3DRRDB: Super Resolution of Multiple Remote Sensing Images using 3D Residual in Residual Dense Blocks

Mohamed Ramzy Ibrahim[1,2,3], Robert Benavente[2,3], Felipe Lumbreras[2,3], Daniel Ponsa[2,3]
[1]Computer Engineering Department, Arab Academy for Science & Technology, Alexandria, Egypt
[2]Department of Computer Science, Universitat Autònoma de Barcelona, Barcelona, Spain
[3]Computer Vision Center, Campus UAB, Barcelona, Spain
m.ramzy@aast.edu, {mramzy, robert, felipe, daniel}@cvc.uab.es

Abstract

The rapid advancement of Deep Convolutional Neural Networks helped in solving many remote sensing problems, especially the problems of super-resolution. However, most state-of-the-art methods focus more on Single Image Super-Resolution neglecting Multi-Image Super-Resolution. In this work, a new proposed 3D Residual in Residual Dense Blocks model (3DRRDB) focuses on remote sensing Multi-Image Super-Resolution for two different single spectral bands. The proposed 3DRRDB model explores the idea of 3D convolution layers in deeply connected Dense Blocks and the effect of local and global residual connections with residual scaling in Multi-Image Super-Resolution. The model tested on the Proba-V challenge dataset shows a significant improvement above the current state-of-the-art models scoring a Corrected Peak Signal to Noise Ratio (cPSNR) of 48.79 dB and 50.83 dB for Near Infrared (NIR) and RED Bands respectively. Moreover, the proposed 3DRRDB model scores a Corrected Structural Similarity Index Measure (cSSIM) of 0.9865 and 0.9909 for NIR and RED bands respectively.

1. Introduction

Super-resolution (SR) is a task that refers to improving imaging systems resolution. The goal of SR is to generate a high-resolution (HR) image from single or multiple low-resolution (LR) images [1]. In general, SR techniques are a solution when the resolution of available images is not sufficient for the application at hand. This is the case of remote sensing applications, where capturing high-resolution images from satellites is difficult due to constraints such as sensor limitations or exorbitant acquisition expenses [2].

Nowadays, small satellite missions centered on data collection have relatively expensive hardware with higher spatial and spectral resolutions. Consequently, as resolutions improve, onboard equipment on satellites creates larger amounts of data, making compression algorithms increasingly difficult to fulfill bandwidth constraints [3-5]. A possible solution to this problem could be to acquire images at a lower resolution and use SR algorithms for enhancing and reconstructing HR images from LR images captured by sensors [1, 2], and indeed, different SR techniques and models are used in numerous applications of remote sensing for monitoring and mapping Earth's surface [1].

Earliest methods [6-10] were based on reconstructing a HR version of the scene using a single image as input (single-image super-resolution, SISR). However, the quantity of information that a single image can convey is limited. Furthermore, the problems due to the atmospheric conditions (clouds, sun illumination, water, and others), altitude, errors in the sensors or noise, can affect badly the satellite images and cause a loss of data in the images that hinder the performance of SISR algorithms [11].

To overcome such problems, multiple-image super-resolution (MISR), on the other hand, constructs a HR image from multiple LR images of the same scene taken from the same or separate sensors, usually at different times. MISR has a substantial benefit from SISR in that it can extract previously inaccessible information from several image observations of the same scene. In the case of remote sensing applications, MISR can obtain missing information (e.g., data occluded by clouds) from the rest of the input images of the scene.

In recent years, different deep learning (DL) methods were proposed for SISR achieving considerable progress in the field. In particular, Y. Zhang et al. [12] used dense networks and K. Zhang et al. [13] exploited the use of residual connections at different levels to solve the SISR problem. On the other hand, MISR has received much less attention. Only recently, some approaches have been proposed, but existing methods do not completely exploit the relationships existing between the input images.

In this paper, we propose a novel DL model to generate a HR image from multiple low-resolution images. Our proposal is based on the combination of dense networks and residual connections, which have been successfully applied for SISR previously, and the use of 3D convolutions [14] to take advantage of the existing correlations between pixels of the same point of the scene in the different input images. Hence, the main contributions of the model can be summarized as,

- The use of 3D convolutions to focus on correlation between multiple LR images for each scene.
- Dense blocks that stack large amounts of 3D feature maps for better generation of SR by establishing maximum information flow between blocks.
- Fusion of global and local residual connections with residual scaling that solves the problem of vanishing gradient.

2. Related Work

In the last decades, different work was provided for both approaches of SISR and MISR. In this section, we review different SISR models followed by recent advances in MISR.

2.1. SISR Techniques

SISR is a technique for reconstructing a HR image from a single LR image. Different SISR techniques have been introduced in the literature. Interpolation-based methods such as Lanczos filters [6], optimization-based methods, and learning-based methods are the three primary categories of SISR techniques [1]. Low total-variation priors [7], gradient-profile priors [15], and non-local similarity [8] are examples of optimization-based algorithms that specifically convey prior knowledge about real images in order to address this ill-posed inverse issue. Prior knowledge narrows the solution space, resulting in higher-quality results [1]. However, once the upscaling factor is increased, many optimization-based methods' performance rapidly degrades, and these methods are typically computationally expensive.

Pixel-based or example-based learning approaches are both learning-based methods. The latter are the most common, and they generate HR patches by modeling the relationship between LR and HR patches. Different machine learning techniques were used at the beginning of learning-based methods such as Freeman et al. [9] that used the k-nearest neighbor algorithm and Schulter et al. [10] that used random forests on LR-HR image pairs to generate a SR image.

