Research Article

Multisensor Remote Sensing Imagery Super-Resolution with Conditional GAN

Junwei Wang¹, Kun Gao¹, Zhenzhou Zhang¹, Chong Ni², Zibo Hu¹, Dayu Chen², and Qiong Wu¹

¹Key Laboratory of Photoelectronic Imaging Technology and System, Ministry of Education, Beijing Institute of Technology, Beijing 100081, China

²Institute of Spacecraft System Engineering, CAST, Beijing 100094, China

Correspondence should be addressed to Kun Gao; gaokun@bit.edu.cn

Received 9 April 2021; Accepted 10 August 2021; Published 8 September 2021

Copyright © 2021 Junwei Wang et al. Exclusive Licensee Aerospace Information Research Institute, Chinese Academy of Sciences. Distributed under a Creative Commons Attribution License (CC BY 4.0).

Despite the promising performance on benchmark datasets that deep convolutional neural networks have exhibited in single image super-resolution (SISR), there are two underlying limitations to existing methods. First, current supervised learning-based SISR methods for remote sensing satellite imagery do not use paired real sensor data, instead operating on simulated high-resolution (HR) and low-resolution (LR) image-pairs (typically HR images with their bicubic-degraded LR counterparts), which often yield poor performance on real-world LR images. Second, SISR is an ill-posed problem, and the super-resolved image from discriminatively trained networks with l^p norm loss is an average of the infinite possible HR images, thus, always has low perceptual quality. Though this issue can be mitigated by generative adversarial network (GAN), it is still hard to search in the whole solution-space and find the best solution. In this paper, we focus on real-world application and introduce a new multisensor dataset for real-world remote sensing satellite imagery super-resolution. In addition, we propose a novel conditional GAN scheme for SISR task which can further reduce the solution-space. Therefore, the super-resolved images have not only high fidelity, but high perceptual quality as well. Extensive experiments demonstrate that networks trained on the introduced dataset can obtain better performances than those trained on simulated data. Additionally, the proposed conditional GAN scheme can achieve better perceptual quality while obtaining comparable fidelity over the state-of-the-art methods.

1. Introduction

In recent years, remote sensing satellite imagery has been widely used in various fields [1–4]. It is an important data source for people to understand the earth and has a wide range of applications, which has attracted substantial research attention. Among its applications, high-resolution remote sensing data is in great demand [5, 6]. Due to the undersampling effect caused by the size and alignment density of charge-coupled device (CCD), sensor will confuse high-frequency information with the low-frequency information of the image [7]. However, enhancing resolution via the improvement of imaging hardware is challenging because of the cost and craftsmanship limitations. Therefore, it is more attractive to make software-based image super-resolution (SR) in practice [8].

The goal of SR is to exceed the limit of sensor to improve the resolution of the image, which means increasing the number of pixels of the image and provide better spatial details than the original image obtained by the sensor [9, 10]. The methods of SR can generally be divided into two categories: single-image super-resolution (SISR) [11] and multi-image super-resolution (MISR) [12]. MISR usually requires a set of low-resolution (LR) images with subpixel misalignment to reconstruct the high-resolution (HR) image [13–15]. Practically, this approach is generally computation-consuming and numerically limited only to small increases in resolution (by factors smaller than 2) [16, 17]. These limitations lead to the development of SISR, which can recover a HR image from only one of its LR counterparts. In the field of remote sensing, SISR is usually adopted, because it can super-resolve satellite images without the need for multiple

images from the same scene while avoiding the step of accurate registration or the need for a satellite constellation [18, 19].

SISR is an ill-posed inverse problem. Because the LR image does not have complete image information of the ground truth (or, HR image), therefore, instead of a unique solution, there exist multiple corresponding SR solutions to the same LR image. This fact can be mitigated by adding reliable prior information as a regularization to constrain the solution space [11]. Over the past decades, numerous SISR methods have been developed. The pioneer works on SISR focus on heuristic algorithms and hand-crafted features. They devote in SISR and leverage on interpolation techniques based on spatial structure information, for example, multisurface fitting [11] and wavelet transformation [20]. However, the result is not satisfactory due to the complexity of SISR problems. The learning-based SISR methods are then proposed to improve high-frequency details of image patches, such as sparse coding [21], neighborhood regression methods [22], and mapping functions [23]. However, when the parameters of model are complex, the SR performance deteriorates significantly.

In recent years, with the development of deep learning (DL) in the field of computer vision [24], many deep neural networks (DNNs) have been proposed for SR task, which have outperformed traditional SR methods both in metrics and visual quality. Among them, Dong et al. [25] first introduced convolutional neural network (CNN) in SR task, which is trying to learn a nonlinear mapping from LR to HR in an end-to-end manner. The training data are paired images, i.e., paired LR-HR images where LR images are typically bicubic-degraded versions from the HR images. By adopting deeper networks and well-designed network architectures, a multitude of top-performing methods was proposed and achieved the state-of-the-art performance [26–35], which is the current main stream research in SISR. However, despite their promising performance on benchmark datasets, there are two underlying limitations to existing SR methods.

