International Journal of Applied Earth Observations and Geoinformation 104 (2021) 102543

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag

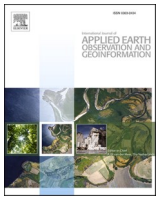

Image super-resolution with dense-sampling residual channel-spatial attention networks for multi-temporal remote sensing image classification

Yue Zhu [a,*], Christian Geiß [b], Emily So [a]

[a] The Department of Architecture, University of Cambridge, CB2 1TN Cambridge, UK
[b] The German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), 82234 Weßling-Oberpfaffenhofen, Germany

ARTICLE INFO

Keywords:
Image super-resolution
Convolutional neural networks
Attention mechanism
Dense connection
Multi-temporal land use classification

ABSTRACT

Image super-resolution (SR) techniques can benefit a wide range of applications in the remote sensing (RS) community, including image classification. This issue is particularly relevant for image classification on time series data, considering RS datasets that feature long temporal coverage generally have a limited spatial resolution. Recent advances in deep learning brought new opportunities for enhancing the spatial resolution of historic RS data. Numerous convolutional neural network (CNN)-based methods showed superior performance in terms of developing efficient end-to-end SR models for natural images. However, such models were rarely exploited for promoting image classification based on multispectral RS data. This paper proposes a novel CNN-based framework to enhance the spatial resolution of time series multispectral RS images. Thereby, the proposed SR model employs Residual Channel Attention Networks (RCAN) as a backbone structure, whereas based on this structure the proposed models uniquely integrate tailored channel-spatial attention and dense-sampling mechanisms for performance improvement. Subsequently, state-of-the-art CNN-based classifiers are incorporated to produce classification maps based on the enhanced time series data. The experiments proved that the proposed SR model can enable unambiguously better performance compared to RCAN and other (deep learning-based) SR techniques, especially in a domain adaptation context, i.e., leveraging Sentinel-2 images for generating SR Landsat images. Furthermore, the experimental results confirmed that the enhanced multi-temporal RS images can bring substantial improvement on fine-grained multi-temporal land use classification.

1. Introduction

The value of high spatial and temporal resolution remote sensing (RS) data has been widely recognized in terms of the improvement in the quality of multi-temporal land use and land cover (LULC) classification maps (Vuolo et al., 2018). However, publicly accessible RS datasets that feature a high spatial resolution mostly do not have long temporal coverage. For instance, the temporal coverage of Sentinel-2 started in 2015. Even for commercially available datasets, their temporal coverage commonly started from the year 2000. The limited temporal coverage of high-resolution or medium-resolution RS datasets substantially restricts the analysis of long time series. As for the RS datasets having much longer temporal coverage (e.g., over 30 years), they usually have a much lower spatial resolution. For example, Landsat datasets have been widely used for time series land dynamic analysis due to their long temporal coverage. However, the spatial resolution of Landsat data is 30 m for the modern platforms in the multi-spectral domain, the relatively coarse spatial resolution considerably limits the research on long-term yet fine-grained land change observations.

Recent advances in the field of deep learning provide new opportunities for improving the quality of long time series LULC maps. With super-resolution (SR) deep networks, the spatial resolution of long time series imagery can be largely improved by taking advantage of newly produced high-resolution RS imagery.

1.1. Convolutional neural networks (CNNs) for image super-resolution

Methods developed for addressing single image super-resolution (SISR) problems have been extensively studied. In general, SISR methods can be mainly categorized into interpolation-based methods (e.g., bicubic interpolation) and learning-based methods. Compared with interpolation-based methods, learning-based methods, especially deep learning-based methods, are less dependent on handcrafted features (Kim et al., 2016). Also, many deep-learning-based SISR methods

* Corresponding author.
E-mail addresses: yz591@cam.ac.uk (Y. Zhu), christian.geiss@dlr.de (C. Geiß), ekms2@cam.ac.uk (E. So).

https://doi.org/10.1016/j.jag.2021.102543
Received 14 May 2021; Received in revised form 25 August 2021; Accepted 12 September 2021
Available online 20 September 2021

presented superior reconstruction ability with efficient end-to-end model structures.

In the field of deep learning, the development of CNN-based SISR methods made substantial progress over recent years. As one of the earliest deep learning methods for SISR, Super-resolution Convolutional Neural Network (SRCNN) (C. Dong et al., 2015) directly mapped images to a higher resolution in an end-to-end fashion. Based on SRCNN, Very Deep Super Resolution (VDSR) (Kim et al., 2016) was proposed with a residual learning mechanism and a deeper model structure. Both SRCNN and VDSR require upscaling schemes before low-resolution images are fed into networks. By contrast, Fast Super-Resolution Convolutional Neural Networks (FSRCNN) (C. Dong et al., 2016) and Enhanced Deep Super-resolution network (EDSR) (Lim et al., 2017) exhibited better performance with an up-sampling scheme embedded in the last part of the model. EDSR outperformed many state-of-the-art residual networks by introducing a residual-in-residual mechanism in the model structure to further deepen and widen the network. Subsequently, Residual Channel Attention Networks (RCAN) (Y. Zhang, Li, et al., 2018) provided another direction to improve the performance of CNN-based SR models, it emphasized the significant benefit of introducing a channel attention mechanism for increasing the channel-wise discriminative ability of the model. The adoption of channel attention in RCAN leads to an improvement of model performance while outperforming many other CNN-based methods, including SRCNN, VDSR, and EDSR. Moreover, efforts have been made in tailoring deep learning-based SISR methods for remote sensing images, including hyperspectral super-resolution (Gao et al., 2021; Zheng et al., 2019, 2021). For example, Zheng et al., (2019) proposed a deep network with separable-spectral convolution module designated for hyperspectral image super-resolution. Furthermore, attempts have been made in applying CNN-based methods to enhance the resolution of RS data for practical analysis. For instance, M. Chen et al., (2020) adopted SRCNN and FSRCNN for monitoring the invasion of exotic plants.

Alternatively, based on Generative Adversarial Networks (GAN) (Goodfellow et al., 2014), many GAN-based methods were developed for SISR applications, such as Super-resolution GAN (SRGAN) (Ledig et al., 2017) and Enhanced Super-Resolution GAN (ESRGAN) (X. Wang et al., 2018). GAN-based methods generally consist of at least two deep networks which act as generator and discriminator respectively. However, the multiple networks in GANs generally increase the difficulty of training, the training process of many GAN-based models tends to be highly unstable (Kodali et al., 2017). Given these considerations, we did not follow a GAN-based approach here.

1.2. Convolutional neural networks (CNNs) for image classification

Recent advances of CNN models, especially Fully Convolutional Networks (FCNs), in semantic segmentation tasks brought new opportunities for RS image classification. As one of the widely recognised FCNs, U-Net (Ronneberger et al., 2015) was initially proposed for biomedical image segmentation but also has achieved success in segmenting remote sensing images (McGlinchy et al., 2019; Pasquali et al., 2019; Schuegraf & Bittner, 2019; Yang et al., 2019; W. Zhang et al., 2021). However, a vanilla U-Net specialises in processing two-dimensional images with single or multiple spectral channels. Consequently, it lacks the capability of processing temporal sequential data efficiently. To deal with spatial–temporal data, Convolutional LSTM (ConvLSTM) network was proposed by X. Shi et al., (2015) for precipitation prediction. Considering the advantages of ConvLSTM in terms of extracting spatial–temporal features, attempts have been made to develop ConvLSTM for multi-temporal classification and prediction (Rußwurm & Körner, 2018; Teimouri et al., 2019; Yeom et al., 2020; Zhu et al., 2021). These developments for multi-temporal classification suggested that integration of ConvLSTM layers with FCN can generally bring improvements in classification accuracy.

1.3. Methods of improving CNN model performance

Much research has been focused on improving the performance of CNN models. Viable approaches include deepening the depth (He et al., 2015), expanding the width of networks (Szegedy et al., 2015), and increasing cardinality (Xie et al., 2017). These approaches generally require redesigns of the model structure to achieve improvements, whereas other methods can promote model performance by lightweight mechanisms that do not demand much network engineerings, such as attention mechanisms and dense connection mechanisms.

1.3.1. Attention mechanisms

Inspired by the importance of attention in human visual experience, numerous attention mechanisms were developed for promoting deep learning networks (F. Wang et al., 2017; Zagoruyko & Komodakis, 2017). Attention mechanisms in deep networks can be regarded as trainable weighted maps, which are functional in terms of guiding models to be more focused on important features in the data. In this manner, models with attention mechanisms can be less affected by the noise in the data, thereby become more efficient and robust.

Attention mechanisms can be mainly categorized into four types: (1) channel attention mechanism (Haut et al., 2019; Panboonyuen et al., 2019; W. Tong et al., 2020; Q. Wang et al., 2020; Y. Zhang, Li, et al., 2018), (2) spatial attention mechanism (W. Shi et al., 2020; Zhao et al., 2018), (3) temporal attention mechanism (Tran et al., 2017), and (4) hybrid attention mechanism (e.g., channel-spatial attention mechanism (J. Chen et al., 2020a; L. Chen et al., 2017; Muqeet et al., 2019; Woo et al., 2018; Xu & Li, 2019), spatial–temporal attention mechanism (Altaf et al., 2018)). Each type of attention mechanism can be effective in providing weighted features along the axis that they are implemented.

The implementation of channel attention boosted the performance of many CNNs, such as Squeeze-and-Excitation Networks (SE-Net) (Hu et al., 2019) and Efficient Channel Attention Networks (ECA-Net) (Q. Wang et al., 2020). Channel attention modules have been frequently adopted together with spatial attention modules. For instance, Residual Attention Network (F. Wang et al., 2017) was proposed with a stack of channel attention and spatial attention modules to produce attention-aware features for image classification. Moreover, Dual Attention Networks (DANet) (Fu et al., 2019) included spatial attention and channel attention modules in two sub-branches to capture global dependencies.

Tentative efforts have been made in developing channel-spatial attention blocks that can be efficiently incorporated into any feed-forward network, two representative examples are Bottleneck Attention Modules (BAM) (Park et al., 2018) and Convolutional Block Attention Module (CBAM) (Woo et al., 2018). BAM and CBAM are different regarding how they combine attention modules and where they are integrated into model structures. BAM arranges channel attention and spatial attention parallelly, whereas CBAM organizes them sequentially. Ablation studies of CBAM showed that sequential connections of channel attention and spatial attention can yield more enhancement than other combination sequences. Moreover, BAM can perform better when being applied at the bottlenecks (Park et al., 2018), whereas CBAM is proposed to be integrated inside residual blocks (Woo et al., 2018).

