Multi-stream Deep Residual Network for Cloud Imputation Using Multi-resolution Remote Sensing Imagery*

1st Yifan Zhao
Department of Computer Science
North Carolina State University
Raleigh, USA
yzhao48@ncsu.edu

2nd Xian Yang
Department of Computer Science
North Carolina State University
Raleigh, USA
xyang45@ncsu.edu

3rd Ranga Raju Vatsavai
Department of Computer Science
North Carolina State University
Raleigh, USA
rrvatsav@ncsu.edu

***Abstract*—For more than five decades, remote sensing imagery has been providing critical information for many applications such as crop monitoring, disaster assessment, and urban planning. Unfortunately, more than 50% of optical remote sensing images are contaminated by clouds severely affecting the object identification. However, thanks to recent advances in remote sensing instruments and increase in number of operational satellites, we now have petabytes of multi-sensor observations covering the globe. Historically cloud imputation techniques were designed for single sensor images, thus existing benchmarks were mostly limited to single sensor images, which precludes design and validation of cloud imputation techniques on multi-sensor data. In this paper, we introduce a new benchmark data set consisting of images from two widely used and publicly available satellite images, Landsat-8 and Sentinel-2, and a new multi-stream deep residual network (MDRN). This newly introduced benchmark dataset fills an important gap in the existing benchmark datasets, which allows exploitation of multi-resolution spectral information from the cloud-free regions of temporally nearby images, and the MDRN algorithm addresses imputation using the multi-resolution data. Both quantitative and qualitative experiments shows that the utility of our benchmark dataset and as well as efficacy of our MDRN architecture in cloud imputation. The MDRN outperforms the closest competing method by 14.1%.**

***Index Terms*—Remote sensing, Cloud imputation, Benchmark, Multi-resolution, Deep learning**

I. Introduction

Remote sensing imagery became a contributory research material since the 1950's until today. Many areas such as agricultural monitoring, urban planning, earth science, and climate change research rely on remote sensing data. However, cloud cover could be a serious problem for the application of remote sensing imagery in these areas and making the applied machine learning methods performed poorly in tasks such as prediction, segmentation, and recognition. Fortunately, the spatial and temporal density of multi-sensor image collections have significantly increased in recent years thus providing the possibility for improving machine learning performance with cloud imputation.

A large number of both cloud imputation techniques and benchmark datasets have been developed in recent years [1]–[9]. However, most of these benchmarks only include images from a single sensor and train the proposed methods with single-sensor images for cloud imputation. The practical use and performance of single-sensor methods are limited since cloud-free images from a single sensor collection could be temporally distant from each other. Hence, there are significant amount of effect-less and disturbing information in single-sensor data and it could be more difficult and confusing for models to learn the real pattern from cloud-free images. In such cases, multi-sensor imagery becomes an obvious option for improving cloud imputation performance.

However, the diversity of spatial and spectral resolutions presented a challenge when dealing with multiple sensor collections. The diversity of resolutions could make exploiting information hard although diverse image collections provided better opportunities for improving cloud imputation performances. Some recent works tried to utilize multi-sensor imagery for imputing cloud-contaminated areas and correspondingly introduced multi-sensor benchmarks [10]–[16]. But they did not explicitly address the multi-resolution issue that arises in multi-sensor imagery collections. Instead, they only artificially down-sampled the high-resolution images to match lowest resolution images in the collection. As a result, the spatial and spectral information contained in remote sensing images cannot be fully exploited after artificial down-sampling. Thus, the cloud imputation performance could be compromised.

Therefore, in this paper, we introduce a new remote sensing benchmark dataset for the multi-resolution cloud imputation task to fill an important gap in the existing benchmark datasets. The new benchmark dataset consists of Landsat-8 (30m resolution) and Sentinel-2 (10m resolution) images. With this new benchmark dataset, the spectral information in cloudy regions can be inferred with geo-registered temporally nearby multi-resolution cloud-free images from either Landsat-8 or Sentinel-2. The temporal gaps between nearest cloud-free images could be smaller with denser images from

(Preprint): Accepted at ICMLA-2022

multi-sensor collections. However, the ground truth under real clouds cannot be evaluated. Thus, real-world cloud patterns from the EarthNet dataset [6] are superimposed to cloud-free images for simulating real cloudy images and evaluating the cloud imputation performance. Additionally, a novel deep learning based imputation technique is proposed for inferring spectral values under the clouds using nearby multi-resolution imagery. The proposed multi-stream deep residual network (MDRN) exploits multi-resolution spectral information from the cloud-free regions of corresponding temporally nearby images. A multi-stream-fusion structure with two-phase losses is proposed to address the multi-resolution inputs and fuse them for exploiting useful information to restore the cloud-contaminated regions. Besides, a composite upsampling structure is proposed for better incorporating and exploiting the spectral information in low-resolution inputs.

Contributions: Overall, the contributions of this paper are two-fold. First, we developed a new cloud imputation benchmark dataset drawn from two widely used satellite image collections (Landsat-8 and Sentinel-2). This benchmark dataset is tailored for training and testing multi-resolution based cloud imputation algorithms. The benchmark offers sufficient coverage (over three geographically different cities) and variation (size and shape) by introducing a wide variety of cloud masks. Second, a multi-stream deep residual network (MDRN) architecture is proposed for imputing cloud-contaminated regions using the new benchmark dataset. We have conducted extensive experiments and compared MDRN against several state-of-the-art deep learning architectures.