First, because it is difficult to obtain paired high- and low-resolution images in the real world for training, existing methods are usually discriminatively trained on simulated datasets, i.e., HR images with their bicubic-degraded LR counterparts. Consequently, the trained SR networks have poor generalization capacity and often yield poor performance when directly used to super-resolved real-world data, for example, remote sensing satellite imagery. Therefore, it is necessary to consider the specific degradation from HR images to LR images according to the application scenario, so as to train a more robust network with better generalization capacity.

Second, typically, the purpose of SR is to restore the high-frequency information and details lost in the LR image to achieve high fidelity and good visual perceptual quality. Fidelity measures the degree of distortion between super-resolved images and HR images, which matches on the pixel-wise aspect, while the perceptual quality metrics describe the difference in their distribution. Note that high perceptual quality does not need necessarily meet the high fidelity. Although most research still aims at fidelity, perceptual super-resolution research began to flourish. Their

results are less blurry and more realistic [36]. Theoretically, it can be proved that a super-resolved image from a network trained discriminatively with l^p norm loss is the average of the solution-space corresponding to the LR image [37]. Searching in the whole solution-space is not a trivial work. Even worse, the super-resolved images are usually overly blurry and smooth, which lack high-frequency details and textures in high variance regions. Ledig et al. [38] introduced generative adversarial network (GAN) for SR to drive the reconstruction towards the natural image manifold and producing perceptually more convincing solutions. However, although the generated images are more realistic, it usually results in a decrease in the fidelity of super-resolved images. Therefore, how to further reduce the solution space and obtain super-resolved images with high perceptual quality while keeping high fidelity is still a problem.

In this paper, we propose a novel conditional GAN scheme to super-resolved multisensor remote sensing satellite imagery. There are many works which apply GANs in the conditional setting, such as discrete labels [39], text [40], and medical images [41]. With the global availability of Multispectral Instrument (MSI, on-board Sentinel-2 satellite) data and Operational Land Imager (OLI, on-board Landsat-8 satellite) data which have 10 m and 30 m ground sample distance (GSD), respectively, it is conceivable that we can construct a dataset of LR-HR image pairs with OLI data as the LR image and MSI data as the HR image. This dataset, termed OLI2MSI, is a collection of paired real-world multisensor LR-HR data by carefully selecting relatively cloud-free MSI and OLI image in the same location obtained within a suitable temporal window. Experiments demonstrate that networks trained on this dataset can have better performance than that trained on simulated dataset (e.g., bicubic down-sampling), which properly addresses the first limitation. To address the second limitation, we introduce conditional GAN to further reduce the solution space. Experiments show the superiority of our method on fidelity and perceptual quality of the super-resolved images.

Our contributions can be summarized as follows:

- (i) Compared with the current mainstream SR methods which use simulated LR-HR paired data for training, we focus on real-world application and use a multisensor dataset, OLI2MSI, to train SR model, which can be applied to super-resolve Landsat-8 data and can get better performance than that trained on simulated data
- (ii) We introduce a new dataset, OLI2MSI, a real-world multisensor dataset for remote sensing imagery SR with an upscale factor of 3. Images taken from Sentinel-2 MSI serve as ground truth of the LR images which are taken from Landsat-8 OLI
- (iii) We propose a novel conditional GAN scheme for SR task which can further reduce the solution-space. Thus, the super-resolved images can not only have high fidelity but also be more realistic and have high perceptual quality

TABLE 1: Some specifications of OLI2MSI dataset.

	OLI	MSI
Satellite	Landsat 8	Sentinel 2B
Selected common bands	2 (blue), 3 (green), 4 (red)	2 (blue), 3 (green), 4 (red)
Spatial resolution	30 m, 30 m, 30 m	10 m, 10 m, 10 m
Footprints/granules	Path: 126, row: 38; Path: 126, row: 39; Path: 126, row: 40; Path: 126, row: 41; Path: 126, row: 42;	49RBM, 49RCQ, 48RYQ, 48RYS, 49RBL, 49RCN, 49RDM, 49RBK, 49SDR, 49RBQ, 49RDN, 49RDQ, 49RCH, 49RCP, 49SCR, 49SDS, 48RYR, 49RCM, 49RBN, 49RCK, 49RBP, 49SCS, 49RBH, 48RYP, 49RCL, 49RDL, 49RBJ, 49RDP, 49RCJ, 48RYN
Sensing time range	September 23, 2019, 03:14:46-03:16:21 (UTC)	September 23, 2019, 03:28:55-03:31:15 (UTC)

2. Materials and Methods

2.1. OLI2MSI Dataset. We use a real-world multisensor LR-HR dataset, i.e., OLI2MSI, to train SR models instead of simulated datasets. The OLI2MSI dataset is composed of OLI and MSI images, where OLI images serve as LR images and MSI images are regarded as ground truth HR images.