Regarding the applications of attention mechanisms in SISR methods, although RCAN and Multiscale Attention Network (MSAN) (S. Zhang et al., 2020) leveraged channel attention mechanisms to promote their reconstruction capability, they neglected the role of spatial attention mechanisms. Some other SISR methods attempted to value both channel-wise and spatial-wise attention. For instance, Multi-Grained Attention Networks (MGAN) (Wu et al., 2021) incorporated a multi-grained attention mechanism that can generate and multi-scale feature maps considering both channel-wise dependencies and spatial locations. Yao et al., (2020) proposed a cross-attention mechanism that can bridge the spatial importance of high-resolution images and the spectral importance of the low-resolution images for SISR performance gains.

1.3.2. Dense connection mechanism

The shortcut connections introduced in ResNet (He et al., 2015) enable models to exploit deeper structures and become easier to train. Based on ResNet, Dense Convolutional Network (DenseNet) (Huang et al., 2018) incorporated a dense connection mechanism, in which each layer was connected to every subsequent layer to formed a densely connected network. These dense skip connections are beneficial in mitigating gradient vanishing and enhancing feature propagation (Huang et al., 2018). Such promising effects have been widely observed in a variety of CNNs for diverse tasks.

Regarding the implementations of dense connections in SISR methods, SRDenseNet (T. Tong et al., 2017) implemented dense skip connections to propagate encoded feature maps to every subsequent layer, thereby low-level features can be integrated with high-level features to promote reconstruction ability. Moreover, Residual Dense Network (RDN) (Y. Zhang, Tian, et al., 2018) was proposed with stacked residual dense blocks (RDB), inside which every convolutional layer is densely connected. RDBs can facilitate the model to extract local dense features and achieve contiguous memory. Similarly, Wen et al., (2018) proposed densely connected residual networks (DRNet) with dense skip connections in residual blocks, they yielded higher PSNR values with relatively fewer parameters compared with EDSR.

The above mentioned densely connected CNN-based SISR methods applied the dense connection mechanism before up-sampling layers, whereas Dong et al., (2020) developed a dense-sampling super-resolution network (DSSR) which used skip connections between earlier layers and up-sampling layers. In this framework, each prior residual group was upscaled then densely connected with a corresponding up-sampling layer. Such type of dense connections is termed as a dense-sampling mechanism. According to their experimental results, the dense-sampling mechanism is particularly effective for SISR models due to its capability of integrating the features learned at various depths for constructing better high-frequency information. Therefore, we adopted and tested this mechanism as one of the integrated modules for achieving performance gains for our considered models.

Identifying a research gap, this paper aims to address the limitation of low-resolution historic RS images on fine-grained multi-temporal LULC classification through leveraging newly produced data. To achieve this aim, we propose a novel CNN-based super-resolution method for multi-temporal image classification. Specifically, two extension mechanisms, i.e., channel-spatial attention and densely sampling, were integrated to improve the performance of the considered SISR methods. The proposed SISR method can improve the spatial resolution of historic RS images, which can further lead to accuracy gains of image classification. Moreover, we implemented state-of-the-art CNN-based classification methods on super-resolution Landsat data to examine the extent to which the proposed method can benefit the classification accuracy on time series data. It is worth noting that most research on CNN-based SISR methods was developed on datasets that consist of natural images with three spectral bands (i.e., red, blue, green). However, given multiple spectral bands in RS images are critical for LULC classification, we developed the proposed framework with four spectral bands (i.e., red, blue, green, and NIR). Furthermore, the proposed SISR methods were developed based on publicly available RS datasets (i.e., Sentinel-2 and Landsat) to gain extensive practical values.

The remainder of this paper is organized as follows: section 2 introduces the proposed method. The experiment datasets and setup are described in section 3. Then results are reported in section 4 and the main findings are concluded in section 5.

2. Proposed SISR methods

Given RCAN exhibited superior performance than many other state-of-the-art CNN-based SISR methods, the proposed method employed RCAN as the backbone structure. However, we substantially redesigned the structure in three aspects. Firstly, the original channel attention mechanism was replaced with a hybrid channel-spatial attention mechanism (Fig. 1 (c)). Secondly, a dense-sampling mechanism was applied between each residual group and the upscaling block of the model (Fig. 1 (a)). Lastly, based on the original skip connections implemented in the inner structures of an RCAN, additional skip connections were added from the low-level feature maps to the following output of residual groups (Fig. 1 (a)). As such, our proposed method is termed as Dense-sampling Residual Channel-spatial Attention Network (D-RCSAN). The details are introduced as follows.

2.1. Channel-spatial attention mechanism

The channel-wise attention in RCAN enables the network to weigh the importance of each channel and thereby be more focused on prioritized channels than the remaining channels. However, RCAN treated the feature maps in spatial dimension homogenously. The proposed method replaced the channel-wise attention mechanism with CBAM (Woo et al., 2018) in every residual block to produce weighted feature maps. In a CBAM module, channel-wise and spatial-wise attention mechanisms are sequentially combined as follows:

$$F^{'} = A_c(F) \otimes F$$

$$F^{''} = A_s(F^{'}) \otimes F^{'}$$

where the feature map $F \in \mathbb{R}^{c\times h\times w}$ represents the input for a CBAM module, $A_c \in \mathbb{R}^{c\times 1\times 1}$ refers to the channel attention sub-module and $A_s \in \mathbb{R}^{1\times h\times w}$ refers to the spatial attention sub-module. Also, c, h, and w are the number of channels, height, and width of the feature map, respectively. Moreover, $\otimes$ denotes the operation of element-wise multiplication.

$$A_c(F) = \sigma\left(MLP\left(f^c_{avg}(F)\right) + MLP\left(f^c_{max}(F)\right)\right)$$

$$= \sigma\left(\left(W_1\left(W_0\left(f^c_{avg}(F)\right)\right)\right) + \left(W_1\left(W_0\left(f^c_{max}(F)\right)\right)\right)\right)$$

where f^c_{avg} and f^c_{max} denote the operation of average-pooling and max-pooling in the channel-attention sub-module respectively, MLP refers to a multilayer perception network containing one hidden layer, $W_0 \in \mathbb{R}^{c/r\times c}$ and $W_1 \in \mathbb{R}^{c\times c/r}$ denote the weights in the MLP, in which r refers to a ratio that changes the number of channels. σ refers to the sigmoid activation function. Ablation studies (Woo et al., 2018) claimed that the max-pooling is functional in encoding the most salient information, such encoding can compensate for some over-softened features caused by global average-pooling (Woo et al., 2018).

$$A_s(F^{'}) = \sigma\left(f^{7*7}\left(CAT\left(f^s_{mean}(F^{'}), f^s_{max}(F^{'})\right)\right)\right)$$

where f^s_{mean} and f^s_{max} denote the operation of getting the mean and maximum value of the feature maps through channel dimension. The results of $f^s_{mean}(F^{'})$ and $f^s_{max}(F^{'})$ are two 2D spatial attention maps. CAT denotes the operation of concatenating the two spatial attention maps. Then f^{7*7} a convolutional operational, which has a filter size of 7 by 7. As in the equation for channel attention sub-module, σ refers to the sigmoid activation function.

2.2. Dense sampling mechanism

Given deeper layers with dense connections can promote the performance of deep neural networks for SISR tasks (X. Dong et al., 2020; X. Wang et al., 2018; Wen et al., 2018), we integrated dense connections in the up-sampling part of the model architecture. The original RCAN model follows a late up-sampling scheme, which upscales the output of the last residual group to the target upscaled size before feeding it into the final output layer. Based on the original model structure, a dense-sampling structure is introduced in the up-sampling part of the

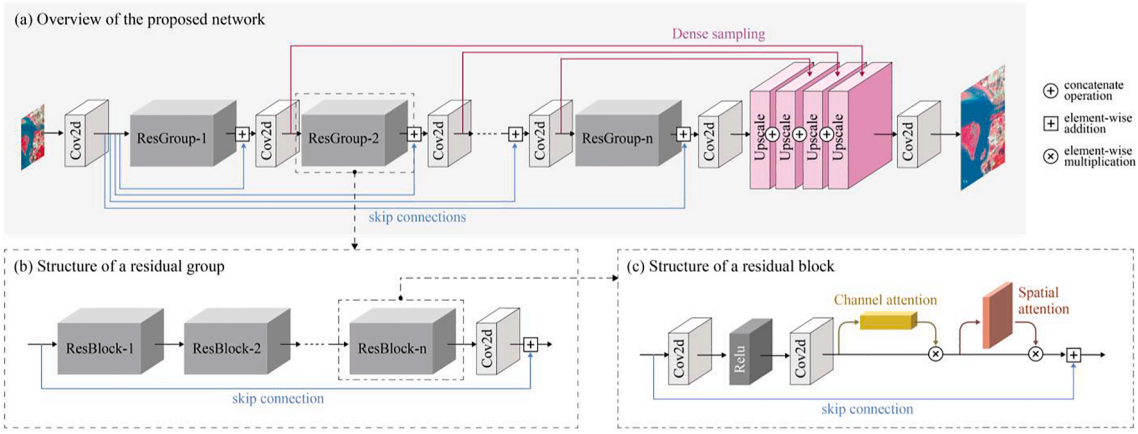

Fig. 1. The structure of the proposed method for image super-resolution. (a) overview of the proposed network. (b) structure of a residual group in the network. (c) structure of a residual block in a residual group with embedded spatial-channel attention mechanism.

network. To be more specific, the output of each residual group is upscaled to the target size, then concatenated together to form a deeper upscaled feature map before the final output layer. In this way, the lower-level features from prior convolutional groups can be deployed for reconstructing larger sizes of feature maps. The operation of the dense-sampling concatenation can be expressed as follows:

$$F_{dense} = CAT\left(f_{up}(F_{res0}), f_{up}(F_{res1}), f_{up}(F_{res2}), \cdots, f_{up}(F_{res5})\right) F^{SR}$$
$$= f_{out}^{3*3}\left(f^{3*3}(F_{dense})\right)$$

where F_{dense} denotes a feature map that is generated by a concatenation of all the upscaled outputs of residual groups. f_{up} refers to the operation of upscaling, and F_{res0} to F_{res5} are the outputs of residual groups. Then the final output F^{SR} is computed after applying two convolutional layers on F_{dense} with a filter size of 3 by 3.

2.3. Residual in residual structure

As discussed in section 1.1, residual-in-residual structures have been extensively proven to be an effective mechanism to improve the performance of a very deep neural network. Especially considering that the depth of networks can substantially influence the performance of SR models, the proposed method preserves the residual-in-residual structure adopted in RCAN, but more skip connections are added to the model.