After that, DL methods gained popularity and were used in SISR tasks. Dong et al. [16] proposed a Super-Resolution Convolutiosn (SRCNN) that was used to learn an end-to-end mapping between LR and HR image pairs. Moreover, Multi-scale ResNet CNN architecture was adopted by Li et al. [17] for Image SR. Tai et al. [18] proposed a Deep Recursive Residual Convolution Neural Network (DRRCNN) with 52 convolution layers with residual connections to avoid problems of deep CNNs and also proposed a memory network (MemNet) that is based on a recursive unit and a gate unit for explicitly mining persistent memory via an adaptive learning process [19]. Under different receptive fields, the recursive unit learns multi-level representations of the current state. The preceding memory blocks' representations and outputs are combined and passed to the gate unit, which selects how much to reserve from the previous state and how much should be updated by the current state [19].

A SR Generative Adversarial Network (SRGAN) was proposed by Ledig et al. [20]. Sajjadi et al. [21] proposed a GAN model that focuses on combining perceptual loss and automatic texture synthesis to create realistic textures without focusing on correct ground truth pixel generation.

The common weak point of the previous SISR technique is that it only depends on one LR for each scene to get SR image, which can result in missing different inaccessible information that affects the quality of the enhanced image.

2.2. MISR Techniques

MISR technique elucidated the importance of feature fusion from several LR images of the same scene to generate a HR image. It is easier to produce a more accurate reconstruction than with SISR approaches since more data from many observations of the scene is available. Tsai and Huang [22] implied the first work of MISR techniques which is based on utilizing a frequency-domain approach to combine multiple images with subpixel displacement to enhance image spatial resolution. The first proposed method had several flaws in terms of fusion of information from HR images. Therefore, different spatial domain MISR strategies were introduced.

Some of the most proposed popular spatial domain MISR strategies were iterative back-projection (IBP) [23], projection onto convex sets (POCS) [24], non-uniform interpolation [25], regularized methods [26], and sparse coding [27]. Most of the previous MISR methods needed a priori knowledge of the motion model, noise level, and blur kernel where a processing step of image registration and blur identification must be done as a reconstruction preprocessing stage [1]. However, knowing the image degradation process or properly forecasting it might be difficult in different situations. As a result, numerous investigations on blind SR image reconstruction have been conducted.

Different DL approaches had been adopted in the last previous year in the MISR strategies for video SR [28, 29]. On the other hand, MISR is addressed rarely in remote sensing satellite imagery [2]. A model proposed by Kawulok et al. [30] does not fully harness the benefits of DL, limiting their CNN's ability to tackle a SISR problem. The median shift-and-add approach is used to fuse the upsampled LR pictures, resulting in an SR image that is utilized as an initial guess for a classic regularized procedure [1].

Recently, The European Space Agency (ESA) issued a challenge to super-resolved multiple PROBA-V satellite imagery [31, 32]. Deudon et al. [33] introduced HighRes-net, a DL-based system for MISR of remotely sensed

PROBA-V satellite data. Exploiting both spatial and temporal connections, an end-to-end learning strategy was developed. They suggested an end-to-end learning process for the MISR sub-tasks of co-registration, fusion, upsampling, and registration-at-the-loss [2]. Molini et al. recently[1, 34] presented DeepSUM and DeepSUM++, two new CNNs for super resolving multi-temporal PROBA-V imaging. Salvetti et al. [2] proposed RAMS model that uses feature attention at various phases.

The previous models did not focus on the importance of the fusion of 3D Convolutions, Deep Dense Blocks, Global, Local Residual connections and residual scaling in MISR which are exploited in our proposed model.

3. Methods

In this section, we propose the 3DRRDB model to extract multiple features from multiple low-resolution satellite images of the same scene to generate a high-resolution image. The model adopts the use of dense blocks [35] that have been successfully used in residual dense networks [36] for single-image super-resolution. The main novelty of the model is replacing the 2D convolutions with 3D convolutions to capture the correlated information of the multiple low-resolution input images of the same scene. Moreover, the model introduces the importance of fusion of global and local residual connections with the residual scaling. This fusion allows keeping information between layers of the proposed deep 3D model and avoiding vanishing gradient through it during the training process. It is done by allowing the flow of low-frequency information through multiple skip connections to enhance the generation of SR images. Figure 1 shows the full architecture of the proposed model that consists of two main phases which are preprocessing and the 3DRRDB CNN.

3.1. Preprocessing

First, all the low resolution (LR) images for a certain scene are registered to a reference image with maximum clearance where clearance refers to the part of the image with valid information (e.g., free of clouds). The registration process is important because the images for a certain scene are taken at a different time with different conditions and have a misalignment with the other LR images [2]. Resampling each pixel value is necessary to create a cohesive reference image. Translation as a transformation model during the registration process is used, which calculates the necessary shifts to register each image on both axes. In order to avoid improper registration caused by misaligned pixels, clearance masks are considered throughout this process. The registration is done using normalized cross-correlation in the Fourier domain [2].

Second, T frames (in our case, T = 9) with a clearance rate higher than 0.85 (i.e., 85% of pixels clear in the image) are selected to avoid any unclear scene. If a scene has a clearance rate lower than 0.85, it is removed completely. These values are fixed to have the same settings as in previous works [2, 37].

Third, augmentation is applied for the previously selected T LR images for each scene. As a result, in the T input images, there is no reason to prefer one order over another. The training dataset is pre-augmented by performing n_p random temporal permutations of the selected T input images to improve generalization. This allows us to train the algorithm to find the best temporal image regardless of where its position is in the input tensor.

Finally, each image forming the pack of T images is normalized by subtraction of the mean pixel intensity value and dividing it by the standard deviation derived across the entire dataset [2].