The Landsat series of sensors have acquired the longest continuous series of image observations of the Earth's surface [42]. The sensor OLI, on-board Landsat-8, collects image data for 9 shortwave spectral bands over a 190 km swath with a 30 meter (m) spatial resolution for all bands except the 15 m Pan band [43]. Sentinel-2 is an earth observation mission from the Copernicus program developed by the European Space Agency (ESA). The mission is a constellation with two identical satellites, Sentinel-2A and Sentinel-2B, to meet the high frequent revisits requirement, which are phased 180 degrees from each other on the same orbit. The on-board instrument MSI has 13 spectral channels in the visible/near-infrared and short wave infrared spectral range: four bands at 10 m, six bands at 20 m, and three bands at 60 m spatial resolution [44]. The orbital swath width is 290 km. Within the 13 bands, the 10 m spatial resolution allows for continued collaboration with the SPOT-5 and Landsat-8 missions. Both OLI and MSI have stringent radiometric performance requirements and can provide well-calibrated sensor data. Barsi et al. [45] demonstrated that OLI and MSI showed stable radiometric calibration, with consistency between matching spectral bands to approximately 2.5%. This creates a prerequisite for us to make a paired dataset based on OLI and MSI images.

In order to build the paired dataset for SR, it is necessary for the two sensors to image common ground targets in the same location and spectral bands of the electromagnetic spectrum. For the sensors investigated in this work, i.e., OLI and MSI, there are 6 common bands (band 2, 3, 4, 5, 6, and 7 for OLI and band 2, 3, 4, 8a, 11, and 12 for MSI, respectively). Because the spatial resolution of the 6 bands for MSI is not the same, we finally choose the three bands of blue, green, and red, resulting in an upscale factor of 3 from OLI to MSI. We select the southwest region of China as study area, which contains abundant surface and ecological types, such as forest, farmland, lake, and urban residential area.

In order to minimize the difference in atmospheric conditions and environmental changes, we search all the OLI level-1c data and MSI level-1c data in the study area within a temporal window less than 1 hour and then filter by selecting the data less contaminated by cloud. The final selected scenes (for Sentinel-2 are granules) are listed in Table 1 with their footprints shown in Figure 1. All the Landsat-8 OLI data used in this study were downloaded from the United States Geological Survey (USGS) Earth Resources Observation and Science (EROS) Data Center (<https://earthexplorer.usgs.gov/>). Also, Sentinel-2 MSI data can be accessed from the Copernicus Open Access Hub (<https://scihub.copernicus.eu/>). The digital number (DN) values of OLI and MSI level-1c data are converted to top-of-atmosphere (TOA) reflectance. Atmosphere correction is not performed because atmosphere conditions are similar on the condition that the temporal windows between matched OLI and MSI scenes are less than one hour. Since Landsat-8 and Sentinel-2 have spatial misalignment that varies regionally depending on ground control point quality [46], all the OLI data are resampled to 10 m resolution using the bilinear upsampling method followed by a registration step to MSI images by ECC algorithm [47]. All data are then cropped to 480×480 pixels images without overlapping. We finally get 5325 images and randomly divided them into two parts: 5225 images for training set and 100 images for testing set. Some LR-HR image-pair samples in OLI2MSI training set can be seen from Figure 2, where the HR images are less blurry and have more details than the LR ones.

2.2. Methodology. A generative adversarial network (GAN) is composed of a generator G and a discriminator D , which aims to model a distribution by forcing the generated samples from random noise by G to be indistinguishable from the real samples by D . The GAN objective is to find a Nash equilibrium to the following two players min-max problem as described in its original form [48]:

$$\min_G \max_D \mathbb{E}_{y \sim q_{\text{data}}(y)} [\log D(y)] + \mathbb{E}_{z \sim p(z)} [\log (1 - D(G(z)))], \quad (1)$$

where z is a random noise vector drawn from distribution p , and y is a real sample drawn from data distribution q_{data} . For our task, the objective of the generator G is to

FIGURE 1: Landsat8 and Sentinel2 footprints on the region of interest.

FIGURE 2: LR-HR image-pair samples in OLI2MSI training set. The first row is LR images from Landsat8 OLI, while the second row is the corresponding HR images from Sentinel2 MSI.

map the input LR images to super-resolved images, while the discriminator D aims to distinguish HR images from the super-resolved images. In this context, the input of G is not random noise vector z but an observed LR image termed as x thereafter, and y is the corresponding real HR image. Several works have used GANs for SR task [38, 49], but only applied in an unconditional mode. These works usually rely on other terms (e.g., L2 or L1 norm) to force the output to be conditioned on the input, which will cause blurry effects and low perceptual quality. Though the GAN framework has been adopted to reduce the solution-space, the coefficient of GAN loss is usually so small that the capacity to reduce the solution-space is limited. In this work, we apply GAN in the conditional setting, which can further reduce the solution space. Consequently, the super-resolved images can not only have high fidelity but also be more realistic and have high perceptual quality.