II. Related Work

A. Remote sensing benchmarks

Single-image or single-resolution remote sensing benchmarks have been introduced in several recent works [2], [3], [6]–[9]. [2], [3] considered single-image data for training and evaluating the model performance on the cloud imputation task. The tasks of object detection and scene classification were considered in [8], [9]. However, benchmarks for object detection and scene classification lacks the temporal relationship between images and thus cannot be applied to the cloud imputation task. A multi-image yet single-resolution benchmark named EarthNet was proposed in [6] for forecasting climate impacts. EarthNet collected Sentinel-2 imagery in 10m resolution and tiled it to 128×128 patches for cloud imputation problems. The geo-registered patches were organized as data cubes by temporal order. The time step between each patch is fixed as all data were from a single remote sensing imagery collection. While EarthNet provided stable and well-formulated data series, it is possible that a data cube contains too many cloudy images and any two cloud-free images could be temporally distant from each other. Hence, there are significant amount of effect-less and disturbing information in the data and it could be more difficult and confusing for the model to learn the real pattern from cloud-free images.

In contrast, multi-sensor remote sensing benchmarks were introduced in [11], [15], [16]. Although the multi-sensor image pairs were used for addressing cloud imputation and land cover classification tasks, they did not explicitly address the multi-resolution issue or fully exploit the information contained by multi-sensor data. Instead, the high-resolution data in these benchmarks were artificially down-sampled for processing jointly with the low-resolution data. Such processing could limit the capacity of the information contained in the benchmarks while being applied to the cloud imputation task and could lead to suboptimal performance. Hence in this paper, we introduce a new remote sensing benchmark dataset to fill the gap for the multi-resolution cloud imputation task. The original resolution of data from every sensor collection was preserved in our benchmark for fully exploiting the unique spatial and spectral information contained and improving the cloud imputation performance.

B. Cloud imputation methods

Besides the benchmark, the remote sensing cloud imputation methods has also been primarily considered in single-sensor setting previously in [1]–[4]. [2] employed a GAN architecture with contextual attention mechanism proposed by [17] for restoring cloud-contaminated sea surface temperature images. [3] proposed a spatial-temporal-spectral (STS) convolution network with multi-scale features for predicting missing areas in remote sensing images. Although these works made significant improvements on cloud imputation tasks, single-sensor, single-image setting can only provide limited information and be adopted to limited practical situations compared to multi-sensor, multi-image setting.

Some recent works have considered multi-sensor imagery in the area of remote sensing. Particularly, cloud imputation with multi-sensor data were experimented in [10]–[14], [18]. [10]–[12], [14] used optical and SAR channels for cloud imputation tasks. The SAR sensor could penetrate clouds and always provide high-resolution cloud-free images. Particularly, MSOP$_{unet}$ proposed by [14] employed a similar tri-stream structure to us to encode the optical and SAR images separately. However, the three encoder streams in MSOP$_{unet}$ shared the same weight values whereas our proposed multi-stream structure with three independent streams could exploit multi-sensor inputs more efficiently. Besides, they did not explicitly address the multi-resolution issue between SAR and optical images but artificially down-sampled the SAR images to the lower resolution as optical images instead. There is loss of information caused by artificial down-sampling in their processing compared to the original multi-resolution data.

The multi-resolution issue in remote sensing imagery was considered while addressing other problems such as land cover classification and segmentation in [19]–[26]. A deep learning fusion model, Multi^{3}Net, for multi-resolution remote sensing imagery was proposed in [19] for image segmentation. The contextual information of each resolution of images were extracted with a Pyramid Scene Parsing (PSP) module [27]. While performing our cloud imputation task, Multi^{3}Net could restore the rough and overall contextual information well. However, comparing to the ground truth and our proposed model's predictions, Multi^{3}Net's predictions could be inconsistent to the cloud-free background due to the massive pooling performed in PSP modules.

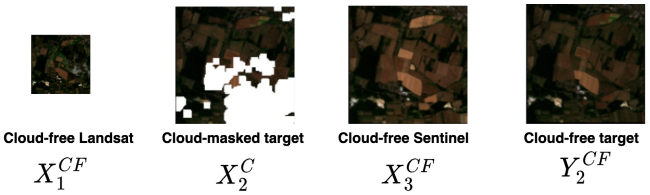


Fig. 1. The illustration of the data setting in multi-resolution cloud imputation task. The cloud-free Landsat-8 image (low resolution) is denoted as X_1^{CF}, the cloud-masked target Sentinel-2 image (high resolution) is denoted as X_2^{C}, the cloud-free Sentinel-2 image (high resolution) is denoted as X_3^{CF}, and the cloud-free target Sentinel-2 image (high resolution) is denoted as Y_2^{CF}.

III. Benchmark

In this section, we introduce the data benchmark we use for training and testing our proposed model and compare it with existing benchmarks that have been used in similar cloud imputation tasks before.

A. Satellite collections

The remote sensing data products we used are from two satellite collections, Landsat-8 and Sentinel-2.

1) Landsat-8:: Equipped with Operational Land Imager (OLI) sensor and Thermal Infrared Sensor[1]. Provides 11 bands, 8 bands at 30m, 1 band at 15m, and 2 bands at 100m spatial resolution. Revisits the same area every 16 days. The Landsat-8 data is from the Level-1 product that can be rescaled to top-of-atmosphere reflectance product. In the cloud imputation task, we use its RGB bands with 30m resolution.