3.2. 3DRRDB CNN

The proposed 3DRRDB model consists mainly of eight Residual Dense Blocks (RDB). Each RDB is composed of a dense block. A global residual connection connects the input of the first RDB to the output of the last one. The global residual connection helps in passing gradient (information) from the input to the last layers and that will prevent the vanishing gradient problem from occurring in very deep CNNs [13]. Another addition to each RDB is the residual scaling to make training more stable [36, 38]. The pipeline of CNN ends with a convolutional layer followed by a bi-linear upsampling layer to scale by a factor of 3. The whole model architecture is shown in Figure 2.

The RDB consists of Dense Block, Local Residual Connection and Residual Scaling Factor as shown in Figure 2. Each Dense Block has five 3D convolutional layers and a local residual connection. In Dense Block, each convolutional layer has a Leaky Rectified Linear Unit (LReLU) activation function except the last convolutional layer as shown in Figure 3.

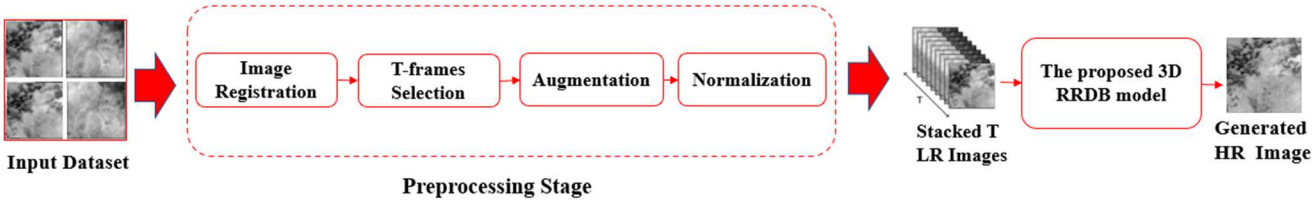

Figure 1. Full system architecture.

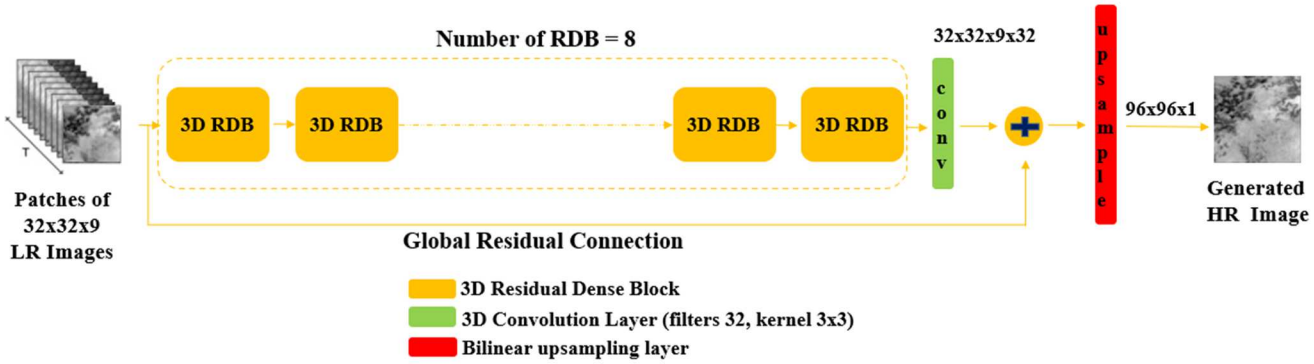

Figure 2. The proposed 3DRRDB DL model.

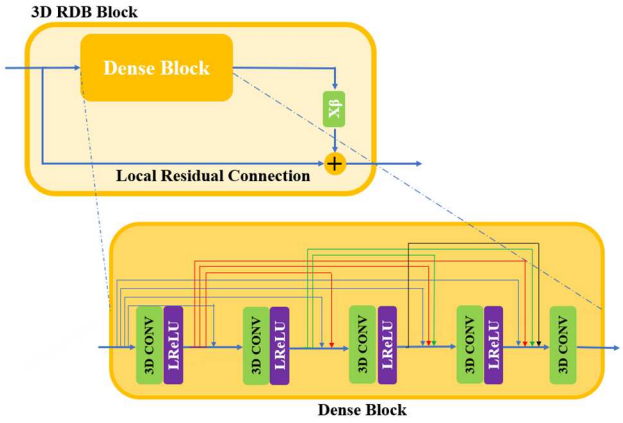

Figure 3. 3D Residual Dense Block structure.

The dense block is a connectivity pattern that optimizes the information flow between layers by applying direct connections from any layer to all following layers as shown in Figure 3 [35]. As a result, the feature maps of all the preceding layers, $x_0, \dots, x_{l-1}$ are sent to the l^{th} layer as shown in Eq. 1.

$$x_l = H_l([x_0, x_1, \dots, x_{l-1}]) , \quad (1)$$

where $[x_0, x_1, \dots, x_{l-1}]$ denotes the concatenation of feature-maps generated in layers $0, \dots, l-1$. In each dense block, a local residual connection is connected between the input of the block and the output of the block. The local residual connection is introduced to optimize the information flow because of the several 3D convolutions in each RDB. If we assume that the input of RDB is F_d and the output after the five 3D convolution layers is, F_{d+5}, then the final output of each RDB can be represented by Eq. 2.

$$RDB_{output} = F_d + Residual\ scaling(F_{d+5}) , \quad (2)$$

where the Residual scaling (β) is added to increase the stability of the output during training [38]. Different values of residual scaling are applied between 0.1 and 0.3 [38]. The best results were obtained at Residual scaling (β = 0.2). The RDB consists of 2 main parts which are the 3D Convolution Layers and the LReLU activation function.