In contrast to vanilla GANs, conditional GANs [39] learn a mapping from observed LR images x to the super-resolved ones $G(x)$, while the discriminator is modeling a

conditional distribution to distinguish from “real” HR images y to the “fake” generated images on the condition of x [50]. Practically, we adopt the LSGAN [51] variant of this optimization problem, thus, the objective of a conditional GAN can be expressed as follows:

$$\mathcal{L}_{\text{cGAN}}(G, D) = \mathbb{E}_{x,y} [(D(y|x) - 1)^2] + \mathbb{E}_x [D(G(x)|x)^2], \quad (2)$$

where the purpose of training G is to minimize this objective against that of D to maximize it, i.e., $G^* = \arg \min_G \max_D \mathcal{L}_{\text{cGAN}}(G, D)$. In order to test the importance of conditioning the D , we compare to SR models trained discriminatively (i.e., not GAN framework) and an unconditional variant in which the D does not observe x :

$$\mathcal{L}_{\text{cGAN}}(G, D) = \mathbb{E}_y [(D(y) - 1)^2] + \mathbb{E}_x [D(G(x))^2]. \quad (3)$$

In practice, training GANs only using $\mathcal{L}_{\text{cGAN}}(G, D)$ may result in mode collapse. Previous GAN-based SR works

FIGURE 3: Network architecture of the conditional patchGAN.

would mix the GAN objective with a content loss, such as L2 or L1 loss. For the proposed conditional GAN scheme for SR task, the job of D remains unchanged, i.e., to distinguish the “real” HR images from the “fake” generated images, except in the fact that the conditional GAN is trying to model a conditional probability distribution while the vanilla GAN is unconditional. For the G , it aims not only to fool the D but also to generate results which are near to the ground truth, that is what the L2 or L1 loss does as a fidelity term. We also adopted this strategy, using L1 loss as the fidelity term as it encourages less blurring [52]:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y}[\|y - G(x)\|_1]. \quad (4)$$

This is the most widely used optimization target for image SR, because it can effectively capture low-frequency information. Though it can achieve high peak signal-to-noise ratio (PSNR) [53] and structural similarity (SSIM) [54] score, the super-resolved images often lack high-frequency contents and have low perceptual quality with overly smooth textures. Zhang et al. [55] have demonstrated that the PSNR and SSIM are simple, shallow functions and proposed the Learned Perceptual Image Patch Similarity (LPIPS) metric which can better account for many nuances of human perception. In this work, we introduce LPIPS as a perceptual loss term, which can drive the SR solution-space towards the natural image manifold producing perceptually more convincing solutions:

$$\mathcal{L}_{LPIPS}(G) = \mathbb{E}_{x,y}[LPIPS(y, G(x))], \quad (5)$$

where the LPIPS is the pretrained network from [55]. The overall loss function can be formulated as follows:

$$G^* = \arg \min_G \max_D \mathcal{L}_{L1}(G) + \lambda_1 \mathcal{L}_{cGAN}(G, D) + \lambda_2 \mathcal{L}_{LPIPS}(G), \quad (6)$$

where the λ_1 and λ_2 are the corresponding weight coefficients of $\mathcal{L}_{cGAN}(G, D)$ and $\mathcal{L}_{LPIPS}(G)$, respectively.

2.3. Network Architectures. In this work, we focus on the conditional GAN scheme to further reduce the solution-

space which allows the generator can learn to create solutions images that not only have high fidelity but also are more realistic and have high perceptual quality. Thus, we directly adopt the state-of-the-art SR network DRN [35] as our generator G . As for the discriminator D , it aims to learn the conditional probability distribution of patches of the input and discriminate between real HR patches from this distribution and fake super-resolved patches generated by G . Therefore, we design a conditional patch discriminator architecture termed as conditional PatchGAN to capture the statistics of local patches only.

For the SR task, i.e., a low-level computer vision task which mainly concerns the textures of local patches, it is not necessary to capture the global contextual information, but only information of each patch with relatively suitable size in an image. To discriminate patches of an image, we use a fully convolutional patch discriminator as introduced in [50]. We use convolution with strides and no pooling layers to achieve a relatively big receptive field of a $N \times N$ patch (where N can be much smaller than the full size of the image). Therefore, the discriminator can implicitly classify each $N \times N$ patch separately to be real or fake. The output of D is a heat map where each pixel in this map indicates how likely its surrounding patch is to be drawn from the learned patch distribution. See Figure 3 for architecture details of the discriminator.

Such a conditional PatchGAN discriminator which penalizes structures at the scale of $N \times N$ patches is sufficient to discriminate between HR images and super-resolved images by G on the condition of the input LR images. Compared to the discriminators with a fully-connective layer at the last and output a scalar that indicates how the whole image fits the distribution of natural images, the conditional PatchGAN can model the image as a Markov random field with an independent pixel diameter of a patch size which can be regarded as a form of textual loss [56–58], with fewer parameters and can be applied to arbitrary size images. In addition, the patch size N can be flexibly adjusted by the stride of convolution layers or the number of convolution layers with strides.

3. Results

3.1. Datasets and Implementation Details. We compare different methods on both simulated data and OLI2MSI

TABLE 2: Performance comparisons with state-of-the-art algorithms in terms of 4 evaluation metrics on simulated dataset and the introduced OLI2MSI dataset. Note that a higher value of PSNR and SSIM is better, while for LPIPS and NIQE score, a lower value is better.