2) Sentinel-2:: Equipped with Multispectral Imager (MSI)[2]. Provides 13 bands, 4 bands at 10m, 6 bands at 20m, 3 bands at 60m spatial resolution. Revisits the same area every 10 days. Currently comprises two identical polar-orbiting satellites in the same orbit, phased at 180 degrees to each other, thus the closest possible time difference between two images is 5 days. The Sentinel-2 data is from the Level-1C top-of-atmosphere reflectance product and has values in the range of [0, 10,000]. In the cloud imputation task, we use its RGB bands with 10m resolution.

B. Multi-resolution benchmark

In this paper, we propose a new benchmark to allow deep learning models to exploit the information in multi-resolution remote sensing imagery collections and achieve better cloud imputation performance.

As noted above in III-A, both widely used satellite collections, Landsat-8 and Sentinel-2, have their independent fixed revisiting frequencies and spatial resolutions. This frequency difference provides us with possibly close remote sensing images for any given spatial extents. Images with smaller temporal gaps could be obtained with multi-sensor collections. For example, the closest possible time difference between two Sentinel-2 images is 5 days. But it is possible that a Landsat-8 image is 2 days away from a Sentinel-2 image since the two satellite collections are independent from each other. In this case, temporally closer cloud-free images could be available from Landsat-8 for any cloudy Sentinel-2 image. In the meantime, more spectrally close information could still be provided by images from the same sensor collection. Therefore, the cloud imputation performance on cloudy Sentinel-2 images could be improved with cloud-free images from both Sentinel-2 and Landsat-8.

More specifically, the cloud imputation benchmark dataset introduced in this paper consists of multi-resolution image triplets from both Landsat-8 and Sentinel-2 with the smallest temporal gap between images similar to [5]. Considering Sentinel-2 has higher resolution, the target image for cloud imputation task is set as Sentinel-2 images. The temporally closest cloud-free Landsat-8 image and another temporally closest cloud-free Sentinel-2 image covering the same spatial extents are searched and extracted as informative inputs with the extended STAC proposed by [28]. Then the remote sensing images covering a large area are tiled into small 384 × 384 pixel patches for Sentinel-2 and 128 × 128 pixel patches for Landsat-8 without overlapping for the ingestion into deep learning models. Furthermore, the same cloud detection method as [11] and Google Earth Engine[3] is employed for cloud coverage filtering. Each patch with less than 10% of cloud coverage is considered as cloud-free and could be used for training and evaluating cloud imputation models. In total, 5003 cloud-free triplets are obtained in our benchmark dataset.

Fig. 1 shows the data setting of our experiments. Let X_1^{CF} (CF stands for cloud-free) denote the cloud-free Landsat-8 patch, Y_2^{CF} denote the target cloud-free Sentinel-2 patch, and X_3^{CF} denote the cloud-free Sentinel-2 patch. For training and evaluating purposes, Y_2^{CF} is artificially cloud-masked by random real cloud masks from the EarthNet dataset [6]. The cloud-masked Y_2^{CF} is denoted as X_2^{C} (C stands for cloudy) and is used as an input patch. Therefore, the multi-resolution cloud imputation task would be to get a mapping such that:

$$G(X_1^{CF}, X_2^{C}, X_3^{CF}, M) :\rightarrow Y_2^{CF} \tag{1}$$

where $G(\cdot)$ is the mapping, that is, the model will be trained and M is the transplanted cloud mask as extra input.

IV. Methodology

Given the data and training framework described in Section III, we introduce our multi-stream deep residual network (MDRN) architecture in this section. MDRN is inspired by the single-resolution cloud imputation network, EDSR [11]. We propose the multi-stream-fusion structure and the composite upsampling structure to address multi-resolution tasks. The newly proposed structures and some other existing components of MDRN are introduced in detail below.