3.2.1 3D Convolution Layers

In the area of the SR of satellite imagery, the focus is on the idea of the fusion of different features in temporal dimensions. Features are extracted from a certain scene through multiple captured spectral images for the same area [14]. This concept is defined as MISR which is more informative in capturing and finding the correlation of spatial and temporal features in resolving the low-resolution images 32×32 (LR) to have a 96×96 high-resolution Image (HR) than the SISR. For the previous reason, 3D convolution layers with 32 filters and a kernel size of 3×3×3 for T=9 were used in the model. The 3D convolution works by convolving 3D kernels to the stacked multi-images (frames) of the same scene captured at different times. By this construction, the feature maps in the convolution layer are connected to multiple contiguous frames in the previous layer [14]. For a unit value at position (x, y, z) on the j^{th} feature map in the i^{th} layer is defined as in Eq. 3.

$$v_{ij}^{xy} = ACT\left(b_{ij} + \sum_{m} \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} \sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right) (3)$$

where $ACT(\,)$ means the activation function, b_{ij} is the bias for the feature map, m indexes over the set of feature maps in the $(i-1)^{th}$ layer connected to the current feature map. w_{ijm}^{pqr} is the position value (p, q, r) of the kernel connected to the m^{th} feature map in the previous layer where P_i, Q_i and R_i are the height, width and depth of 3D kernel respectively [14]. Figure 4 shows how the kernel of 3D convolutions works.

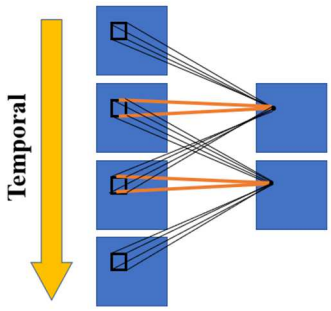

Figure 4. 3D kernel of 3D convolutional layers.

3.2.2 LReLU activation function

LReLU is used as a non-saturated activation function [39]. It is a variant of ReLU that overcomes the problem of "dying ReLU" which comes from the fact that it is unlikely for a neuron to recover once it has gone negative. Such neurons are essentially worthless because they do not play any part in discerning the input. Over time, a big portion of the network becomes idle where all the inputs with negative values are dropped and only the positive part is kept [39]. The enhancement added in the LReLU is that the negative input would generate a non-zero output and that can be defined as $f(x) = \max(ax, x)$ where x is the input to the activation function and α is a specified parameter that falls between (0,1). It should be emphasized that ReLU maps negative input to zero, whereas LReLU compresses negative input using a predetermined linear function [39].

4. Experiments Setup

In this section, dataset description, all the experimental settings, environment, packages, loss function and performance metrics used will be explained.

4.1. Dataset Description

The European Space Agency's (ESA) Advanced Concept Team provided a dataset for MISR Problem [2, 31, 32]. It can be found in [32]. There was no evidence of a clear temporal order for the captured image for one region [2]. The data consists of 300 m and 100 m resolution radiometrically and geometrically corrected Top-Of-Atmosphere (TOA) reflectance's in Plate Carrée projection for the RED and NIR spectral bands. For the 300 m low resolution (LR) images, the data was delivered in 128×128 grey-scale pixels, and for the 100 m High Resolution (HR) images, it was given in 384×384 grey-scale pixels [31]. A quality map was included with each image, indicating which pixels were hidden (clouds, cloud shadows, ice, water, missing pixels, etc.) and which should be regarded clearly. For HR images, at least 75% of the pixels must be clear, and for LR images, at least 60% of the pixels must be clear [31]. Each scene consisted of several 128×128 pixels LR pictures (varying from 9 to 35 depending on the scene) and a single 384×384 pixels HR ground truth. Even though the real signal bit-depth is 14 bits, the images were encoded as 16-bit png files [2, 31]. The dataset was divided into two parts: the train section contains both LR and HR images while the test part contains only the LR images without ground truth or HR images. The experiments were done on the training data only as in [1, 2, 33, 34, 37] to validate the effectiveness of our approach. As a result, the train section was divided into training and validation sets. The dataset is divided into 396 training scenes and 144 validation scenes for the NIR band while the RED band scenes were divided into 415 training scenes and 146 validation scenes. The same validation pictures as in [1, 2, 33, 34, 37] were utilized to make comparisons with earlier approaches easier.

4.2. Experimental Settings

Our model implementation is done using Python language with Keras package (TensorFlow backend). The model is proposed to upscale the LR image by a factor of three (s = 3) because of the Proba-V dataset specifications described in section 4.1. Experiments are carried out on a core i7 CPU running at 3.80 GHz with 32 GB of RAM and an Nvidia RTX 3090 graphics card with 24 GB of RAM. The model is trained for 200 epochs using the Adam optimizer with Nesterov Momentum with a starting learning rate of 0.0001. Inputs are divided into batches of size 32. For each epoch, the Peak Signal to Noise Ratio (PSNR) is measured. The best model is characterized as having the highest PSNR, and it is then saved and used on the testing set.

In the preprocessing stage detailed in Section 3.1, we apply data augmentation by generating 10 permutations (n_p = 10) of the T input images for each scene in the dataset. This results in a total of 3930 NIR and 4150 RED training images. Augmentation is not applied to validation data to keep it with the same size and establish a fair comparison. After that, 16 patches of each LR training image are extracted with a size of 32×32 pixels and the corresponding clearance masks and HR images with a size of 96×96 pixels. After a further improvement is applied by removing the patches with a clearance rate lower than 0.85. The total number of training image patches for NIR and RED bands becomes 62880 image patches and 66400 image patches, respectively while the validation image patches for NIR and RED are 2720 image patches and 2816 image patches respectively.