Algorithms	Trained on simulated dataset				Trained on OLI2MSI dataset			
	PSNR	SSIM	LPIPS	NIQE	PSNR	SSIM	LPIPS	NIQE
Bicubic	33.24	0.879	0.351	22.24	33.24	0.879	0.351	22.24
SRCNN	33.66	0.896	0.237	18.12	35.26	0.907	0.218	18.46
ESPCN	33.65	0.895	0.228	18.17	35.41	0.910	0.212	18.05
VDSR	33.64	0.895	0.215	17.86	35.87	0.919	0.181	16.61
SRResNet	33.59	0.896	0.217	17.55	37.13	0.939	0.131	13.25
EDSR	33.56	0.895	0.212	16.76	37.68	0.946	0.113	12.08
DBPN	33.57	0.895	0.214	17.11	37.60	0.945	0.116	12.26
RCAN	33.58	0.896	0.215	17.05	37.85	0.947	0.111	11.86
DRN	33.60	0.896	0.217	17.31	37.54	0.944	0.119	12.40
SRGAN	32.70	0.878	0.145	6.06	33.63	0.879	0.081	2.75
ESRGAN	32.32	0.878	0.151	5.90	33.82	0.888	0.075	2.45
cDRSRGAN(ours)	33.59	0.896	0.204	16.83	37.41	0.943	0.057	9.81

dataset. For the simulated data, we directly down-sample (bicubic) HR images in OLI2MSI dataset with a scale factor of 3 to the LR images. We train all networks on PyTorch framework with TITAN xp GPU. In case of memory issues for deeper networks, we use two TITAN xp GPUs to keep the same batch size.

During the training process, we randomly crop a 192×192 patch from HR image and the corresponding 64×64 patch from LR image in OLI2MSI dataset (for simulated data, the LR patch is the down-sampled version of the HR patch). We augment the input data with random flip and rotation before feeding to networks with a batch size of 16. For optimization, we use Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. All the networks are trained for 200,000 iterations (≈ 900 epochs) with an initial learning rate of 3×10^{-4} and decayed to 1×10^{-7} by a cosine annealing strategy. It takes about 2 days to train the proposed method with 2 TITAN xp GPUs. Compared with the non-GAN network, it will take almost 2 times the time to train a GAN-based one.

In the process of testing, the D is discarded. It takes ≈ 0.5 second to super-resolve a 160×160 image with a TITAN xp GPU. More details of testing and evaluation will be introduced in the next two sections.

3.2. Evaluation Metrics. We adopt four image quality metrics to quantitatively evaluate the quality of super-resolved images. Besides the commonly used peak signal-to-noise ratio (PSNR) [53] and structure similarity index (SSIM) [54] which can be regarded as fidelity metrics that measure the difference between ground-truth HR images and super-resolved images, we also adopt two perceptual metrics, i.e., the learned perceptual image patch similarity (LPIPS) [55] metric and naturalness image quality evaluator (NIQE) [59] which can better account for many nuances of human perception. Note that NIQE is a no-reference image quality evaluator.

For fair comparison, all the PSNR and SSIM measurements are calculated on the Y-channel of image, with a border crop of 5-pixels wide for each border. The version of the

LPIPS metric we used is 0.1, and the default settings are adopted during training and inference. Before calculating the NIQE score, we train a custom NIQE model using all the HR images in OLI2MSI to model the distribution of remote sensing images. Among the four image quality metrics, a higher value of PSNR and SSIM indicates a higher fidelity for HR images, while a lower value of LPIPS and NIQE score means better visual quality and more agree with human perception.

3.3. Comparison with State-of-the-Art Methods. We extensively evaluate the SR performance from different networks trained on both the simulated paired remote sensing data (bicubic down-sampling) and multisensor satellite imagery, i.e., the introduced OLI2MSI dataset. We choose a few representative super-resolution (SR) methods for comparisons: the discriminatively trained networks which demonstrate different network architectures for SR, including SRCNN [25], ESPCN [26], VDSR [27], SRResNet [38], EDSR [28], DBPN [29], RCAN [33], DRN [35], and also GAN-based SR models, e.g., SRGAN [38] and ESRGAN [49]. For the sake of fair comparisons, all models are retrained from scratch on the used datasets.

In order to validate the effectiveness of the introduced OLI2MSI dataset in remote sensing imagery SR task, we train all the networks with the simulated paired SR dataset in the way proposed in their papers and then use the testset of OLI2MSI to test. Average values of the 4 quantitative evaluation metrics of the reconstructed results over the entire testset are given in Table 2. From Table 2, it can be seen that all the models trained on OLI2MSI dataset can have better performance than those trained on simulated dataset in terms of both fidelity and visual quality evaluation metrics, which further illustrates the necessity and effectiveness of adopting a multisensor dataset in remote sensing imagery SR task. The proposed conditional GAN scheme for SR, which adopts the DRN network [35] as the generator (termed as cDRSRGAN), is actually a GAN-based SR model. On the one hand, from the results trained on OLI2MSI

FIGURE 4: Visual comparisons of different methods trained on OLI2MSI dataset, and the numbers in the bracket refer to the PSNR, SSIM, LPIPS, and NIQE values of the corresponding result. (a) LR(-/-/-/-). (b) SRCNN (26.44/0.693/0.490/17.54). (c) ESPCN (26.54/0.701/0.484/17.24). (d) VDSR (26.90/0.721/0.438/15.25). (e) EDSR (28.72/0.819/0.289/10.91). (f) DBPN (28.43/0.808/0.314/11.32). (g) RCAN (28.66/0.819/0.292/10.88). (h) DRN (28.28/0.803/0.335/11.50). (i) SRGAN (24.61/0.610/0.200/4.95). (j) ESRGAN (24.65/0.626/0.183/3.45). (k) cDRSRGAN (28.39/0.814/0.160/9.43). (l) HR(-/-/-/4.11).

