CNR-IREA), Naples, Italy; Fudan University, Shanghai, China; the University of Tokyo, Tokyo, Japan; and the University of California at Los Angeles, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. She is currently a Visiting AI Professor at ESA's Phi-Lab. Her main research interests are remote sensing and Earth observation, signal processing, machine learning, and data science, with their applications in tackling societal grand challenges, for example, global urbanization, the UN's SDGs, and climate change.

Dr. Zhu has been a member of the Young Academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She serves on the Scientific Advisory Board in several research organizations, among others the German Research Center for Geosciences (GFZ, from 2020 to 2023) and the Potsdam Institute for Climate Impact Research (PIK). She is an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and *Pattern Recognition* and serves as the Area Editor responsible for special issues of *IEEE Signal Processing Magazine*.