[1] https://landsat.gsfc.nasa.gov/satellites/landsat-8/

[2] https://sentinel.esa.int/web/sentinel/missions/sentinel-2

[3] https://earthengine.google.com/

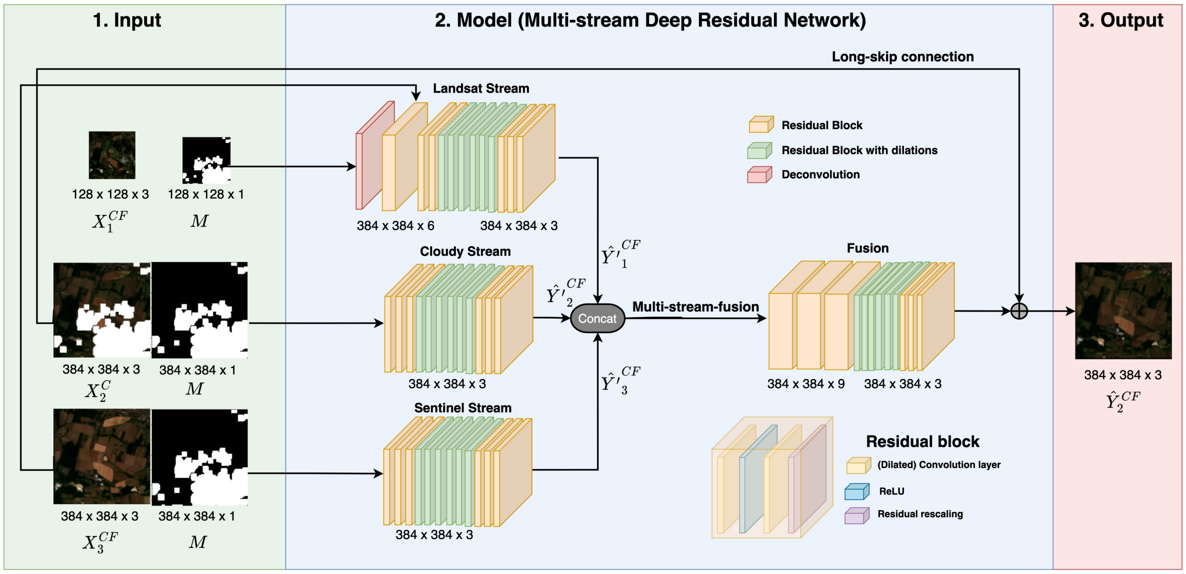


Fig. 2. The overall dataflow and network architecture of our proposed method. The inputs are formed by 3 images, X_1^{CF}, X_2^C, and X_3^{CF}. The cloud mask feature M is incorporated to each image as an extra input feature. The three inputs are first processed with three separate streams consisting of a series of residual blocks. Additionally, the Landsat stream is processed with a composite upsampling structure with the extra information from the Sentinel stream to increase its resolution and dimension to 10m (384×384). After the separate stream processing, the three images are fused to one vector and processed by a series of residual blocks and compressed back to a 3-feature RGB image with a 3-filter dilated residual block. Then the predicted output is formed by the 3-feature RGB image along with the long skipped cloudy input. The bottom-right corner shows the basic residual block that formed the network.

TABLE I
THE DETAILED SPECIFICATIONS OF THE ARCHITECTURE. EXCEPT THE COMPOSITE UPSAMPLING STRUCTURE IN THE LANDSAT STREAM, THAT IS, THE FIRST THREE LAYERS IN LANDSAT STREAM, THE NUMBER OF LAYERS, KERNEL SIZE, AND DILATION RATES OF EACH STREAM ARE ALL EQUAL. THE VARIABLE i INDICATES THE INCREASING RATE OF DILATED CONVOLUTION LAYERS, $i = 1, \ldots, 6$.

Landsat_Stream X_1^{CF} 128*128*3	**Cloudy_Stream** X_2^C 384*384*3	**Sentinel_Stream** X_3^{CF} 384*384*3
Deconvolution layer, kernel = 3×3, output size = 384*384*3		
Concatenating with the cloud-free Sentinel-2 features, output size = 384*384*6		
Convolution layer, kernel = 3×3, output size = 384*384*3		
(Residual block, kernel = 3×3) * 3, output size = 384*384*3	(Residual block, kernel = 3×3) * 3, output size = 384*384*3	(Residual block, kernel = 3×3) * 3, output size = 384*384*3
(Dilated residual block, kernel = 3×3, dilation = 2^i) * 6, output size = 384*384*3	(Dilated residual block, kernel = 3×3, dilation = 2^i) * 6, output size = 384*384*3	(Dilated residual block, kernel = 3×3, dilation = 2^i) * 6, output size = 384*384*3
(Residual block, kernel = 3×3) * 3, output size = 384*384*3	(Residual block, kernel = 3×3) * 3, output size = 384*384*3	(Residual block, kernel = 3×3) * 3, output size = 384*384*3
Concatenation		
(Residual block, kernel = 3×3) * 3, output size = 384*384*9		
(Dilated residual block, kernel = 3×3, dilation = 2^i) * 6, output size = 384*384*3		
(Residual block, kernel = 3×3) * 3, output size = 384*384*3		

Overall architecture: Fig. 2 shows the overall dataflow and network architecture of our proposed method, MDRN. The architecture consists of three panels, 1. input, 2. model (MDRN), and 3. output. The inputs are formed by 3 images, X_1^{CF}, X_2^C, and X_3^{CF}. Each image contains 3 features, R, G, and B. The cloud mask feature M is incorporated to each image as an extra input feature. Then the inputs are transferred to the model. The output of the model, $\hat{Y}_2^{CF}$ is a single RGB image in the same resolution as Sentinel-2 images. Then $\hat{Y}_2^{CF}$ could be quantitatively evaluated and compared with the cloud-free ground-truth Y_2^{CF} to verify the performance of the model. The detailed specifications of the overall architecture are listed in Table I.

Residual block: The basic component in our model is a residual block [11] consisting of four layers: convolution layer, ReLU activation, convolution layer, then a residual rescaling layer [29]. Besides, a shortcut connection over these four layers is deployed to overcome performance degradation possibly caused by the depth of the network [30]. The residual rescaling layer can stabilize the features in the residual block by leveraging the features to an appropriate scale [29].

Dilated convolution: Although the residual blocks with rescaling and shortcut connection can provide stable and well-converged results, the performance of the model is limited by the small receptive fields in traditional convolution layers. Small size of receptive fields can only capture information from limited number of pixels. Whereas large size of receptive fields would be a redundancy and heavy burden in sense of computing. Therefore, we adopt some of the residual blocks in our network with dilated convolution layers [31] in place of the traditional convolution layers. The dilated receptive fields could efficiently capture information from distant pixels to learn the macro pattern well without tremendously increasing the number of parameters in the convolution layer.