dataset in Table 2, we can see that cDRSRGAN can achieve comparable PSNR and SSIM scores with the SOTA network (i.e., DRN) when compared to other discriminatively trained networks. On the other hand, the cDRSRGAN can get better visual quality evaluation metrics while ensuring the fidelity. As for the two other GAN-based methods, although they have better performance in visual quality evaluation metrics, they often behave poorly in fidelity and may cause artifacts in many cases. The visual comparisons in Figure 4 show that our model produces sharper edges and shapes, while other baselines may give more blurry ones. Especially for the SRGAN and ESRGAN, they have much lower scores in LPIPS and NIQE. This is mainly due to the artifacts produced by GAN, which are not real textures in the scene.

One point that needs to be emphasized again is that although cDRSRGAN does not have best performance in either fidelity terms (PSNR, SSIM) or visual quality terms (LPIPS, NIQE), it can produce super-resolved images that not only have high fidelity but also be more realistic with high perceptual quality due to the conditional GAN training scheme to further reduce the solution-space. The visual

comparisons results in Figure 4 demonstrate the effectiveness of the proposed scheme in generating more accurate and visually promising super-resolved images.

In order to demonstrate the generalization capacity of the proposed method, we apply our method to other Landsat8-OLI images representing distinct surface types. These images are neither in the training set nor in the testset of OLI2MSI dataset. From Figure 5, we can see that the super-resolved images by our method can obtain a promising performances, which have more sharp textures and details. This is because what the network learns is the mapping from the low-resolution images to high-resolution images, which represents the degradation from Sentinel2-MSI images to Landsat8-OLI images and is not limited to any land surface type of remote sensing images.

3.4. Ablation Study on Conditional GAN Scheme. We conduct an ablation study on the conditional GAN scheme and the introduced LPIPS loss and report the results in Table 3. Compared to the baseline (i.e., DRN), we adopt the vanilla GAN scheme to train the DRN (i.e., set the

FIGURE 5: Visual results of the proposed method in Landsat8-OLI images with distinct land surface types. Note that these images are neither in the training set nor in the testset of OLI2MSI dataset. The first row is the Landsat8-OLI LR images, the second row is the super-resolved images by our method, and the third row is the corresponding HR images from Sentinel2-MSI.

DRN as the generator, termed as DRSRGAN) and get a poorer (lower) PSNR and SSIM score while a better (lower) LPIPS and NIQE score as expected. It adopts the L1 loss as the content loss which can effectively capture low-frequency information while the high-frequency details and textures in the super-resolved images mainly come from the artifacts caused by GAN. Once the conditional GAN scheme is adopted (termed as cDRSRGAN), the model yields an even higher PSNR& SSIM and a lower LPIPS& NIQE score than the baseline due to the fact that the conditional GAN scheme is different from L1 loss and can reduce the solution-space in another aspect which results in better performance in both fidelity and visual quality. It can be seen that introducing the LPIPS loss as the perceptual loss can yield more visually pleasing results with a small drop in PSNR& SSIM. These results suggest that the conditional GAN scheme can effectively improve the reconstruction of HR images by introducing an additional constraint to reduce the solution space.

To test the effect of receptive field size in the conditional PatchGAN discriminator, we vary the patch size N by changing the number of blocks in D . From Table 4, it demonstrates that the super-resolution performance is not better as the receptive field size increases. Using a PatchGAN with small receptive field size will lead to some artifacts, therefore, resulting in a low PSNR value. As the N increases, the artifacts can be alleviated slightly and get better scores. However, an excessive large N will get

TABLE 3: The impact of the proposed conditional GAN scheme and LPIPS loss on super-resolution performance in terms of PSNR, SSIM, LPIPS, and NIQE score on the OLI2MSI dataset.

Model	PSNR	SSIM	LPIPS	NIQE
DRN (baseline)	37.54	0.944	0.119	12.40
DRSRGAN	36.36	0.933	0.054	5.63
cDRSRGAN	37.71	0.947	0.099	11.43
cDRSRGAN+LPIPS	37.41	0.943	0.057	9.81

TABLE 4: Super-resolution performance varying the receptive field size of conditional PatchGAN discriminator in terms of PSNR, SSIM, LPIPS, and NIQE score on the OLI2MSI dataset. Note that the input images are 192×192 and larger receptive fields are padding with zeros.

Receptive field	Blocks	PSNR	SSIM	LPIPS	NIQE
15×15	1	36.18	0.926	0.046	3.65
47×47	3	37.41	0.943	0.057	9.81
255×255	5	36.95	0.937	0.052	8.57

considerably lower scores. This is because a relatively large N is sufficient. Additionally, the conditional PatchGAN has more parameters as the number of blocks grows, thus would be harder to train.