4.3. Loss Function

A special loss function is adopted because different images in the same scene can have quite different circumstances. It is critical to make the loss function independent of possible intensity biases between the target

image $I_{u,v}^{HR}$ and super-resolved $I_{u,v}^{SR}$. The loss function can be defined in Eq. 4 [2] as the smallest mean absolute error (L1 loss) between $I_{u,v}^{HR}$ and $I_{u,v}^{SR}$ and the Mean absolute error as proposed in [2] due to its good results.

$$L = min_{u,v \in [0,2d]} \frac{\left\| I_{u,v}^{HR} - \left(I_{u,v}^{SR} + b_{u,v} \right) \right\|}{(sH - 2d)(sW - 2d)} \ , \qquad (4)$$

where a super-resolved output cropped of d pixels on each border is defined as $I_{u,v}^{SR}$ and each possible patch is considered as $I_{u,v}^{HR}$, u,v ϵ [0,2d] of size (sH – 2d) x (sW – 2d) extracted from the ground truth I^{HR}. ||…|| describe L1 Norm (absolute value summation). $b_{u,v}$ represents the mean biases between $I_{u,v}^{HR}$ and $I_{u,v}^{SR}$ patches, defined as shown in Eq. 5.

$$b_{u,v} = \frac{\sum_{i=1}^{sH-2d} \sum_{j=1}^{sW-2d} \left[I_{u,v}^{HR} - I_{u,v}^{SR} \right](i,j)}{(sH - 2d)(sW - 2d)} \ . \qquad (5)$$

4.4. Performance metrics

First, the proposed model is evaluated using two quantitative evaluation metrics which are Corrected Peak Signal-to-Noise ratio (cPSNR) and Corrected Structural Similarity Index Measure (cSSIM). The two metrics are adopted by state-of-the-art [1, 2, 33, 34].

cPSNR is a modified version of PSNR that handles the disadvantage of high sensitivity of PSNR towards biases in brightness [32]. cPSNR is calculated using Corrected Mean Squared Error (cMSE) proposed by [31]. cMSE is the minimum Mean Squared Error (MSE) between $I_{crop}^{SR} + b_{u,v}$ and HR patches $I_{u,v}^{HR}$ that can be defined in Eq. 6 [2].

$$cMSE = \min_{u,v \in [0,6]} MSE_{clear}(I_{u,v}^{HR}, I_{u,v}^{SR} + b_{u,v}) \ , \qquad (6)$$

where $b_{u,v}$ is the mean biases defined in Eq. 5 and where MSE_{clear} denotes the mean squared error computed exclusively on pixels in the clearance mask [2]. The cPSNR can be computed as in Eq. 7.

$$cPSNR = 10 \log_{10} \frac{(2^{16} - 1)^2}{cMSE} \ , \qquad (7)$$

where $2^{16} - 1$ is the highest possible pixel intensity for a 16-bit encoded image [2].

Similarly, cSSIM is the maximum SSIM between $I_{crop}^{SR} + b_{u,v}$ and HR patches $I_{u,v}^{HR}$ multiplied for the clearance mask $M_{u,v}^{HR}$ and can be defined as in Eq. 8.

$$cSSIM = \max_{u,v \in [0,6]} SSIM(I_{u,v}^{HR} . M_{u,v}^{HR}, I_{u,v}^{SR} . M_{u,v}^{HR} + b_{u,v}) \ . \ (8)$$

Second, the proposed model 3DRRDB is evaluated and compared to state-of-the-art methods using a qualitative metric which is the quality map [1]. The quality maps are generated by the absolute difference between the HR target and the SR reconstructions for the Bicubic.

Finally, the efficiency of the best models in state-of-the-art is compared to our proposed model.

5. Results

First, a partial ablation study is conducted to evaluate the importance of 3D convolution layers in RRDB model compared to 2D convolution layers.

Second, two experiments are conducted using the proposed 3DRRDB model. The two experiments are applied on the NIR band and RED band separately that is provided by Proba-V dataset [31]. The proposed 3DRRDB is compared to different models of the state-of-the-art which are Bicubic [2], IBP [23], RCAN [12], Dynamic Adaptive Filter (VDR-DUF) [40], HighRes-net [33], DeepSUM [1], DeepSum++ [34], RAMS [2] and RAMS+$_{20}$ [2].

Finally, number of learning parameters for the proposed 3DRRDB is compared to the highest model results, RAMS [2] and RAMS+$_{20}$ [2], to have a clear view on the complexity of both models.

5.1. 2DRRDB vs 3DRRDB

The importance and effectiveness of the 3D convolutions in our proposal has been evaluated by replacing them by 2D convolutions.

As shown in Table 1, the proposed 3DRRDB model overwhelms the 2DRRDB in all bands (NIR and RED). The proposed 3DRRDB scores an increase in cPSNR of 1.23 dB in the NIR band and 1.24 dB in the RED band compared to the 2DRRDB. Moreover, the 3DRRDB achieves a cSSIM greater than 2DRRDB.

Table 1. Ablation Study: Quantitative Comparison between the proposed 3DRRDB model and 2DRRDB cPSNR and cSSIM metrics for NIR and RED bands.

	NIR Band		**RED Band**	
	cPSNR	**cSSIM**	**cPSNR**	**cSSIM**
2DRRDB	47.56	0.9818	49.59	0.9874
Our model	**48.79**	**0.9865**	**50.83**	**0.9909**

5.2. NIR Band Experiment Results

The proposed 3DRRDB model outperforms all state-of-the-art models in cPSNR metric for NIR band images as shown in Table 2. It scores an increase of 0.28 dB, 0.56 dB, 0.86 dB and 0.95 dB in comparison to NIR band images results of the best four models RAMS+$_{20}$ [2], RAMS [2], DeepSUM++ [34], DeepSUM [1] respectively. Furthermore, the 3DRRDB model surpasses the Bicubic

[2], IBP [21], RCAN [19] and Dynamic Adaptive Filter (VDR-sDUF) [41] models in terms of cSSIM metric for NIR band images. The proposed 3DRRDB model scores a slight increase to DEEPSUM++ model [32], DEEPSUM [1] and HighRes-net [31] models in SSIM of NIR band images with a comparable cSSIM to RAMS+$_{20}$ [2] and RAMS [2] models as shown in Table 2.