Multi-stream-fusion structure: For efficiently exploiting the unique information in each input image, we propose a new multi-stream-fusion structure here. With the multi-stream-fusion structure, the three input images with each concatenated with cloud mask M are processed with three independent streams. Each stream consists of 6 residual blocks and 6 dilated residual blocks. The intermediate output features of each stream are reshaped to the same dimension ($384\times384\times3$) of the ground truth image Y_2^{CF}. Then the mean square error (MSE) loss between each stream output, $\hat{Y'}_1^{CF}$, $\hat{Y'}_2^{CF}$, $\hat{Y'}_3^{CF}$ and the ground truth image Y_2^{CF} is computed for evaluating each stream's preliminary performance and contributions while reconstructing the ground truth image,

$$Loss_{\hat{Y'}_1^{CF}} = MSE(\hat{Y'}_1^{CF}, Y_2^{CF}),$$

$$Loss_{\hat{Y'}_2^{CF}} = MSE(\hat{Y'}_2^{CF}, Y_2^{CF}),$$

$$Loss_{\hat{Y'}_3^{CF}} = MSE(\hat{Y'}_3^{CF}, Y_2^{CF}).$$

After the separate processing by three streams, the three images can be concatenated to one vector as they have the same resolution (10m) and dimension (384×384) now. Then the concatenated vector with 9 features is processed by three residual blocks first. Then the 9-feature vector is compressed back to a 3-feature RGB image with a 3-filter dilated residual block. After that, the 3-feature RGB image is processed with 5 more dilated residual blocks and 3 residual blocks. Additionally, a long skip connection is used over the whole neural network to add the cloudy input X_2^C directly to the final output. This long skip connection could provide a reference for the prediction on the given partly-cloudy image and enable the model to focus on the cloud-masked area [32].

Then the MSE loss $Loss_{Fusion}$ between the fused restored image $\hat{Y}_2^{CF}$ and the ground truth image Y_2^{CF} is computed for evaluating the final performance of the network,

$$Loss_{\hat{Y}_2^{CF}} = MSE(\hat{Y}_2^{CF}, Y_2^{CF}).$$

This way, the network is trained and back-propagated by a tri-stream composite loss,

$$Loss = Loss_{\hat{Y}_2^{CF}} + \lambda(Loss_{\hat{Y'}_1^{CF}} + Loss_{\hat{Y'}_2^{CF}} + Loss_{\hat{Y'}_3^{CF}}) \tag{2}$$

where λ is a hyperparameter controlling the weight of preliminary stream-wise losses.

Composite upsampling structure: As the multi-resolution cloud imputation problem is tackled here, the Landsat-8 image X_1^{CF} with its concatenated cloud mask has lower spatial resolution than the Sentinel-2 features. Thus we propose a lightweight composite upsampling structure in MDRN for upsampling the Landsat-8 features to the same spatial resolution and dimensions as the Sentinel-2 features.

The Landsat-8 stream features are first deconvoluted to 384×384 for the convenience of further processing. Then a copy of the cloud-free Sentinel-2 RGB features X_3^{CF} are concatenated to the Landsat-8 stream as they both have the same dimensions. The concatenated X_3^{CF} are expected to provide spatial information for the Landsat stream while X_1^{CF} could provide spectral information for the later upsampled features. After that, a convolution layer is employed for reshaping the concatenated two-image features to the shape of one-image features. Then the Landsat-8 stream is preliminarily upsampled to the same dimensions as the Sentinel-2 stream with unique information from Landsat-8 represented accurately.

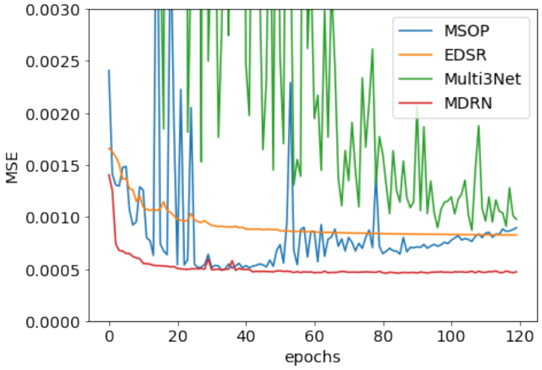


Fig. 3. The validation mean squared error (MSE) curve for our MDRN, and other competing methods: Multi^{3}Net, EDSR, and MSOP$_{unet}$. It shows that our proposed method significantly outperforms the other three state-of-the-art compared here. MSOP$_{unet}$ is the closest model to MDRN but it tends to overfit and has increasing MSE. MDRN also out performs EDSR (see the gap between curves). Multi^{3}Net is oscillating and still not fully converged after 120 epochs.

V. Experimental results and analysis

In this section, we present experiments and comparisons for MDRN and three other closest state-of-art deep learning models, EDSR [11], Multi^{3}Net [19], and MSOP$_{unet}$ [14] on our multi-resolution cloud imputation benchmark dataset. Because EDSR is essentially a single-resolution cloud imputation model, we add one deterministic interpolation layer for the low-resolution Landsat input to let the inputs for EDSR have the same resolution. Multi^{3}Net is a multi-resolution model for remote sensing image segmentation, we adopted it to image restoration task by adding a final layer for adjusting the output dimensions while the inputs stay the same as our proposed model. MSOP$_{unet}$ is the most comparable model experimented here since MSOP$_{unet}$ is a recent multi-sensor cloud imputation model as MDRN. However, MSOP$_{unet}$ did not explicitly address the multi-resolution issue. So we have to add one deterministic interpolation layer for the low-resolution Landsat input to let the inputs have the same resolution.

Environmental settings: All the experiments are performed on a TITAN RTX GPU with 24GB memory. Our training and validations are implemented with the PyTorch framework. The benchmark dataset is split to a training set with 4,003 triplets and a validation set with 1,000 triplets. All the models are trained with batch-size as 16, 120 epochs, mean squared error (MSE) loss, ADAM optimizer, and a step learning rate scheduler starting from 0.01 and every 10 epochs decrease by the rate of 0.75.