4. Conclusion

In this paper, we introduce a new multisensor paired super-resolution dataset (i.e., OLI2MSI) and proposed a novel conditional GAN scheme to super-resolved real-world remote sensing satellite imagery. The OLI2MSI is a satellite remote sensing imagery dataset composed of Landsat8-OLI and Sentinel2-MSI images, where OLI images serve as LR images and MSI images are regarded as ground truth HR images. Experiments demonstrate that networks trained on this dataset can have better performance than that trained on simulated dataset (e.g., bicubic down-sampling). Furthermore, the proposed conditional GAN scheme can further reduce the solution-space of SR. Thus, the super-resolved images can not only have high fidelity but also be more realistic and have high perceptual quality. Extensive experiments show the superiority of our method on fidelity and perceptual quality over the considered baseline methods.

Data Availability

All the Landsat-8 OLI images used in this study can be downloaded from the United States Geological Survey (USGS) Earth Resources Observation and Science (EROS) Data Center (<https://earthexplorer.usgs.gov/>). Also, the Sentinel-2 MSI images can be accessed from the Copernicus Open Access Hub (<https://scihub.copernicus.eu/>). The OLI2MSI dataset introduced in this study can be download in <https://github.com/wjwjww/OLI2MSI>.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Authors' Contributions

J. Wang conceived the idea, conducted the datasets, and designed the experiments. Z. Zhang conducted the experiments. Z. Hu and Q. Wu helped data processing. K. Gao, C. Ni, and D. Chen revised the paper. All authors contributed equally to the writing of the manuscript.

Acknowledgments

This work was supported by the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology (grant number GZZKFJJ2020004); the National Natural Science Foundation of China (grant numbers 61875013 and 61827814); and the Natural Science Foundation of Beijing Municipality (grant number Z19J00064).

References

- [1] C. M. Viana, I. Girão, and J. Rocha, "Long-term satellite image time-series for land use/land cover change detection using refined open source data in a rural region," *Remote Sensing*, vol. 11, no. 9, p. 1104, 2019.
- [2] C. Yang, J. H. Everitt, Q. du, B. Luo, and J. Chanussot, "Using high-resolution airborne and satellite imagery to assess crop growth and yield variability for precision agriculture," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 582–592, 2012.
- [3] H. T. Verstappen, "Aerospace technology and natural disaster reduction," *Advances in Space Research*, vol. 15, no. 11, pp. 3–15, 1995.
- [4] Y. Xu, L. Lin, and D. Meng, "Learning-based sub-pixel change detection using coarse resolution satellite imagery," *Remote Sensing*, vol. 9, no. 7, p. 709, 2017.
- [5] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE signal processing magazine*, vol. 20, no. 3, pp. 21–36, 2003.
- [6] K. Nasrollahi and T. B. Moeslund, "Super-resolution: a comprehensive survey," *Machine vision and applications*, vol. 25, no. 6, pp. 1423–1468, 2014.
- [7] I. Demir, K. Koperski, D. Lindenbaum et al., "A challenge to parse the earth through satellite images," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 18–22, Salt Lake City, UT, USA, 2018.
- [8] N. Hajlaoui, C. Chaux, G. Perrin, F. Falzon, and A. Benazza-Benyahia, "Satellite image restoration in the context of a spatially varying point spread function," *Journal of the Optical Society of America A*, vol. 27, no. 6, pp. 1473–1481, 2010.
- [9] F. Li, X. Jia, and D. Fraser, "Universal hmt based super resolution for remote sensing images," in *2008 15th IEEE International Conference on Image Processing*, pp. 333–336, San Diego, CA, USA, 2008.
- [10] B. C. Tom, N. P. Galatsanos, and A. K. Katsaggelos, "Reconstruction of a high resolution image from multiple low resolution images," in *Super-Resolution Imaging*, pp. 73–105, Springer, 2002.
- [11] C.-Y. Yang, C. Ma, and M.-H. Yang, "Single-image super-resolution: a benchmark," in *European conference on computer vision*, pp. 372–386, Springer, 2014.
- [12] F. Salvetti, V. Mazzia, A. Khaliq, and M. Chiaberge, "Multi-image super resolution of remotely sensed images using residual attention deep neural networks," *Remote Sensing*, vol. 12, no. 14, p. 2207, 2020.
- [13] R. Tsai, "Multiframe image restoration and registration," *Advance Computer Visual and Image Processing*, vol. 1, pp. 317–339, 1984.
- [14] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graphical models and image processing*, vol. 53, no. 3, pp. 231–239, 1991.
- [15] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE transactions on image processing*, vol. 13, no. 10, pp. 1327–1344, 2004.
- [16] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1167–1183, 2002.
- [17] Z. Lin and H. Y. Shum, "Fundamental limits of reconstruction-based superresolution algorithms under local translation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 83–97, 2004.
- [18] D. Yang, Z. Li, Y. Xia, and Z. Chen, "Remote sensing image super-resolution: challenges and approaches," in *2015 IEEE International Conference on Digital Signal Processing (DSP)*, pp. 196–200, Singapore, 2015.
- [19] L. Alparone, B. Aiazzi, S. Baronti, and A. Garzelli, *Remote Sensing Image Fusion*, Crc Press, 2015.