As shown in Fig. 1, additional skip connections are set between the output of the first convolutional layer and the output of each residual group. In this manner, features extracted at various depths can be passed through hidden layers. For instance, low-level features (e.g., location information) can therefore be integrated with high-level features for enhancing the reconstruction ability of networks.

3. Data sets and experimental setup

3.1. Datasets

3.1.1. Datasets for SISR experiments

The RS images for SISR experiments were collected from Sentinel-2

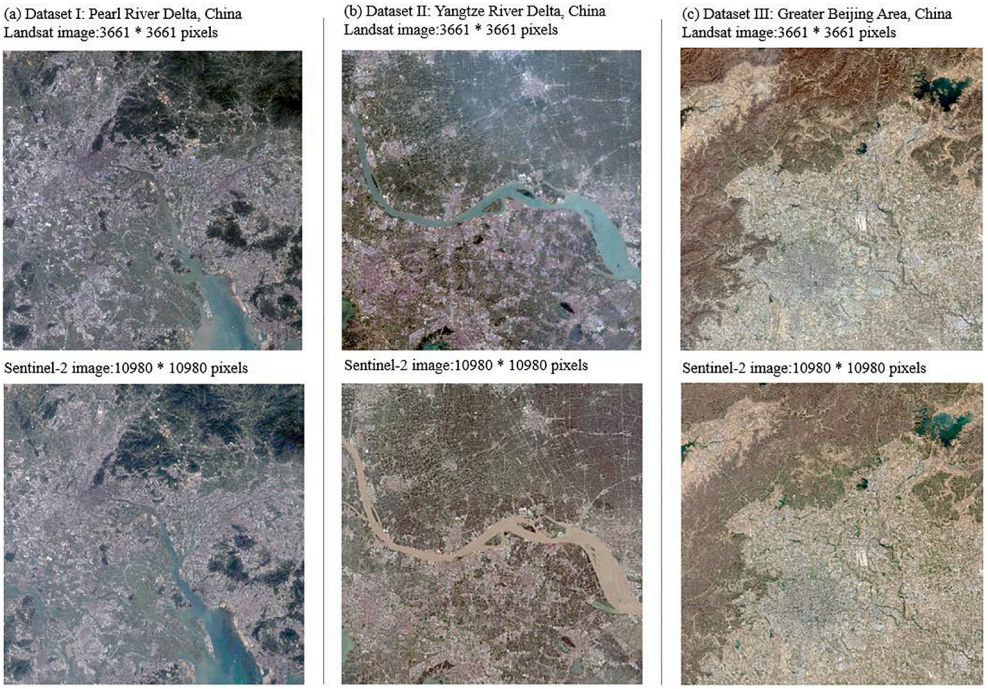

Fig. 2. Overview of two datasets (a) dataset I: Pearl River Delta. (b) dataset II: Yangtze River Delta, (c) dataset III: Greater Beijing Area.

and Landsat 8. In the experiments, 10 m resolution Sentinel-2 images were used as high-resolution targets, and 30 m resolution Landsat images were employed as low-resolution inputs. Four spectral bands were adopted in the SISR experiments, including blue, green, red, and NIR bands.

Three datasets for experiments have different geographic locations (Fig. 2). The location of the first dataset was Pearl River Delta, China, the second dataset was the Yangtze River Delta, China, and the third dataset was the Greater Beijing Area, China. All three datasets had a likewise spatial coverage of around 10,000 km^2, the pixel sizes of the collected Landsat images and Sentinel images in each dataset were 3661 by 3661 pixels and 10,980 by 10,980 pixels, respectively. These datasets consisting of paired Landsat and Sentinel-2 images were regarded as cross-sensor datasets. It is worth noting that, for training SISR models on cross-sensor datasets, the domain of Landsat inputs was transferred to the domain of the target Sentinel-2 data, meaning the considered SISR methods for the cross-sensor datasets involved the issue of domain adaptation.

Much SISR research tended to use downscaled high-resolution images as low-resolution images for model testing, such datasets can be referred to as same-sensor datasets. Following this way of generating paired datasets, an alternative approach to achieving SR Landsat is to train a SISR model based on purely Sentinel-2 data then apply the trained model on Landsat images. To compare with this alternative approach, the Sentinel-2 images in each cross-sensor dataset were downscaled to 30 m resolution as low-resolution inputs to form three same-sensor datasets.

Both cross-sensor datasets and same-sensor datasets had the same settings regarding sub-sampling images for training and validation. In each dataset, an original high-resolution image (e.g., Sentinel-2 image) was evenly cropped into small images with a size of 240 by 240 pixels, and the corresponding low-resolution image (e.g., Landsat data and down-scaled Sentinel-2 data) were cropped into small images with 80 by 80 pixels. To achieve more samples for training, each subsampled high-resolution image had an overlap of 10 pixels with its neighbouring images. In total, each dataset consisted of 2,209 cropped images, which were subsequently randomly divided into 1,988 samples for training and 221 samples for validation. It is worth mentioning that all the overlapping pixels were excluded at the stage of model validation and evaluation.

It should be noted that, although we intended to select the same acquisition dates of data from the Sentinel-2 and Landsat datasets, it was almost inevitable to have time lags between the two data sources, ranging from 10 days to four months. For dataset I, the sampled Landsat data was produced in September 2017, whereas the corresponding Sentinel-2 data was produced in November 2017. Regarding dataset II, the acquisition dates of images from these two sensors were May 2020 and September 2020, respectively. As for dataset III, the acquisition dates for Sentinel-2 and Landsat images were both in April 2020.

3.1.2. Datasets for image classification

Both single-temporal LULC classification and multi-temporal classification were conducted with SR Landsat images generated by the proposed SISR methods. The geographic location of datasets for image classification was Shenzhen, China. Shenzhen has a prevalence of urban villages, which are generally in sub-standard conditions compared with the rest of urban built-up areas. Thus, the settlements in urban villages can be regarded as “informal settlements”, whereas other built-up areas can be regarded as “formal settlements”. To test the extent to which the enhanced SR images can contribute to better LULC maps, we preferred a LULC category that is manageable with Sentinel-2 images but challenging for Landsat images. Specifically, “informal settlements” was included as a LULC class to test the performance of SR images in relatively more fine-grained image classification tasks. The datasets for single-temporal and multi-temporal classification shared the same LULC category, which consists of formal settlements, informal settlements, water, barren soil, other impervious surfaces, and vegetation.

Regarding the dataset for single-temporal LULC classification, 10 m resolution SR Landsat images with four spectral bands were adopted as input data for classification (Fig. 3). The dataset included three large images with a size of 2048 by 2048 pixels. Then one of the large images was split into two parts, in which a part of 400 by 2048 pixels was used for validation, the rest was included for training. The large images were randomly cropped into small images with 128 by 128 pixels. In total, there were 321 cropped images for training and 50 for validation.

For multi-temporal LULC classification, the original time series data were collected from Landsat 4–5 and Landsat 8 (Fig. 4), then a trained SR model was applied to these images to generate enhanced multi-temporal SR images. The dataset for multi-temporal LULC classification consisted of six time-steps with a 5-year interval, including the years 1995, 2000, 2005, 2010, 2015 and 2020. Due to the limited availability of zero cloud coverage images in the study area, the Landsat images for multi-temporal classification were collected within the ranges of the winter seasons of 1995/2000, 1999/2000, 2005/2006, 2009/2010, 2014/2015, 2019/2020. The SR image for each time-step had a size of 3600 by 3600 pixels. The multi-temporal images were croped into small patches for training and validation. In all, 179 small patches were used for training and 46 for validation. The patches used for validation were higlighted in Fig. 4 (a).

3.2. Experiment setup

Overall, the whole experiment framework can be mainly divided into two stages, image super-solution stage and image classification stage (Fig. 5).

3.2.1. Experiment setup for SISR

The first stage aimed to train a deep learning-based SR model that can improve the spatial resolution of the original Landsat images from 30 m to 10 m. Three cross-sensor datasets with different geographic locations and building morphologies were tested at the first stage. Five baseline models were compared in SISR experiments, including Bicubic, SRCNN, VDSR, EDSR, and RCAN. Moreover, to investigate the extent to which a dense-sampling module and a channel-spatial module (i.e., CBAM) could benefit the proposed method, two methods incorporated with each module were involved in the comparison with baseline models. The two additional developed methods for comparisons are termed as Residual Channel-spatial Attention Network (RCSAN) and Dense-sampling RCAN (D-RCAN) respectively. Together with the proposed method D-RCSAN, in total 8 models were tested for each dataset.

The optimizer used for model training is Adam, with the initial learning rate set as 8 × 10^-5. Each model was trained for 100 epochs. At every five epochs, the learning rate of each model was updated in the manner of multiplication with a factor of 0.8. Such a decrease in the learning rate continued during the whole training process. All the models were trained with PyTorch on an NVIDIA GeForce RTX 3090. The training time of a D-RCSAN for 100 epochs was about 20 h. Same as the loss function of its backbone structure RCAN, the loss function deployed for the proposed method is L_1 loss, which is a simple but efficient loss function that has been widely applied for SR deep neural networks. The loss between the generated SR image F^{SR} and the corresponding ground truth high-resolution image F^{HR} was calculated as below:

$$Loss_{L1} = \frac{1}{n}\sum_{i=1}^{n} \|F_i^{SR} - F_i^{HR}\|$$

3.2.2. Experiment setup for image classification

The purpose of the experiments at the second stage was to investigate the extent to which the SR images generated by the proposed SISR method can improve the performance of image classification, especially for multi-temporal LULC classification. Therefore, a series of images

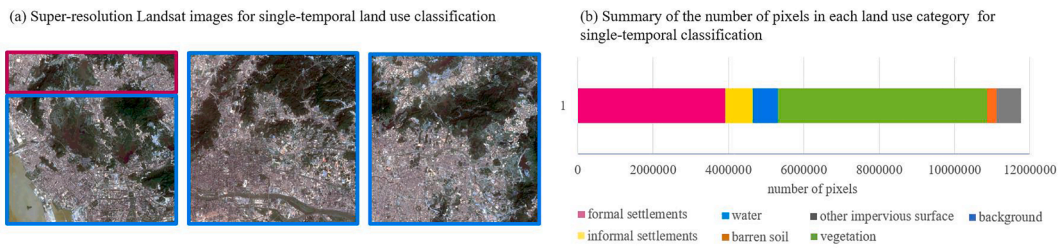

Fig. 3. Dataset settings for single-temporal land use classification tests: (a) super-resolution Landsat images (b) number of pixels for each land use category.