Figure 5 illustrates qualitative visual comparison between the SR images reconstructed by the Bicubic [2], RAMS+$_{20}$ [2] and the proposed 3DRRDB models for the NIR band, respectively. Moreover, the generated quality maps of the suggested 3DRRDB model generate images with sharper edges and finer textures, as well as images that are more artistically detailed. Figure 6 shows the quality maps of the same SR examples in Figure 5. for Bicubic [2], RAMS+$_{20}$ [2] and the proposed 3DRRDB models.

Table 2. Quantitative Comparison between the proposed 3DRRDB model and state-of-the-art models in terms of cPSNR and cSSIM metrics for NIR and RED bands.

	NIR Band		**RED Band**	
	cPSNR	**cSSIM**	**cPSNR**	**cSSIM**
Bicubic [2]	45.12	0.9767	47.63	0.9846
IBP [23]	45.96	0.9796	48.21	0.9865
BTV [41]	45.93	0.9794	48.12	0.9861
RCAN [12]	45.66	0.9798	48.22	0.9870
VDR-DUF [40]	47.20	0.9850	49.59	0.9902
HighRes-net [33]	47.55	0.9855	49.75	0.9904
DeepSUM [1]	47.84	0.9858	50.00	0.9908
DeepSUM++ [34]	47.93	0.9862	50.08	0.9912
RAMS [2]	48.23	0.9875	50.17	0.9913
RAMS+$_{20}$ [2]	48.51	**0.9880**	50.44	**0.9917**
Our model	**48.79**	0.9865	**50.83**	0.9909

5.3. RED Band Experiment Results

The results in the RED band reproduce the same behavior observed in the NIR band and that can be shown in Table 2, Figure 7 and Figure 8. The proposed 3DRRDB surpassed all state-of-the-art models. It scores an increase in cPSNR of RED band images scoring 0.39 dB, 0.66 dB, 0.75 dB and 0.83 dB compared to the best four models RAMS+$_{20}$ [2], RAMS [2], DeepSUM++ [34], DeepSUM [1] respectively. Moreover, the 3DRRDB model surpasses the Bicubic [2], IBP [23], RCAN [12] and Dynamic Adaptive Filter (VDR-DUF) [40] models in terms of cSSIM metric of RED band images as shown in Table 2. The 3DRRDB RED band images results scores a slight increase in cSSIM for RED band images compared to DeepSUM model [1] with comparable cSSIM results to RAMS+$_{20}$ [2], RAMS [2] and DeepSUM++ [34] models as shown in Table 2.

Figure 7 demonstrates a qualitative visual comparison between the SR images reconstructed by the Bicubic [2], RAMS+$_{20}$ [2] and the proposed 3DRRDB models for the RED band, respectively. Furthermore, 3DRRDB model generated quality maps with sharper edges and finer textures, as well as images that are more artistically detailed. Figure 8 shows the quality maps of the same SR examples in Figure 7 for Bicubic [2], RAMS+$_{20}$ [2] and the proposed 3DRRDB models.

5.4. Efficiency

We also compare the number of trainable parameters, which indicates the model's complexity and, as well as the risk of overfitting and the need for a large amount of training data. Table 3 shows the approximate number of trainable parameters for RAMS [2], RAMS+$_{20}$ [2] and the proposed 3DRRDB models. The highest complexity is achieved by the RAMS+$_{20}$ [2] with 19M parameters while our proposed 3DRRDB has the least trainable parameter of 456k. Our proposed 3DRRDB model has about 41 times fewer trainable parameters than RAMS+$_{20}$ [2], and about half the trainable parameters of RAMS [2]. These results prove the efficient complexity provided by the proposed 3DRRDB model compared to the best models in state-of-the-art.

Table 3. Comparison of the number of trainable parameters best state-of-the-art model's result and the proposed 3DRRDB model.

No. of Trainable Parameters		
RAMS [2]	RAMS+$_{20}$ [2]	**Our model**
958k	19M	**456k**

6. Conclusion

In this paper, a novel 3D DL model named 3DRRDB for image super-resolution is introduced which effectively can upsample multiple low-resolution images for each scene to a high-resolution image with a scaling factor of 3. The proposed 3DRRDB model introduces the idea of focusing on correlation between numerous LR pictures for each scene using dense blocks with 3D convolutions that improved super-resolution generation by maximizing information flow. We have proved that 3D convolutions are more relevant for MISR techniques than 2D convolutions on RRDB model. Moreover, the proposed model introduces the idea of mixing global and local residual connections with residual scaling to 3D convolutions to reduce vanishing gradients during the training process. The proposed 3DRRDB model shows great potential in MISR by surpassing the state-of-the-art models in cPSNR results scoring 48.79 dB and 50.83 dB for NIR and RED bands respectively and with comparable cSSIM results. Moreover, our proposed 3DRRDB model has much lower complexity compared to the best resulted state-of-the-art models. In addition, the model can be generalized to different scaling factors and can be extended, as future work, to be applied for multi-spectral remote sensing images.