A. Quantitative metrics

All the models compared here are trained with a simple MSE loss on the patch. Fig. 3 shows the validation MSE loss curve of each model after each epoch. Our proposed model, MDRN, consistently has the lowest loss curve among all the models compared. MSOP$_{unet}$ is the closest model to MDRN but it tends to overfit and has increasing MSE. EDSR is outperformed by MDRN by an obvious gap even though it converged. Multi^{3}Net is oscillating and still not fully converged after 120 epochs. Additionally, we report other quantitative pixel-wise and structural metrics on the validation set. For pixel-wise metrics, we present MSE of the entire patch and MSE of only the cloudy area for evaluating the overall restoring quality, peak signal-to-noise ratio (PSNR) [33] for an approximation to human perception of the restored image, and the spectral angle mapper (SAM) [34] for showing the spectral (RGB) angle between the target pixel and the restored pixel. For structural metrics, we show structural similarity index (SSIM) [35] for measuring the image reconstruction quality from a visual perception standpoint. Table. II shows the comparison between MDRN and the state-of-art models on the metrics listed above. Specifically for cloudy MSE, the percentage reduction of MDRN compared to each of the state-of-the-art methods is presented. MDRN outperforms the state-of-the-art methods on most of both the pixel-wise and structural metrics only except for SAM. Compared to the most comparable model, MSOP$_{unet}$, MDRN outperforms it on cloudy MSE, the most important metric for the cloud imputation task, by 14.1%.

B. Qualitative verification

Fig. 4 shows a few restored results and their residual maps for qualitatively verifying the performance of the models with various types of cloud coverage from scattered to major blocks. The darker the residual map, the closer the corresponding restored image is to the ground truth. We also showed multiple types of land cover that need imputing: city, suburban farms, grasslands, and mountains. MDRN outperformed the other compared methods on all the land cover types shown in Fig. 4. The promising performance that the proposed model has could lead to various impactful applications such as crop monitoring, building recognition, and wildfire detection.

VI. Ablation Studies

In this section, we show the results of an ablation study to verify the contributions of the newly proposed components in MDRN, multi-stream-fusion structure, the composite upsampling mechanism. Two simplified models, each without one component noted above, are trained and validated on the same settings as the complete proposed model.

Additionally, the influence of our proposed multi-resolution benchmark is also evaluated with an experiment on a simplified dataset. The proposed model is trained and validated on the same dataset only except that the low-resolution cloud-free inputs from Landsat-8 are removed.

No multi-stream-fusion: For testing the contributions of the multi-stream-fusion structure in our proposed model, a simplified model without the multi-stream-fusion structure is experimented. The composite upsampling structure is preserved for a fair comparison. The Landsat-8 features are upsampled first with the extra information from the Sentinel-2 features. Then the three inputs are concatenated as they have the same resolution and dimension. The concatenated inputs are processed with the same depth of residual blocks as the full model. The two-phase stream and fusion loss is also replaced with one single MSE loss at the end of the network since it is part of the multi-stream-fusion structure and the stream losses are dependent to the separate streams.

No composite upsampling: For testing the contributions of the composite upsampling structure in our proposed model, a simplified model with the composite upsampling structure is replaced by a single interpolation layer is experimented. The multi-stream-fusion structure is preserved for a fair comparison. Then the three stream of features are fused as in the complete model.

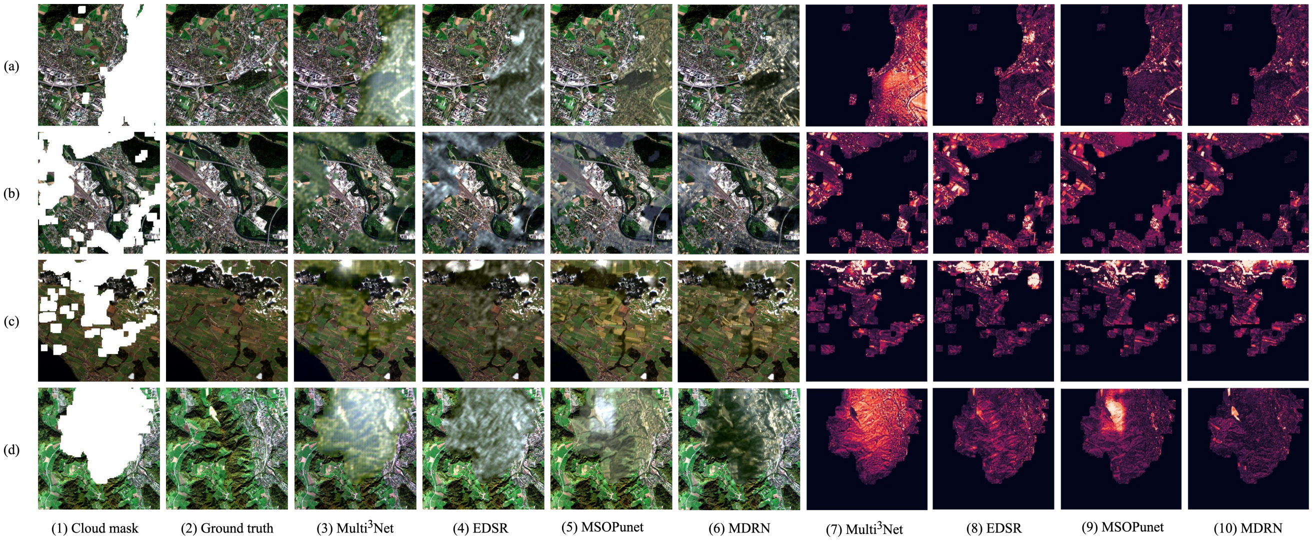
No Landsat-8 image: For evaluating the influence of including a temporally closest low-resolution Landsat-8 image in the cloud imputation problem, a full MDRN model is trained and validated on the same dataset only except that the low-resolution cloud-free inputs from Landsat-8 are removed. The target cloud-masked Sentinel-2 image and neighboring cloud-free Sentinel-2 image stay unchanged as input image pairs for a fair comparison. Our proposed model, MDRN, only the Landsat Stream removed, is trained and validated on the ablate dataset for showing the influence brought by the Landsat-8 input.