- [20] F. Zhou, W. Yang, and Q. Liao, "Interpolation-based image super-resolution using multisurface fitting," *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3312–3318, 2012.
- [21] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.
- [22] R. Timofte, V. De Smet, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *2013 IEEE International Conference on Computer Vision*, pp. 1920–1927, Sydney, NSW, Australia, December 2013.
- [23] G. Polatkan, M. Zhou, L. Carin, D. Blei, and I. Daubechies, "A Bayesian nonparametric approach to image super-resolution," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 2, pp. 346–358, 2015.
- [24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [25] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.
- [26] W. Shi, J. Caballero, F. Huszar et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874–1883, Las Vegas, NV, USA, June 2016.
- [27] J. Kim, J. Kwon Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1646–1654, Las Vegas, NV, USA, June 2016.
- [28] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144, Honolulu, HI, USA, July 2017.
- [29] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1664–1673, Salt Lake City, UT, USA, June 2018.
- [30] W.-S. Lai, J. B. Huang, N. Ahuja, and M. H. Yang, "Fast and accurate image super-resolution with deep Laplacian pyramid networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 11, pp. 2599–2613, 2018.
- [31] J. He, C. Dong, and Y. Qiao, "Modulating image restoration with continual levels via adaptive feature modification layers," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11056–11064, Long Beach, CA, USA, June 2019.
- [32] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11065–11074, Long Beach, CA, USA, June 2019.
- [33] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 286–301, Munich, Germany, 2018.
- [34] R. Feng, J. Gu, Y. Qiao, and C. Dong, "Suppressing model overfitting for image super resolution networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, June 2019.
- [35] Y. Guo, J. Chen, J. Wang et al., "Closed-loop matters: dual regression networks for single image super-resolution," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5407–5416, Seattle, WA, USA, June 2020.
- [36] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 pirm challenge on perceptual image super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018.
- [37] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "Pulse: self-supervised photo upsampling via latent space exploration of generative models," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2437–2445, Seattle, WA, USA, June 2020.
- [38] C. Ledig, L. Theis, F. Huszar et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4681–4690, Honolulu, HI, USA, July 2017.
- [39] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, <https://arxiv.org/abs/1411.1784>.
- [40] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International Conference on Machine Learning*, pp. 1060–1069, PMLR, 2016.
- [41] Z. Liu, H. Chen, and H. Liu, "Deep learning based framework for direct reconstruction of pet images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 48–56, Springer, 2019.
- [42] USGS, "Landsat missions," <https://www.usgs.gov/core-science-systems/nli/landsat>.
- [43] "Landsat 8 data users handbook," <https://www.usgs.gov/core-science-systems/nli/landsat/landsat-8-data-users-handbook>.
- [44] ESA, "Sentinel-2 user handbook," https://sentinels.copernicus.eu/documents/247904/685211/Sentinel-2_User_Handbook.
- [45] J. A. Barsi, B. Alhammoud, J. Czaplá-Myers et al., "Sentinel-2a msi and landsat-8 oli radiometric cross comparison over desert sites," *European Journal of Remote Sensing*, vol. 51, no. 1, pp. 822–837, 2018.
- [46] J. Storey, D. P. Roy, J. Masek, F. Gascon, J. Dwyer, and M. Choate, "A note on the temporary misregistration of landsat-8 operational land imager (oli) and sentinel-2 multi spectral instrument (msi) imagery," *Remote Sensing of Environment*, vol. 186, pp. 121–122, 2016.
- [47] G. D. Evangelidis and E. Z. Psarakis, "Parametric image alignment using enhanced correlation coefficient maximization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1858–1865, 2008.
- [48] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, pp. 2672–2680, 2014.
- [49] X. Wang, K. Yu, S. Wu et al., "Esrgan: enhanced super-resolution generative adversarial networks," *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [50] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1125–1134, Honolulu, HI, USA, July 2017.
- [51] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *2017 IEEE*

- International Conference on Computer Vision (ICCV)*, pp. 2794–2802, Venice, Italy, October 2017.
- [52] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, “Autoencoding beyond pixels using a learned similarity metric,” in *International conference on machine learning*, pp. 1558–1566, PMLR, 2016.
- [53] A. Hore and D. Ziou, “Image quality metrics: Psnr vs. ssim,” in *2010 20th International Conference on Pattern Recognition*, pp. 2366–2369, Istanbul, Turkey, August 2010.
- [54] 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.
- [55] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, Salt Lake City, UT, USA, June 2018.
- [56] C. Li and M. Wand, “Precomputed real-time texture synthesis with markovian generative adversarial networks,” in *European conference on computer vision*, pp. 702–716, Springer, 2016.
- [57] A. A. Efros and T. K. Leung, “Texture synthesis by non-parametric sampling,” in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 1033–1038, Kerkyra, Greece, 1999.
- [58] L. Gatys, A. S. Ecker, and M. Bethge, “Texture synthesis using convolutional neural networks,” *Advances in neural information processing systems*, vol. 28, pp. 262–270, 2015.
- [59] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a “completely blind” image quality analyzer,” *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.