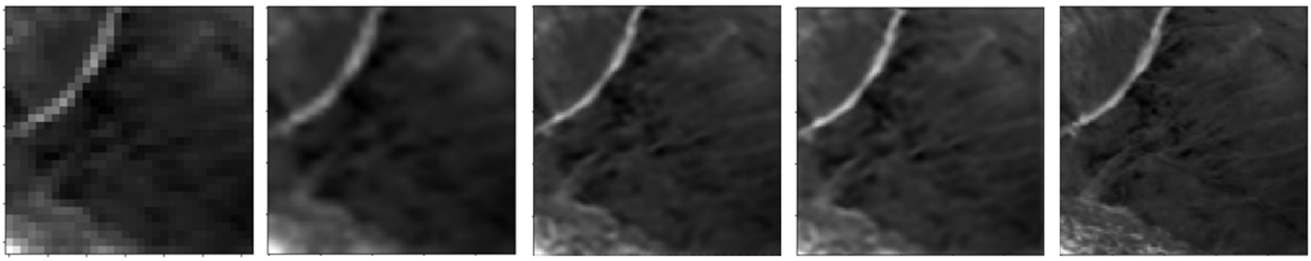

Figure 5. An example of NIR band image qualitative comparison. Left to right: LR image, Bicubic SR image, RAMS+20 SR image, the proposed 3DRRDB SR image and HR image.

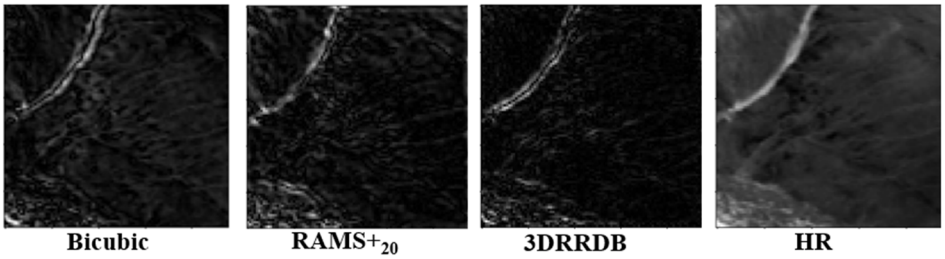

Figure 6. An example of NIR band quality maps. Left to right: Bicubic quality map, RAMS+20 quality map, the proposed 3DRRDB quality map and HR image.

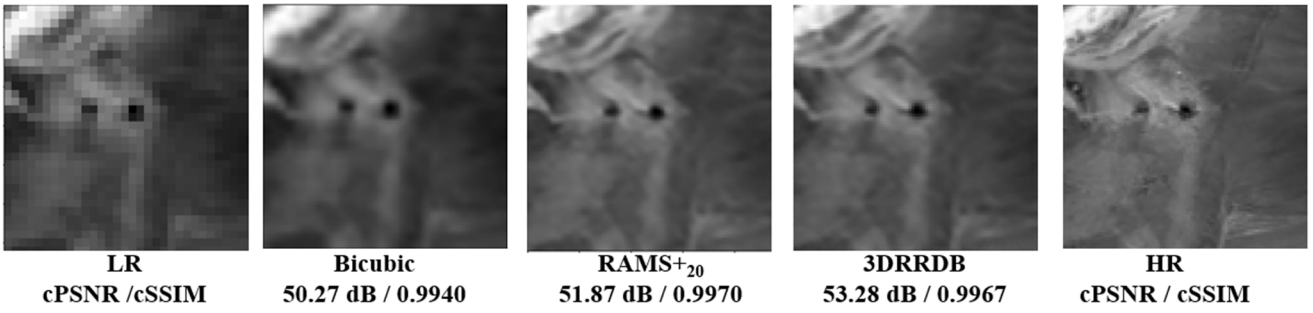

Figure 7. An example of RED band images qualitative comparison. Left to right: LR image, Bicubic SR image, RAMS+20 SR image, the proposed 3DRRDB SR image and HR image.

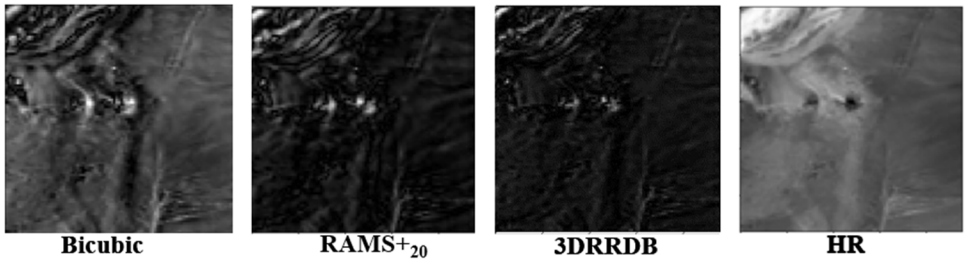

Figure 8. An example of RED band quality maps. Left to right: Bicubic quality map, RAMS+20 quality map, the proposed 3DRRDB quality map and HR image.

Acknowledgement

This work has been supported by the Spanish Ministry of Science and Innovation under project BOSSS TIN2017-89723-P.

References

[1] A. Bordone Molini, D. Valsesia, G. Fracastoro, and E. Magli, "DeepSUM: Deep Neural Network for Super-Resolution of Unregistered Multitemporal Images," *IEEE*

Transactions on Geoscience and Remote Sensing, vol. 58, no. 5, pp. 3644-3656, 2020.

[2] 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, pp. 1-20, 2020.

[3] D. Valsesia and E. Magli, "A Novel Rate Control Algorithm for Onboard Predictive Coding of Multispectral and Hyperspectral Images," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 52, no. 10, pp. 6341-6355, 2014.

[4] P. Benecki, M. Kawulok, D. Kostrzewa, and L. Skonieczny, "Evaluating super-resolution reconstruction of satellite images," *Acta Astronautica,* vol. 153, pp. 15-25, 2018.

[5] D. Valsesia and P. T. Boufounos, "Universal encoding of multispectral images," presented at the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016. Available: https://doi.org/10.1109/ICASSP.2016.7472519

[6] C. Duchon, "Lanczos Filtering in One and Two Dimensions," *Journal of Applied Meteorology - J APPL METEOROL,* vol. 18, pp. 1016-1022, 1979.