Fig. 4. Dataset settings for multi-temporal land use classification tests: (a) super-resolution Landsat images (b) number of pixels for each land use category.

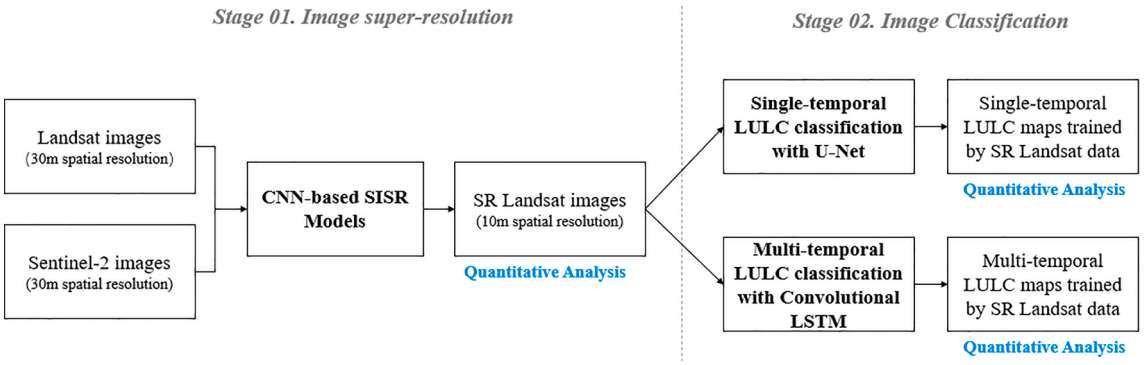

Fig. 5. Overview of the experiment setup.

classification tests were conducted to gain a comprehensive evaluation, including both single-temporal land use classification and multi-temporal land use classification. Firstly, the improved SR Landsat images were deployed to form datasets with single-temporal settings and multi-temporal settings respectively. The classification method employed for single-temporal classification was U-Net (Ronneberger et al., 2015), and the multi-temporal classification method being adopted was a UNet-Convolutional LSTM model (UNet-ConvLSTM) (Zhu et al., 2021), which is a hybrid framework proposed for multi-temporal image segmentation, it incorporates 2D Convolutional LSTM layers in an UNet-like encoder-decoder structure.

Quantitative evaluations were conducted on the classification results of each set of image classification experiments. In the test of single-temporal land use classification, original Sentinel-2 images and the images produced by SR baseline methods were deployed as a benchmark. For multi-temporal land use classification, since the temporal coverage Sentinel-2 datasets were very limited, only upscaled images produced by SR baseline methods were used as a benchmark.

Regarding the training details of single-temporal and multi-temporal land use classification, the optimizers adopted were Adam, the initial learning rates were set to $5 \times 10\hat{}$-4. Throughout the subsequent training process, the learning rates decreased with an adjusting strategy of multiplying with a factor of 0.8 when the validation loss stopped decreasing for more than 20 epochs. Both single-temporal and multi-temporal classification models were trained for 50 epochs with PyTorch on an NVIDIA GeForce RTX 3090.

3.3. Evaluation metrics

Two evaluation methods adopted for SISR experiments, structural similarity index measure (SSIM) and peak signal-to-noise ratio (PSNR), are the evaluation methods that have been widely used in SR problems.

$$PSNR = 10 \times \log_{10}\left(\frac{MAX^2}{\mathrm{MSE}}\right)$$

$$MSE = \frac{\sum_{M,N}[SR(m,n) - HR(m,n)]^2}{M \times N}$$

where M and N refer to the number of rows and columns in the image that needs to be compared, and m and n refer to the m^{th} row and the n^{th} column of the image. MAX refers to the maximum value in the image data type.

$$SSIM(sr, hr) = \frac{(2\mu_{sr}\mu_{hr} + C_1)(2\sigma_{srhr} + C_2)}{(\mu_{sr}^2 + \mu_{hr}^2 + C_1)(\sigma_{sr}^2 + \sigma_{hr}^2 + C_2)}$$

where μ_{sr} and μ_{hr} are means of images, σ_x and σ_y are the standard deviation of images, σ_{srhr} is cross-covariance for images. The default settings for C_1 and C_2 are as follows:

$$C_1 = (0.01 \times L)^2, C_2 = (0.03 \times L)^2$$

where L is the maximum pixel value of images.

As for the tests in the second stage, overall pixel accuracy (OA), per-class accuracy, and K-statistics were employed to evaluate the performance of both single-temporal and multi-temporal LULC classifications.

4. Results and discussion

4.1. Performance of the SR models trained by cross-sensor datasets

The evaluation results of the SR models trained on cross-sensor datasets were presented in Table 1. In general, the backbone structure RCAN showed superior performance than other baseline models, and the three proposed methods (RCSAN, D-RCAN and D-RCSAN) all presented different levels of improvements compared with their backbone structure RCAN among the three datasets.

Regarding the overall performance, although D-RCAN achieved the best quality of SR images in dataset I, D-RCSAN outperformed all the other models in both dataset II and dataset III.

Specifically, it can be observed that when adopting the channel-spatial attention module (i.e., RCSAN) and dense-sampling module (i.e., D-RCAN) separately, the expected improvement on the backbone model is not stable. For instance, although D-RCAN achieved the best result in dataset I, it did not yield improvement in dataset II. As for the performance of RCSAN, its improvement on SR image quality was less than D-RCAN in dataset I, and less than D-RCSAN in dataset II and III. As such, it can be argued that, for the SISR methods trained by the cross-sensor RS datasets, D-RCSAN achieved the best overall performance among all the experimental methods.

4.2. Visual comparisons of the SR images generated by cross-sensor SR models

Regarding the experiments of cross-sensor SISR, the visual comparisons of the SR images from validation datasets produced by models trained on dataset I, dataset II and dataset III were presented in Fig. 6, Fig. 7, and Fig. 8, respectively. Enlarged details were presented below the SR images, and the sampling areas of these detailed images were highlighted by yellow bounding boxes in the large pictures presented in Fig. 6, Fig. 7, and Fig. 8. The corresponding original Landsat images and the Sentinel-2 images were also included in the comparisons as raw inputs and ground truth.

Comparing with the SR images generated by baseline models, the three proposed methods generally showed better performance in reconstructing more detailed features of multi-spectral RS images. Especially in terms of the outlines of objects, the SR images of proposed methods present shaper edges and more regular shapes, such as the informal settlements in Fig. 6 and the formal buildings in Fig. 7. This effect is more pronounced in objects that can be distinctively distinguished by their spectral values, for instance, the boundaries between water bodies and vegetation, individual settlements in the field, and the outlines of roads. However, it seems that the tested models heavily rely on the intricate texture provided in the low-resolution images to construct SR information. To be more specific, in the enlarged areas of Fig. 8, the gaps between buildings in the Sentinel-2 image can hardly be detected in the original Landsat image, thus a lack of corresponding information results in the blurring and merging of building blocks in the produced SR images.

4.3. Comparisons with the SR images trained by same-sensor datasets

As discussed in 3.1.1, to compare with the alternative approach of training SR models on the same-sensor dataset then applying the trained models on Landsat images, a series of model experiments on same-sensor datasets were conducted and evaluated.

Comparing Table 2 with Table 1, the PSNR and SSIM of same-sensor models applied on Landsat data are substantially lower than their cross-sensor counterparts. This is very likely due to the circumstance that the data domain of the input and target are identical in same-sensor tests, but different in cross-sensor tests. It also can be observed that, after eventually applying trained models on Landsat images (as presented in Table 2.), D-RCAN and D-RCSAN yielded the highest PSNR and SSIM values among the three datasets. However, these scores were still lower than their cross-sensor counterparts. Arguably, the SR models on same-sensor datasets and cross-sensor datasets can be regarded as different tasks as SR models on same-sensor datasets generally do not involve domain adaptation during the model training process.

Visual comparisons of the SR Landsat images generated by same-sensor models and cross-sensor models were presented in Fig. 9. The first row showed SR Landsat images generated by same-sensor models; the second row presented the SR Landsat images constructed by cross-sensor models. It is worth noting that, before validating these trained models, a preprocessing method of histogram matching was applied on Landsat images to match their domains with Sentinel-2 images. As shown in Fig. 9, the SR Landsat images by same-sensor models generally had poorer visual representations than the SR Landsat images by cross-sensor models, the former generally failed to reconstruct details of features, including blurring edges and fuzzy objects in the images. Arguably, when the final application of a trained SISR model includes domain adaptation issues, training the model on cross-sensor datasets would achieve better overall performance than the alternative approach of training on same-sensor datasets.

Table 1
Comparison of PSNR and SSIM of super-resolution models trained on cross-sensor data in dataset I, dataset II, and dataset III.

Dataset		Bicubic	SRCNN	VDSR	EDSR	RCAN	RCSAN	D-RCAN	D-RCSAN
Dataset I	PSNR	23.6139	27.2829	27.0803	27.6560	27.7439	27.7665	**27.8029**	27.7628
	SSIM	0.93090	0.96338	0.96228	0.96612	0.96650	0.96658	**0.96689**	0.96662
Dataset II	PSNR	23.0485	28.6598	28.5889	29.1869	29.2461	29.2704	29.2206	**29.2819**
	SSIM	0.87258	0.90972	0.91000	0.92000	0.92088	0.92087	0.92050	**0.92124**
Dataset III	PSNR	23.9020	28.4166	27.9897	29.2601	29.3341	29.3465	29.3712	**29.3867**
	SSIM	0.91048	0.96352	0.96120	0.96820	0.96829	0.96837	0.96842	**0.96857**