Fig. 5 and Table III show the quantitative metrics for the ablation study with our complete proposed method, without multi-stream-fusion structure, without composite upsampling structure, and without Landsat-8 input, respectively. The full MDRN model significantly outperforms the simplified (various components removed) framework. The improved performance suggests that all three components tested here have positive contributions to the full model. The multi-stream-fusion structure is the most contributing component among the three tested here.

TABLE II

The comparison on MSE, cloudy MSE, PSNR, SSIM, and SAM for Multi^{3}Net, EDSR, MSOP$_{unet}$, and MDRN. Specifically for cloudy MSE, the percentage reduction of MDRN compared to each of the state-of-the-art methods is presented. The best result of each metric is bolded.

Methods	MSE (10^{-4})	Cloudy MSE (10^{-4})	Cloudy MSE reduced by (%)	PSNR (dB)	SSIM	SAM (rad 10^{-2})
Multi^{3}Net	8.7208	18.6389	25.7456%	35.4146	0.9585	9.1130
EDSR	8.4067	23.7172	41.6449%	39.0115	0.9796	**5.1581**
MSOP$_{unet}$	4.8696	16.1064	14.0702%	39.3106	0.9816	5.2328
MDRN	**4.5868**	**13.8402**	×	**40.1791**	**0.9826**	5.4678



(1) Cloud mask (2) Ground truth (3) Multi^{3}Net (4) EDSR (5) MSOPunet (6) MDRN (7) Multi^{3}Net (8) EDSR (9) MSOPunet (10) MDRN

Fig. 4. The comparison of restored RGB images and residual maps. The darker the residual map, the closer the corresponding restored image is to the ground truth. From the left to the right, (1) the cloud mask; (2) the ground truth; (3)-(6) the restored images by Multi^{3}Net, EDSR, MSOP$_{unet}$, and MDRN, respectively. (7)-(10) the residual maps of Multi^{3}Net, EDSR, MSOP$_{unet}$, and MDRN, respectively. From top to bottom shows various types of land cover, (a) city; (b) suburban farms; (c) grassland; (d) mountains. We can see obviously from the comparison that MDRN yields the closest restored images to the ground truths and the darkest residual maps.

TABLE III

The comparison on MSE, PSNR, SSIM, and SAM for MDRN, without the multi-stream-fusion structure, without the composite up-sampling structure, and without the low-resolution Landsat-8 input. Specifically for cloudy MSE, the percentage reduction of MDRN compared to each simplified model is presented. The best result of each metric is bolded.

Methods	MSE (10^{-4})	Cloudy MSE	Cloudy MSE reduced by (%)	PSNR (dB)	SSIM	SAM (rad 10^{-2})
No multi-stream-fusion	10.1125	30.5855	54.7491%	34.9582	0.9762	5.0239
No composite upsampling	4.9981	15.8522	12.6922%	37.3975	0.9817	**4.8934**
No Landsat-8	6.6171	19.1968	27.9036%	39.9644	0.9819	5.0791
Complete	**4.5868**	**13.8402**	×	**40.1791**	**0.9826**	5.4678

VII. Conclusion

In this paper, we introduced a new multi-resolution satellite image benchmark dataset for cloud imputation. This new benchmark dataset facilitates the development and thorough validation of cloud imputation algorithms. We also proposed the multi-stream deep residual network (MDRN) that exploits multi-resolution remote sensing images for cloud imputation. MDRN addressed the multi-resolution issue with two newly-proposed components: a multi-stream-fusion structure and a composite upsampling structure. Our experiments showed that MDRN outperforms other state-of-the-art methods using both quantitative and qualitative measures. Our full MDRN model has achieved an improvement of 14.1 to 41.6% cloudy MSE as compared to various state-of-the-art methods. Currently, our method addresses only the multi-resolution issue in multi-sensor cloud imputation problem. In the future, we will work on addressing the generic spectral heterogeneity among multi-sensor images.

VIII. Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Contract #2021-21040700001. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce

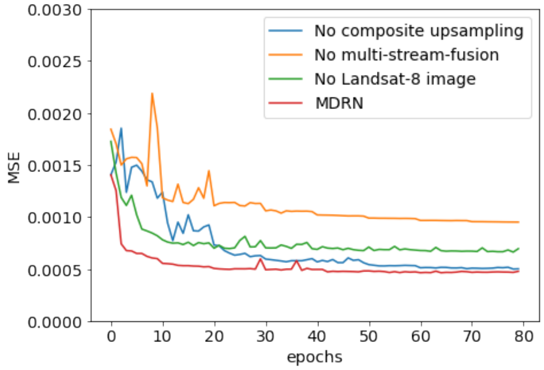


Fig. 5. The validation mean squared error (MSE) curve for the ablation study for MDRN, without the multi-stream-fusion structure, without the composite up-sampling structure, and without the low-resolution Landsat-8 input, respectively after each epoch.

and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] A. Stock, A. Subramaniam, G. L. Van Dijken, L. M. Wedding, K. R. Arrigo, M. M. Mills, M. A. Cameron, and F. Micheli, "Comparison of cloud-filling algorithms for marine satellite data," *Remote Sensing*, vol. 12, no. 20, p. 3313, 2020.