[7] M. K. Ng, H. Shen, E. Y. Lam, and L. Zhang, "A Total Variation Regularization Based Super-Resolution Reconstruction Algorithm for Digital Video," *EURASIP Journal on Advances in Signal Processing,* vol. 2007, no. 1, p. 074585, 2007.

[8] M. Protter, M. Elad, H. Takeda, and P. Milanfar, "Generalizing the Nonlocal-Means to Super-Resolution Reconstruction," *IEEE Transactions on Image Processing,* vol. 18, no. 1, pp. 36-51, 2009.

[9] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Computer Graphics and Applications,* vol. 22, no. 2, pp. 56-65, 2002.

[10] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3791-3799.

[11] Mapwaredev. (2020, Accessed on: Mar. 20, 2022. [Online]). *Understanding Errors and Distortion in Remote Sensing*. Available: https://mapware.ai/blog/understanding-errors-and-distortion-in-remote-sensing/

[12] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. R. Fu, "Image Super-Resolution Using Very Deep Residual Channel Attention Networks," in *European Conference on Computer Vision*, 2018, pp. 286-301.

[13] K. Zhang *et al.*, "Residual Networks of Residual Networks: Multilevel Residual Networks," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 28, no. 6, pp. 1303-1314, 2018.

[14] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 35, no. 1, pp. 221-231, 2013.

[15] J. Sun, J. Sun, Z. Xu, and H. Y. Shum, "Gradient Profile Prior and Its Applications in Image Super-Resolution and Enhancement," *IEEE Transactions on Image Processing,* vol. 20, no. 6, pp. 1529-1542, 2011.

[16] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a Deep Convolutional Network for Image Super-Resolution," in *European Conference on Computer Vision*, 2014, pp. 184-199.

[17] Li, J., Fang, F., Mei, K., Zhang, G., "Multi-scale Residual Network for Image Super-Resolution," in *European Conference on Computer Vision*, 2018, pp. 527-542

[18] Y. Tai, J. Yang, and X. Liu, "Image Super-Resolution via Deep Recursive Residual Network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2790-2798.

[19] J. Y. Y. Tai, X. Liu and C. Xu, "MemNet: A Persistent Memory Network for Image Restoration," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4549-4557.

[20] C. Ledig *et al.*, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 105-114.

[21] M. S. M. Sajjadi, B. Schölkopf, and M. Hirsch, "EnhanceNet: Single Image Super-Resolution Through Automated Texture Synthesis," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4501-4510.

[22] R. Y. Tsai and T. S. Huang, "Multiframe image restoration and registration," *Advances in Computer Vision and Image Processing,* pp. 313-339, 1984.

[23] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graphical Models and Image Processing,* vol. 53, no. 3, pp. 231-239, 1991.

[24] H. Stark and P. Oskoui, "High-resolution image recovery from image-plane arrays, using convex projections," *Journal of the Optical Society of America A,* vol. 6, no. 11, pp. 1715-1726, 1989.

[25] S. Lertrattanapanich and N. K. Bose, "High resolution image formation from low resolution frames using Delaunay triangulation," *IEEE Transactions on Image Processing,* vol. 11, no. 12, pp. 1427-1441, 2002.

[26] H. Shen, M. K. Ng, P. Li, and L. Zhang, "Super-Resolution Reconstruction Algorithm To MODIS Remote Sensing Images," *The Computer Journal,* vol. 52, no. 1, pp. 90-100, 2009.

[27] T. Kato, H. Hino, and N. Murata, "Double sparsity for multi-frame super resolution," *Neurocomputing,* vol. 240, pp. 115-126, 2017.

[28] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video Super-Resolution With Convolutional Neural Networks," *IEEE Transactions on Computational Imaging,* vol. 2, no. 2, pp. 109-122, 2016.

[29] J. Caballero *et al.*, "Real-Time Video Super-Resolution with Spatio-Temporal Networks and Motion Compensation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4778-4787.

[30] M. Kawulok, P. Benecki, S. Piechaczek, K. Hrynczenko, D. Kostrzewa, and J. Nalepa, "Deep Learning for Multiple-Image Super-Resolution," *IEEE Geoscience and Remote Sensing Letters,* vol. 17, no. 6, pp. 1062-1066, 2020.

[31] M. Märtens, D. Izzo, A. Krzic, and D. Cox, "Super-resolution of PROBA-V images using convolutional neural networks," *Astrodynamics* vol. 3, pp. 387-402, 2019.

[32] E. S. Agency. (Nov, 2018, Accessed on: Dec 25, 2021. [Online]). *PROBA-V Super Resolution*. Available: https://kelvins.esa.int/proba-v-super-resolution/home/

[33] M. Deudon *et al.*, "HighRes-net: Recursive Fusion for Multi-Frame Super-Resolution of Satellite Imagery," *ArXiv,* vol. abs/2002.06460, 2020.

[34] A. Bordone Molini, D. Valsesia, G. Fracastoro, and E. Magli, "DeepSUM++: Non-local Deep Neural Network for Super-Resolution of Unregistered Multitemporal Images," in *International GeoScience and Remote Sensing Symposium*, 2020, pp. 609-612.

[35] Z. L. G. Huang, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261-2269.

[36] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual Dense Network for Image Super-Resolution," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2472-2481.

[37] D. Valsesia and E. Magli, "Permutation Invariance and Uncertainty in Multitemporal Image Super-Resolution," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 60, pp. 1-12, 2022.

[38] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 4278–4284.

[39] X. Zhang, Y. Zou, and W. Shi, "Dilated convolution neural network with LeakyReLU for environmental sound classification," in *2017 22nd International Conference on Digital Signal Processing (DSP)*, 2017, pp. 1-5.

[40] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, "Deep Video Super-Resolution Network Using Dynamic Upsampling Filters Without Explicit Motion Compensation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3224-3232.

[41] 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.