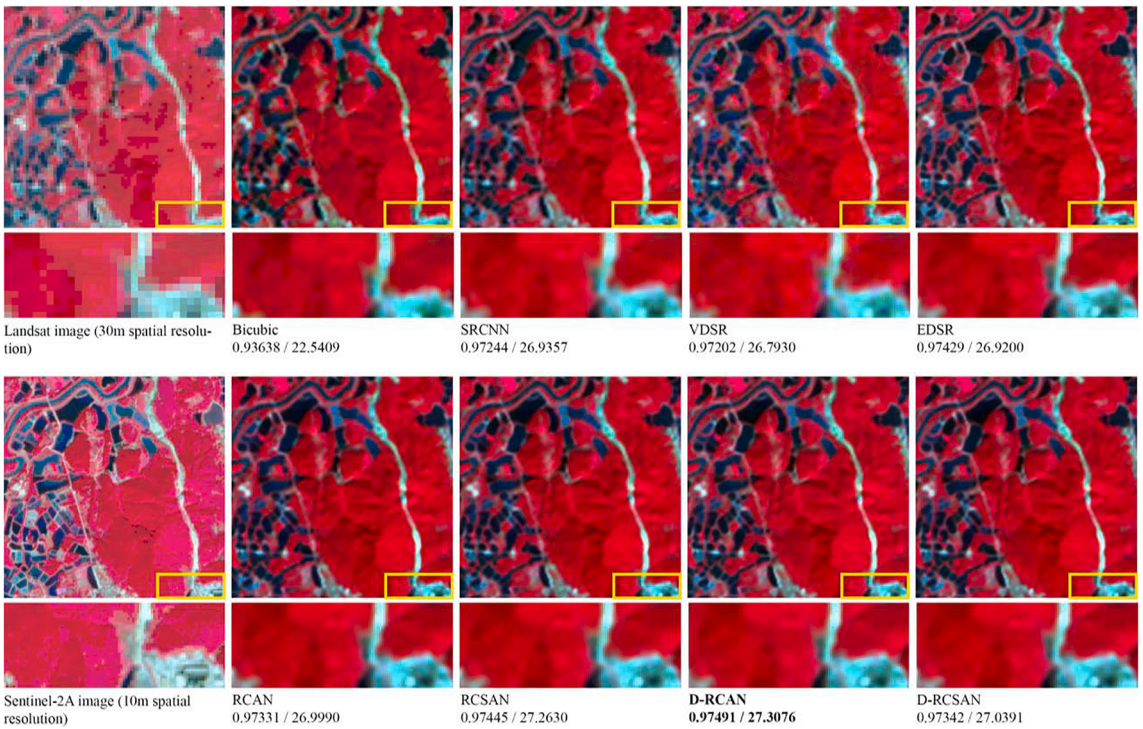

Fig. 6. Visual comparisons of false color images of dataset I.

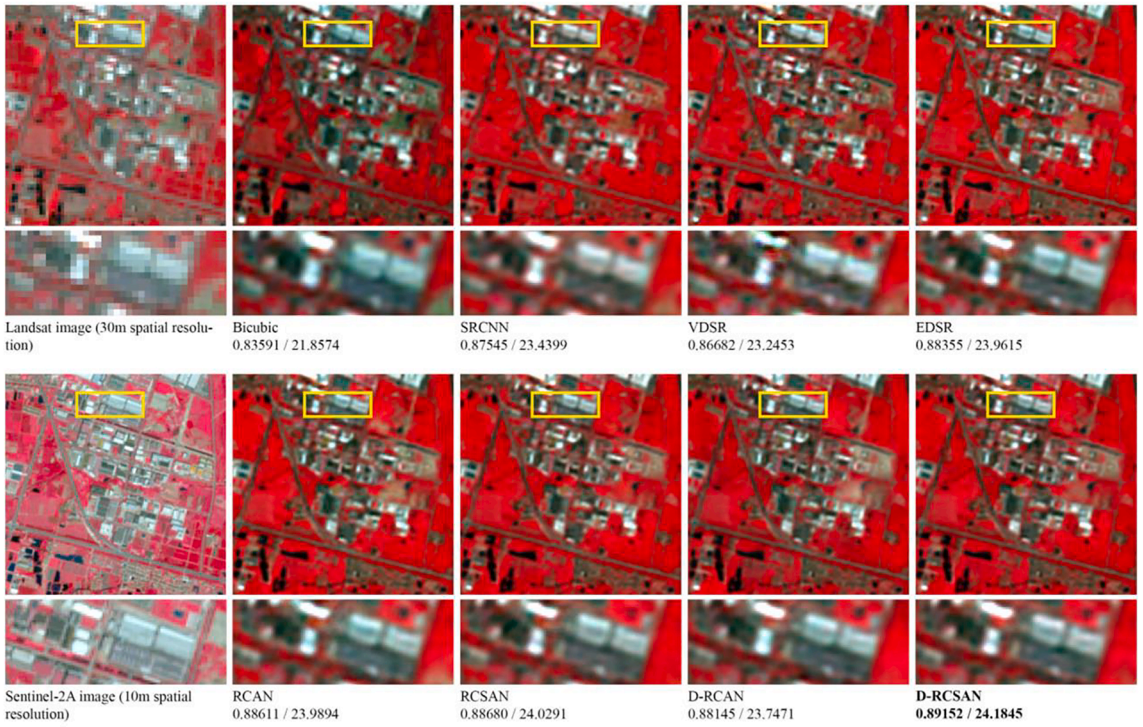

Fig. 7. Visual comparisons of false color images of dataset II.

4.4. Evaluation of SR images for land use classification

Since the D-RCSAN trained by cross-sensor datasets achieved the best overall performance regarding constructing SR Landsat images, we proceeded to the second stage experiments (i.e., single-temporal and multi-temporal LULC classification) with SR Landsat images generated by proposed SISR methods. In the evaluation of SR images for single-temporal and multi-temporal LULC classification, images upscaled by the baseline SISR methods were also deployed for measuring the effects of the proposed SISR method in terms of classification accuracy.

4.4.1. Single-temporal image classification

As can be observed in Fig. 10, in the experiments of single-temporal classification, the OA and k-statistics achieved by D-RCSAN improved

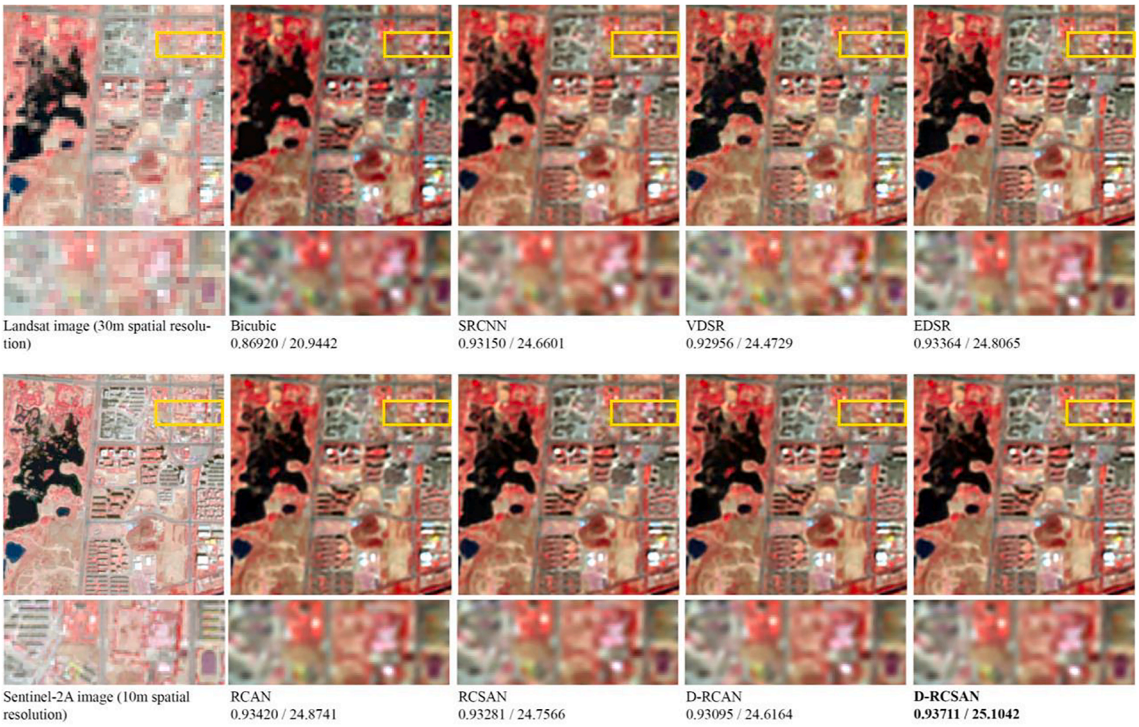

Fig. 8. Visual comparisons of false color images trained on dataset III.

Table 2
Comparison of PSNR and SSIM of same-sensor SR models applied on Landsat data in dataset I, dataset II, and dataset III.

Dataset		Bicubic	SRCNN	VDSR	EDSR	RCAN	RCSAN	D-RCAN	D-RCSAN
Dataset I	PSNR	23.6139	22.2758	22.2047	22.2074	22.2528	22.2592	**22.3157**	22.2958
	SSIM	0.93090	0.92331	0.92062	0.91941	0.91953	0.92057	**0.92034**	0.92028
Dataset II	PSNR	23.0485	22.0043	21.9310	21.9154	21.9611	21.8978	21.8772	**22.0408**
	SSIM	0.87258	0.85792	0.84511	0.84174	0.84682	0.84817	0.84900	**0.85383**
Dataset III	PSNR	23.9020	22.8085	22.6714	22.6396	22.7210	22.7050	**22.7264**	22.7060
	SSIM	0.91048	0.89368	0.88836	0.88589	0.88825	0.88922	**0.88905**	0.88854

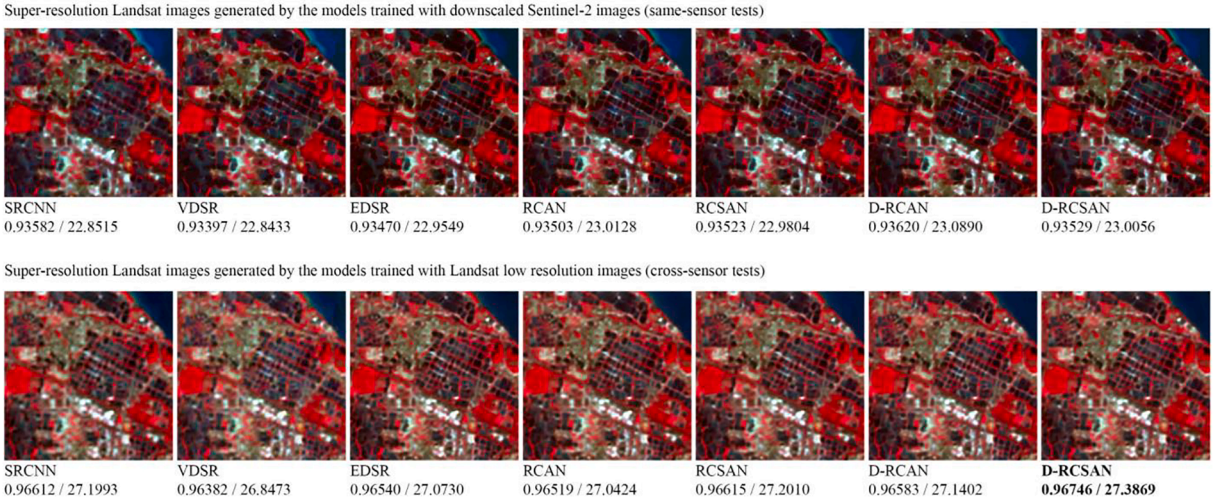

Fig. 9. Visual comparisons of SR Landsat images generated by models trained on same-sensor dataset and cross-sensor dataset.