[2] S.-H. Kang, Y. Choi, and J. Y. Choi, "Restoration of missing patterns on satellite infrared sea surface temperature images due to cloud coverage using deep generative inpainting network," *Journal of Marine Science and Engineering*, vol. 9, no. 3, p. 310, 2021.

[3] Q. Zhang, Q. Yuan, C. Zeng, X. Li, and Y. Wei, "Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 8, pp. 4274–4288, 2018.

[4] D. J. Weiss, P. M. Atkinson, S. Bhatt, B. Mappin, S. I. Hay, and P. W. Gething, "An effective approach for gap-filling continental scale remotely sensed time-series," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 98, pp. 106–118, 2014.

[5] X. Yang, Y. Zhao, and R. R. Vatsavai, "Deep residual network with multi-image attention for imputing under clouds in satellite imagery," in *2022 27th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022.

[6] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021.

[7] S. Verma, A. Panigrahi, and S. Gupta, "Qfabric: multi-task change detection dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1052–1061.

[8] X. Sun, P. Wang, Z. Yan, F. Xu, R. Wang, W. Diao, J. Chen, J. Li, Y. Feng, T. Xu *et al.*, "Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 184, pp. 116–130, 2022.

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

[10] 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 Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5866–5878, 2020.

[11] 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 Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 333–346, 2020.

[12] 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 *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 218–221.

[13] A. N. Srivastava, N. C. Oza, and J. Stroeve, "Virtual sensors: Using data mining techniques to efficiently estimate remote sensing spectra," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 590–600, 2005.

[14] R. Cresson, N. Narçon, R. Gaetano, A. Dupuis, Y. Tanguy, S. May, and B. Commandre, "Comparison of convolutional neural networks for cloudy optical images reconstruction from single or multitemporal joint sar and optical images," *arXiv preprint arXiv:2204.00424*, 2022.

[15] G. Sumbul, A. De Wall, T. Kreuziger, F. Marcelino, H. Costa, P. Benevides, M. Caetano, B. Demir, and V. Markl, "Bigearthnet-mm: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 3, pp. 174–180, 2021.

[16] D. Hong, J. Hu, J. Yao, J. Chanussot, and X. X. Zhu, "Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 178, pp. 68–80, 2021.

[17] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5505–5514.

[18] D. P. Roy, J. Ju, P. Lewis, C. Schaaf, F. Gao, M. Hansen, and E. Lindquist, "Multi-temporal modis–landsat data fusion for relative radiometric normalization, gap filling, and prediction of landsat data," *Remote Sensing of Environment*, vol. 112, no. 6, pp. 3112–3130, 2008.

[19] T. G. Rudner, M. Rußwurm, J. Fil, R. Pelich, B. Bischke, V. Kopačková, and P. Biliński, "Multi3net: segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 702–709.

[20] W. Ma, J. Shen, H. Zhu, J. Zhang, J. Zhao, B. Hou, and L. Jiao, "A novel adaptive hybrid fusion network for multiresolution remote sensing images classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2021.

[21] J. Qu, Y. Shi, W. Xie, Y. Li, X. Wu, and Q. Du, "Mssl: Hyperspectral and panchromatic images fusion via multiresolution spatial–spectral feature learning networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[22] D. Varshney, C. Persello, P. K. Gupta, and B. R. Nikam, "Multiresolution fully convolutional networks to detect clouds and snow through optical satellite images," *arXiv preprint arXiv:2201.02350*, 2022.

[23] Z. Sun, W. Zhou, C. Ding, and M. Xia, "Multi-resolution transformer network for building and road segmentation of remote sensing image," *ISPRS International Journal of Geo-Information*, vol. 11, no. 3, p. 165, 2022.

[24] L. Wang, L. Weng, M. Xia, J. Liu, and H. Lin, "Multi-resolution supervision network with an adaptive weighted loss for desert segmentation," *Remote Sensing*, vol. 13, no. 11, p. 2054, 2021.

[25] L. Wang, C. Zhang, R. Li, C. Duan, X. Meng, and P. M. Atkinson, "Scale-aware neural network for semantic segmentation of multi-resolution remote sensing images," *Remote Sensing*, vol. 13, no. 24, p. 5015, 2021.

[26] H. Zhu, W. Ma, L. Li, L. Jiao, S. Yang, and B. Hou, "A dual–branch attention fusion deep network for multiresolution remote–sensing image classification," *Information Fusion*, vol. 58, pp. 116–131, 2020.

[27] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[28] Y. Zhao, X. Yang, and R. R. Vatsavai, "A scalable system for searching large-scale multi-sensor remote sensing image collections," in *The 6th IEEE International Workshop on Big Spatial Data (BSD 2021)*. IEEE, 2021, pp. 3780–3783.

[29] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[31] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[32] L. Yu, X. Yang, H. Chen, J. Qin, and P. A. Heng, "Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[33] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 2366–2369.

[34] F. A. Kruse, A. Lefkoff, J. Boardman, K. Heidebrecht, A. Shapiro, P. Barloon, and A. Goetz, "The spectral image processing system (sips)—interactive visualization and analysis of imaging spectrometer data," *Remote sensing of environment*, vol. 44, no. 2-3, pp. 145–163, 1993.

[35] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.