SR images were 85.11% and 0.7589, respectively. Although the number was not as high as the performance of using Sentinel-2 images (OA 87.18%, k-statistics 0.8017), it can be regarded as a significant improvement compared with the performance of bicubic images (OA 81.60%, k-statistics 0.7151) and RCAN images (OA 84.54%, k-statistics 0.7556). Moreover, the proposed method achieved higher classification accuracy than all the baseline methods.

The results of per-class accuracy for single-temporal classification

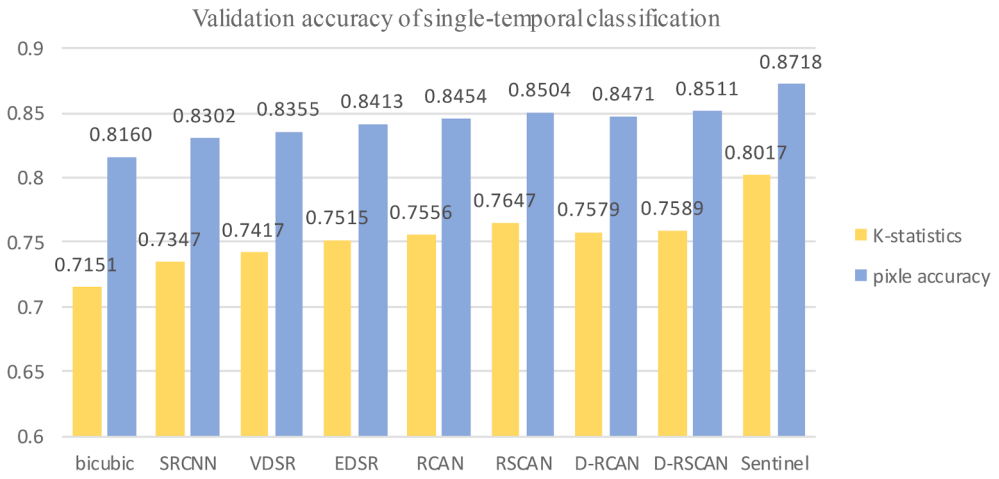

Fig. 10. Overall pixel accuracy and K-statistics of single-temporal classification by super-resolution images.

were presented with producer's accuracy and user's accuracy in Fig. 11. The variations of classification accuracy were significant in the class of informal settlements, all the tested datasets showed higher user's accuracy than producer's accuracy, which indicates more false negatives and fewer false positives. This effect of a large number of false negatives was more significant in most SR-based methods, except the proposed method D-RCSAN.

The visual representations of the single-temporal classification maps generated based on each SR method were shown in Fig. 12. It can be visually detected that the classification maps generated based on the RCSAN and D-RCSAN exhibited better performance than other baseline models. Particularly in delineating informal settlements, the proposed method achieved a very similar pattern compared with Sentinel-2.

4.4.2. *Multi-temporal image classification*

In the experiments of multi-temporal LULC classification, classification maps were produced based on all the baseline and proposed SISR models. The sentinel-2 images were not included in multi-temporal classification due to the circumstance that sentinel-2 can only provide images after the year 2015. The overall pixel accuracy and k-statistics of multitemporal classification maps generated based on D-RCSAN were 83.87% and 0.7118 respectively, which were the highest scores in the comparison (Fig. 13).

The comparison of per-class accuracy can be seen in Fig. 14. In general, the proposed SR method led to higher producer's accuracy and user's accuracy than all the other SR images. Compared with bicubic and RCAN, D-RCSAN achieved substantially better overall performance in the classes of other impervious surfaces and barren soil.

The visual comparisons of multi-temporal land use classification were presented in Fig. 15. It can be observed that the multi-temporal LULC maps improved substantially after adopting effective SR methods, including EDSR, RCAN, as well as the proposed SR methods. As can be observed in the visual comparison, the images enhanced by D-RSCAN led to better performance in the delineation of roads and formal settlements, as well as the recognition of informal settlements.

Furthermore, considering that the potential applications of multi-temporal land-use classification maps mainly include change detection or change prediction, therefore it is critical that the tendency of changes in each land-use class was effectively captured in multi-temporal classification. In this

case, the number of pixels in each land-use class over six-time steps were mapped in Fig. 16. It can be observed that, compared with ground truth, the proposed D-RSCAN showed better performance in capturing the consistent changing trend of the number of pixels in the class of formal settlements. For instance, the ground-truth number of pixels in the class of formal settlements increased 11.04% from 2015 to 2020, D-RSCAN simulated an increase of 6.93%, whereas RCAN projected a decrease of 2.76%.

5. Conclusions

In general, this paper offers a new prospect to improve the quality of multi-temporal LULC classification maps by taking advantage of the newly produced RS images for enhancing the resolution of historic RS imagery. A novel framework was developed for improving multi-temporal LULC classification through a proposed CNN-based SISR

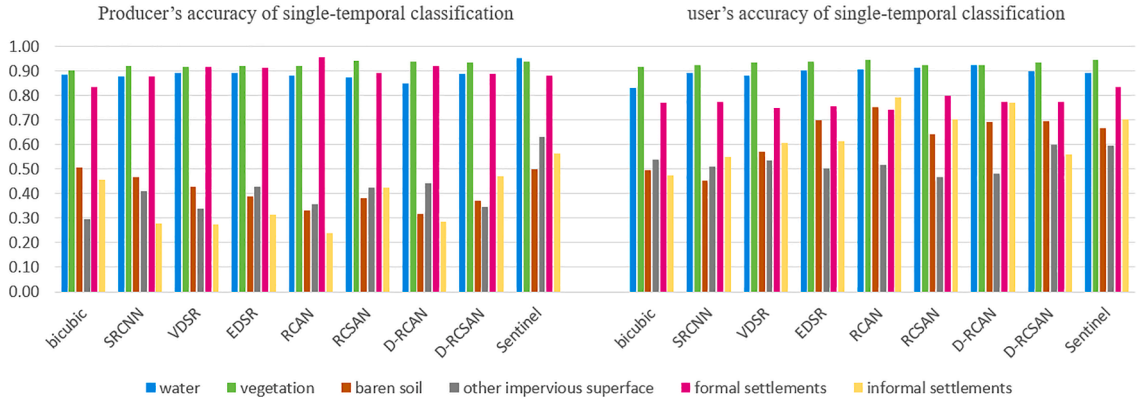

Fig. 11. Per-class accuracy of single-temporal land use classification.

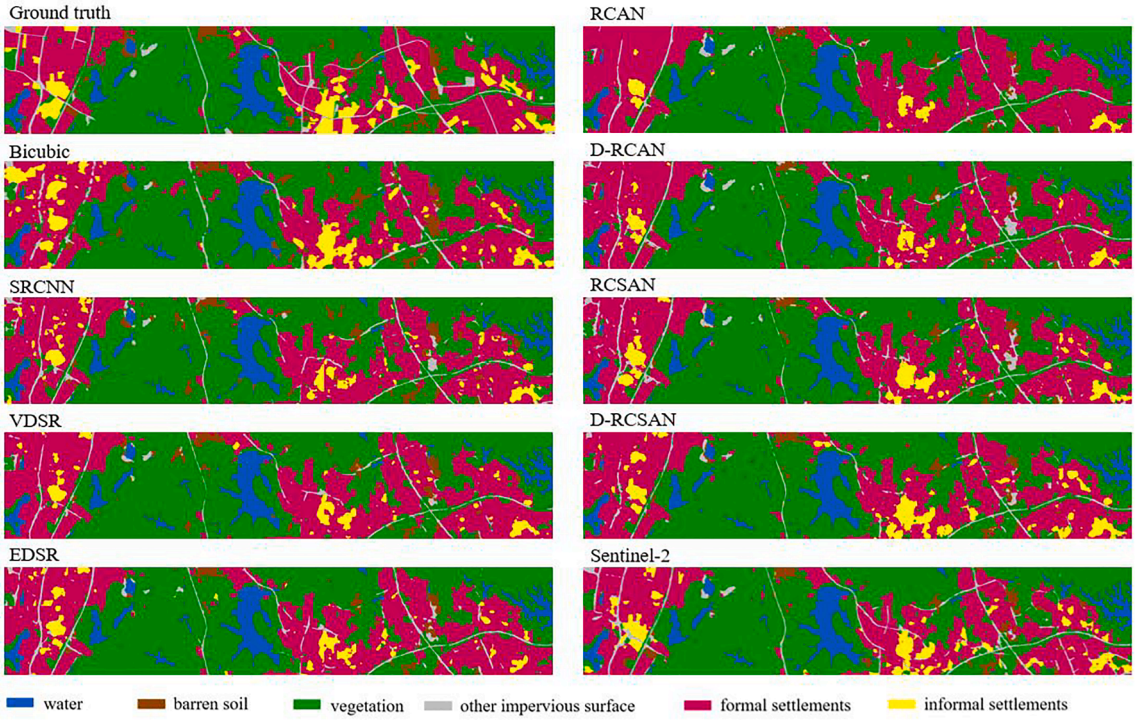

Fig. 12. Visual comparison of the land use maps produced by single-temporal classification.

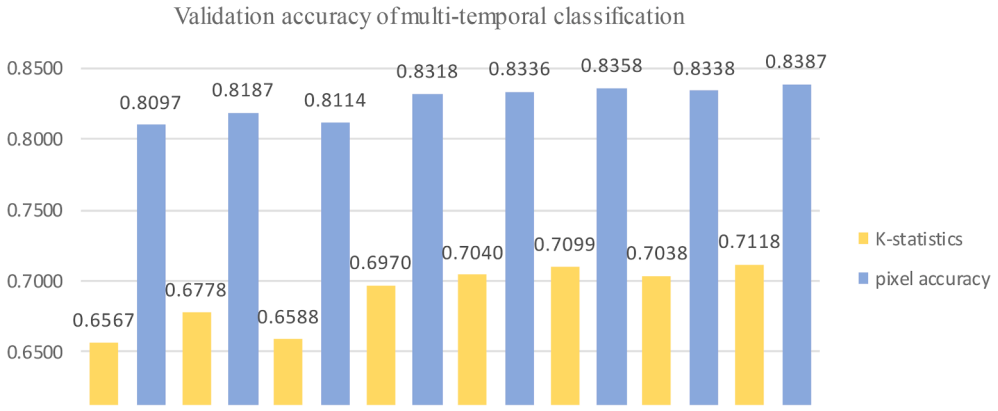

Fig. 13. Overall pixel accuracy and K-statistics of multi-temporal classification by super-resolution images.

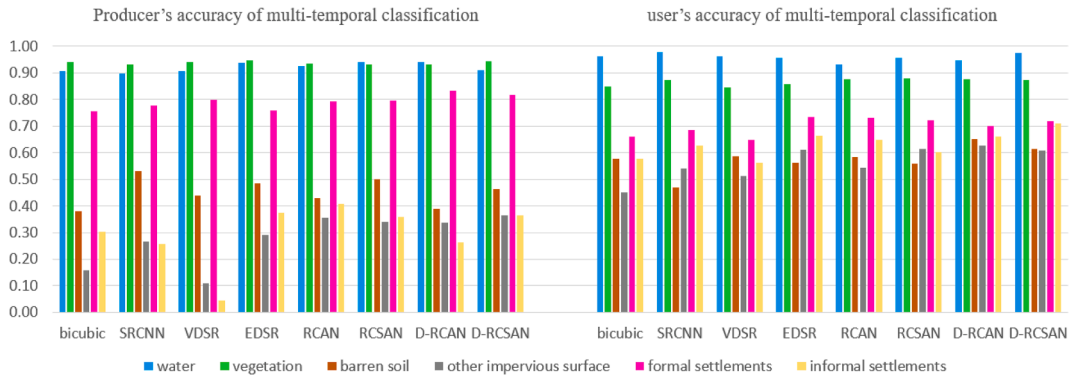

Fig. 14. Per-class accuracy of multi-temporal land use classification.

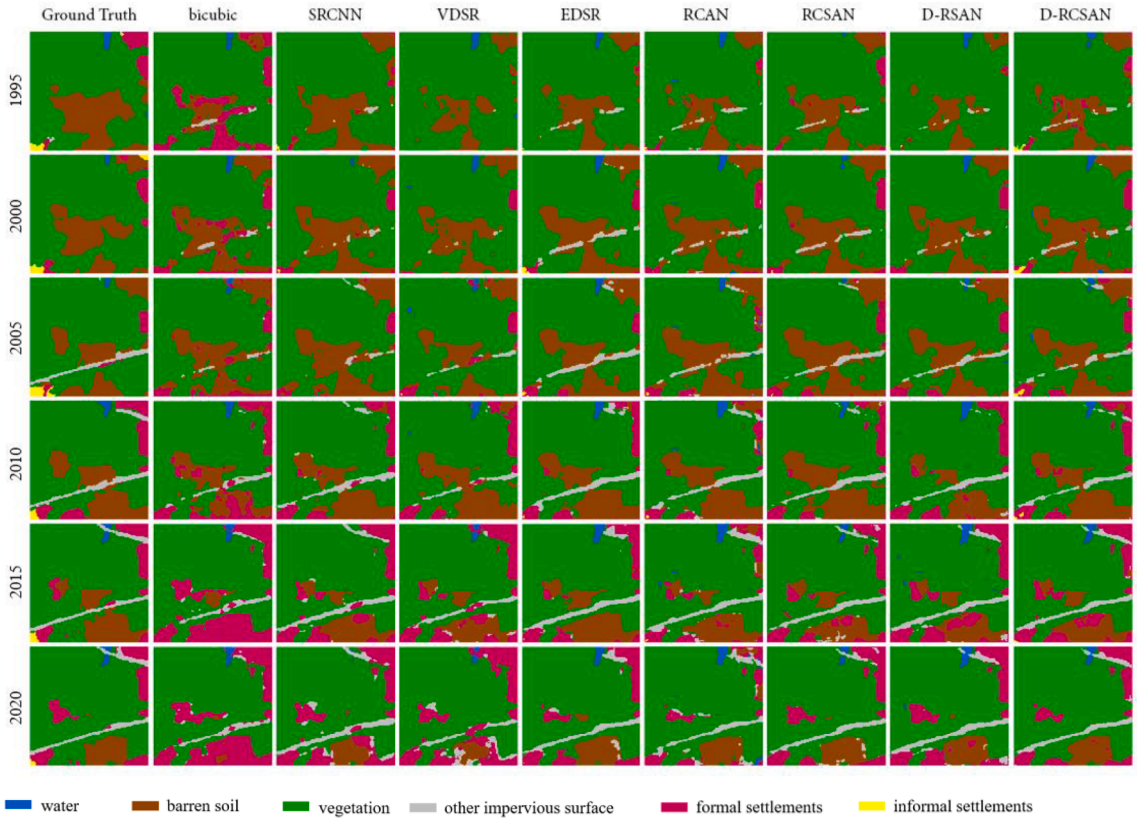

Fig. 15. Visual comparisons of predictions of multi-temporal land use classification with super-resolution images.

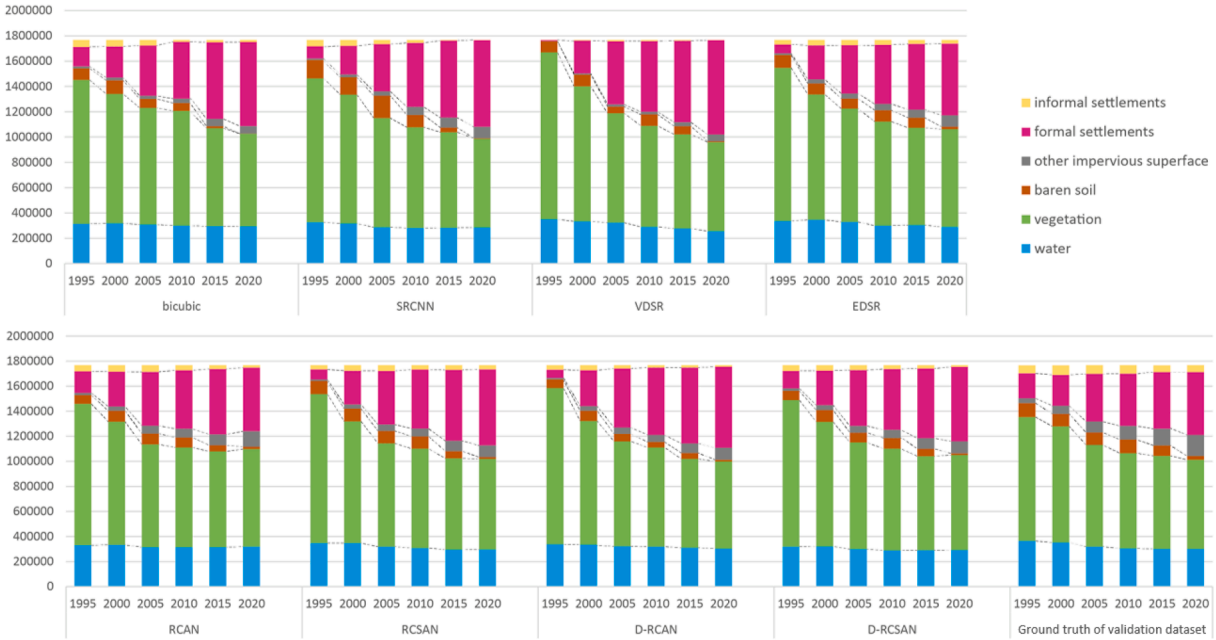

Fig. 16. Number of pixels in each land-use class of multi-temporal classification results.

method.

The research was conducted in two stages: image SR preprocessing and LULC classification. At the first stage, we proposed a SISR method that incorporated a channel-spatial attention module and a dense-sampling module based on an RCAN structure. The results of experiments with baseline SISR methods proved that the two extension modules can both bring improvements to the performance of the proposed SR model. By adopting the proposed SR method, the spatial resolution of multispectral Landsat images (four spectral bands) was significantly enhanced based on multi-spectral Sentinel-2 images. At the second stage, the SR Landsat image generated by the proposed SR method substantially elevated the accuracy of both single-temporal and multi-temporal classification, and the discriminative ability of trained models was distinctively improved.

Furthermore, the proposed framework not only developed an enhanced SR model for better image classification, but also practically highlighted and verified the likelihood of making use of a wealth of previously untapped low-resolution imagery for purposes of urban growth observation, urban planning, or even vulnerability assessment review.

CRediT authorship contribution statement

Yue Zhu: Conceptualization, Methodology, Software, Data analysis, Writing – original draft, Writing – review & editing. **Christian Geiß:** Conceptualization, Writing – review & editing. **Emily So:** Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Altaf, B., Yu, L., Zhang, X., 2018. Spatio-Temporal Attention based Recurrent Neural Network for Next Location Prediction. IEEE International Conference on Big Data (Big Data) 2018, 937–942. https://doi.org/10.1109/BigData.2018.8622218.

Chen, J., Wan, L., Zhu, J., Xu, G., Deng, M., 2020a. Multi-Scale Spatial and Channel-wise Attention for Improving Object Detection in Remote Sensing Imagery. IEEE Geoscience and Remote Sensing Letters 17 (4), 681–685. https://doi.org/10.1109/LGRS.2019.2930462.

Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.-S., 2017. SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017, 6298–6306. https://doi.org/10.1109/CVPR.2017.667.

Chen, M., Ke, Y., Bai, J., Li, P., Lyu, M., Gong, Z., Zhou, D., 2020b. Monitoring early stage invasion of exotic Spartina alterniflora using deep-learning super-resolution techniques based on multisource high-resolution satellite imagery: A case study in the Yellow River Delta, China. International Journal of Applied Earth Observation and Geoinformation 92, 102180. https://doi.org/10.1016/j.jag.2020.102180.

Dong, C., Loy, C. C., He, K., & Tang, X. (2015). Image Super-Resolution Using Deep Convolutional Networks. ArXiv:1501.00092 [Cs]. http://arxiv.org/abs/1501.00092.

Dong, C., Loy, C. C., & Tang, X. (2016). Accelerating the Super-Resolution Convolutional Neural Network. ArXiv:1608.00367 [Cs]. http://arxiv.org/abs/1608.00367.

Dong, X., Sun, X., Jia, X., Xi, Z., Gao, L., Zhang, B., 2020. Remote Sensing Image Super-Resolution Using Novel Dense-Sampling Networks. IEEE Transactions on Geoscience and Remote Sensing 1–16. https://doi.org/10.1109/TGRS.2020.2994253.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019). Dual Attention Network for Scene Segmentation. ArXiv:1809.02983 [Cs]. http://arxiv.org/abs/1809.02983.

Gao, L., Hong, D., Yao, J., Zhang, B., Gamba, P., Chanussot, J., 2021. Spectral Superresolution of Multispectral Imagery With Joint Sparse and Low-Rank Learning. IEEE Transactions on Geoscience and Remote Sensing 59 (3), 2269–2280. https://doi.org/10.1109/TGRS.2020.3000684.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. ArXiv:1406.2661 [Cs, Stat]. http://arxiv.org/abs/1406.2661.

Haut, J.M., Fernandez-Beltran, R., Paoletti, M.E., Plaza, J., Plaza, A., 2019. Remote Sensing Image Superresolution Using Deep Residual Channel Attention. IEEE Transactions on Geoscience and Remote Sensing 57 (11), 9277–9289. https://doi.org/10.1109/TGRS.2019.2924818.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. ArXiv:1512.03385 [Cs]. http://arxiv.org/abs/1512.03385.

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2019). Squeeze-and-Excitation Networks. ArXiv:1709.01507 [Cs]. http://arxiv.org/abs/1709.01507.

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2018). Densely Connected Convolutional Networks. ArXiv:1608.06993 [Cs]. http://arxiv.org/abs/1608.06993.

Kim, J., Lee, J.K., Lee, K.M., 2016. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016, 1646–1654. https://doi.org/10.1109/CVPR.2016.182.

Kodali, N., Abernethy, J., Hays, J., & Kira, Z. (2017). On Convergence and Stability of GANs. ArXiv:1705.07215 [Cs]. http://arxiv.org/abs/1705.07215.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., & Shi, W. (2017). Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. ArXiv:1609.04802 [Cs, Stat]. http://arxiv.org/abs/1609.04802.

Lim, B., Son, S., Kim, H., Nah, S., & Lee, K. M. (2017). Enhanced Deep Residual Networks for Single Image Super-Resolution. ArXiv:1707.02921 [Cs]. http://arxiv.org/abs/1707.02921.

McGlinchy, J., Johnson, B., Muller, B., Joseph, M., Diaz, J., 2019. Application of UNet Fully Convolutional Neural Network to Impervious Surface Segmentation in Urban Environment from High Resolution Satellite Imagery. In: IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium, pp. 3915–3918. https://doi.org/10.1109/IGARSS.2019.8900453.

Muqeet, A., Iqbal, M.T.B., Bae, S., 2019. HRAN: Hybrid Residual Attention Network for Single Image Super-Resolution. IEEE Access 7, 137020–137029. https://doi.org/10.1109/ACCESS.2019.2942346.

Panboonyuen, T., Jitkajornwanich, K., Lawawirojwong, S., Srestasathiern, P., Vateekul, P., 2019. Semantic Segmentation on Remotely Sensed Images Using an Enhanced Global Convolutional Network with Channel Attention and Domain Specific Transfer Learning. Remote Sensing 11 (1), 83. https://doi.org/10.3390/rs11010083.

Park, J., Woo, S., Lee, J.-Y., & Kweon, I. S. (2018). BAM: Bottleneck Attention Module. ArXiv:1807.06514 [Cs]. http://arxiv.org/abs/1807.06514.

Pasquali, G., Iannelli, G.C., Dell'Acqua, F., 2019. Building Footprint Extraction from Multispectral, Spaceborne Earth Observation Datasets Using a Structurally Optimized U-Net Convolutional Neural Network. Remote Sensing 11 (23), 2803. https://doi.org/10.3390/rs11232803.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. ArXiv:1505.04597 [Cs]. http://arxiv.org/abs/1505.04597.

Rußwurm, M., Körner, M., 2018. Multi-Temporal Land Cover Classification with Sequential Recurrent Encoders. ISPRS International Journal of Geo-Information 7 (4), 129. https://doi.org/10.3390/ijgi7040129.

Schuegraf, P., Bittner, K., 2019. Automatic Building Footprint Extraction from Multi-Resolution Remote Sensing Images Using a Hybrid FCN. ISPRS International Journal of Geo-Information 8 (4), 191. https://doi.org/10.3390/ijgi8040191.

Shi, W., Du, H., Mei, W., Ma, Z., 2020. (SARN)spatial-wise attention residual network for image super-resolution. The Visual Computer. https://doi.org/10.1007/s00371-020-01903-8.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W., & WOO, W. (2015). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), Advances in Neural Information Processing Systems 28 (pp. 802–810). Curran Associates, Inc. http://papers.nips.cc/paper/5955-convolutional-lstm-network-a-machine-learning-approach-for-precipitation-nowcasting.pdf.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015, 1–9. https://doi.org/10.1109/CVPR.2015.7298594.

Teimouri, N., Dyrmann, M., Jørgensen, R.N., 2019. A Novel Spatio-Temporal FCN-LSTM Network for Recognizing Various Crop Types Using Multi-Temporal Radar Images. Remote Sensing 11 (8), 990. https://doi.org/10.3390/rs11080990.

Tong, T., Li, G., Liu, X., Gao, Q., 2017. Image Super-Resolution Using Dense Skip Connections. IEEE International Conference on Computer Vision (ICCV) 2017, 4809–4817. https://doi.org/10.1109/ICCV.2017.514.

Tong, W., Chen, W., Han, W., Li, X., Wang, L., 2020. Channel-Attention-Based DenseNet Network for Remote Sensing Image Scene Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13, 4121–4132. https://doi.org/10.1109/JSTARS.2020.3009352.

Tran, D. T., Iosifidis, A., Kanniainen, J., & Gabbouj, M. (2017). Temporal Attention augmented Bilinear Network for Financial Time-Series Data Analysis. ArXiv:1712.00975 [Cs, q-Fin]. Doi: 10.1109/TNNLS.2018.2869225.

Vuolo, F., Neuwirth, M., Immitzer, M., Atzberger, C., Ng, W.-T., 2018. How much does multi-temporal Sentinel-2 data improve crop type classification? International Journal of Applied Earth Observation and Geoinformation 72, 122–130. https://doi.org/10.1016/j.jag.2018.06.007.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X., 2017. Residual Attention Network for Image Classification. ArXiv:1704.06904 [Cs]. http://arxiv.org/abs/1704.06904.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q., 2020. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. ArXiv:1910.03151 [Cs]. http://arxiv.org/abs/1910.03151.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C. C., Qiao, Y., & Tang, X. (2018). ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. ArXiv:1809.00219 [Cs]. http://arxiv.org/abs/1809.00219.

Wen, R., Fu, K., Sun, H., Sun, X., Wang, L., 2018. Image Superresolution Using Densely Connected Residual Networks. IEEE Signal Processing Letters 25 (10), 1565–1569. https://doi.org/10.1109/LSP.2018.2861989.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. ArXiv:1807.06521 [Cs]. http://arxiv.org/abs/1807.06521.

Wu, H., Zou, Z., Gui, J., Zeng, W.-J., Ye, J., Zhang, J., Liu, H., Wei, Z., 2021. Multi-Grained Attention Networks for Single Image Super-Resolution. IEEE Transactions on Circuits and Systems for Video Technology 31 (2), 512–522. https://doi.org/10.1109/TCSVT.2020.2988895.

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated Residual Transformations for Deep Neural Networks. ArXiv:1611.05431 [Cs]. http://arxiv.org/abs/1611.05431.

Xu, X., Li, X., 2019. SCAN: Spatial Color Attention Networks for Real Single Image Super-Resolution. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2019, 2024–2032. https://doi.org/10.1109/CVPRW.2019.00254.

Yang, X., Li, X., Ye, Y., Lau, R.Y.K., Zhang, X., Huang, X., 2019. Road Detection and Centerline Extraction Via Deep Recurrent Convolutional Neural Network U-Net. IEEE Transactions on Geoscience and Remote Sensing 57 (9), 7209–7220. https://doi.org/10.1109/TGRS.2019.2912301.

Yao, J., Hong, D., Chanussot, J., Meng, D., Zhu, X., & Xu, Z. (2020). Cross-Attention in Coupled Unmixing Nets for Unsupervised Hyperspectral Super-Resolution. ArXiv: 2007.05230 [Cs, Eess]. http://arxiv.org/abs/2007.05230.

Yeom, J.-M., Deo, R.C., Adamowski, J.F., Park, S., Lee, C.-S., 2020. Spatial mapping of short-term solar radiation prediction incorporating geostationary satellite images coupled with deep convolutional LSTM networks for South Korea. Environmental Research Letters 15 (9), 094025. https://doi.org/10.1088/1748-9326/ab9467.

Zagoruyko, S., & Komodakis, N. (2017). Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. ArXiv: 1612.03928 [Cs]. http://arxiv.org/abs/1612.03928.

Zhang, S., Yuan, Q., Li, J., Sun, J., Zhang, X., 2020. Scene-Adaptive Remote Sensing Image Super-Resolution Using a Multiscale Attention Network. IEEE Transactions on Geoscience and Remote Sensing 58 (7), 4764–4779. https://doi.org/10.1109/TGRS.2020.2966805.

Zhang, W., Li, J., Hua, Z., 2021. Attention-Based Tri-UNet for Remote Sensing Image Pan-Sharpening. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14, 3719–3732. https://doi.org/10.1109/JSTARS.2021.3068274.

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., & Fu, Y. (2018). Image Super-Resolution Using Very Deep Residual Channel Attention Networks. ArXiv:1807.02758 [Cs]. http://arxiv.org/abs/1807.02758.

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., & Fu, Y. (2018). Residual Dense Network for Image Super-Resolution. ArXiv:1802.08797 [Cs]. http://arxiv.org/abs/1802.08797.

Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C. C., Lin, D., & Jia, J. (2018). PSANet: Point-wise Spatial Attention Network for Scene Parsing. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), Computer Vision – ECCV 2018 (Vol. 11213, pp. 270–286). Springer International Publishing. Doi: 10.1007/978-3-030-01240-3_17.

Zheng, K., Gao, L., Liao, W., Hong, D., Zhang, B., Cui, X., Chanussot, J., 2021. Coupled Convolutional Neural Network With Adaptive Response Function Learning for Unsupervised Hyperspectral Super Resolution. IEEE Transactions on Geoscience and Remote Sensing 59 (3), 2487–2502. https://doi.org/10.1109/TGRS.2020.3006534.

Zheng, K., Gao, L., Ran, Q., Cui, X., Zhang, B., Liao, W., Jia, S., 2019. Separable-spectral convolution and inception network for hyperspectral image super-resolution. International Journal of Machine Learning and Cybernetics 10 (10), 2593–2607. https://doi.org/10.1007/s13042-018-00911-4.

Zhu, Y., Gei, C., So, E., Jin, Y., 2021. Multi-temporal Relearning with Convolutional LSTM models for Land Use Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *PP* 1–1. https://doi.org/10.1109/JSTARS.2021.3055784.