Multi-Image Super Resolution of Remotely Sensed Images Using Residual Attention Deep Neural Networks

remote sensing

MDPI

Article

Multi-Image Super Resolution of Remotely Sensed Images Using Residual Attention Deep Neural Networks

Francesco Salvetti [1,2,3], Vittorio Mazzia [1,2,3,*], Aleem Khaliq [1,2,4] and Marcello Chiaberge [1,2]

[1] Department of Electronics and Telecommunications, Politecnico di Torino, 10129 Turin, Italy; francesco.salvetti@polito.it (F.S.); aleem.khaliq@polito.it (A.K.); marcello.chiaberge@polito.it (M.C.)
[2] PIC4SeR, Politecnico di Torino Interdepartmental Centre for Service Robotics, 10129 Turin, Italy
[3] SmartData@PoliTo, Big Data and Data Science Laboratory, 10129 Turin, Italy
[4] Department of Electrical Engineering, International Islamic University, Islamabad 44000, Pakistan
* Correspondence: vittorio.mazzia@polito.it

Received: 20 June 2020; Accepted: 6 July 2020; Published: 10 July 2020

Abstract: Convolutional Neural Networks (CNNs) consistently proved state-of-the-art results in image Super-resolution (SR), representing an exceptional opportunity for the remote sensing field to extract further information and knowledge from captured data. However, most of the works published in the literature focused on the Single-image Super-resolution problem so far. At present, satellite-based remote sensing platforms offer huge data availability with high temporal resolution and low spatial resolution. In this context, the presented research proposes a novel residual attention model (RAMS) that efficiently tackles the Multi-image Super-resolution task, simultaneously exploiting spatial and temporal correlations to combine multiple images. We introduce the mechanism of visual feature attention with 3D convolutions in order to obtain an aware data fusion and information extraction of the multiple low-resolution images, transcending limitations of the local region of convolutional operations. Moreover, having multiple inputs with the same scene, our representation learning network makes extensive use of nestled residual connections to let flow redundant low-frequency signals and focus the computation on more important high-frequency components. Extensive experimentation and evaluations against other available solutions, either for Single or Multi-image Super-resolution, demonstrated that the proposed deep learning-based solution can be considered state-of-the-art for Multi-image Super-resolution for remote sensing applications.

Keywords: deep learning; multi-image super-resolution; attention networks; 3D convolutional neural networks

1. Introduction

Super-resolution (SR) algorithms serve the purpose of reconstructing high-resolution (HR) images from either single or multiple low-resolution (LR) images. Due to constraints such as sensor limitations and exceedingly high acquisition costs, it is often challenging to obtain HR images. In this regard, SR algorithms provide viable opportunity to enhance and reconstruct HR images from LR images recorded by the sensors. Over more than three decades, progress has steadily been observed in the development of super-resolution, as both multi-frame and single-frame SR now have substantial applications that can use the image generation purposefully.

SR is very significant to Remote Sensing because it provides opportunity to enhance LR images despite the inherent problems often faced in remote-sensing scenarios. The hardware and material costs for smaller missions around data accumulation are very high. Additionally, onboard instruments on satellites continue to generate ever-increasing data as spatial and spectral resolutions also increase,

and this has progressively become challenging for compression algorithms [1], as they try to meet the bandwidth restrictions [2,3]. Remote sensing is fundamental in obtaining images covering most of the globe, permitting many vital projects such as disaster monitoring, military surveillance, urban maps, and vegetation growth monitoring. It is thus imperative that enhancements and progress be made in post-processing techniques to overcome obstacles of increasing spatial resolution.

There are two main methods used in Super-resolution: Single-image SR (SISR) and Multi-image SR (MISR). SISR employs a single image to reconstruct a HR version of it. However, a single image is quite limited in the amount of information that it provides, mainly post the LR image formation process. Contrastingly, MISR involves multiple LR images of the same scene acquired from the same or different sensors to construct an HR image. The significant advantage MISR holds over SISR is in how it can draw out otherwise unavailable information from the different image observations of the same scene. It consequently constructs high spatial resolution image. However, to achieve the additional benefits of MISR, a multitude of problems need to be solved. Conventionally, multiple images are obtained by either a satellite during its multiple orbits or by different satellites at different times or different sensors acquiring images at the same time. With so many variables involved, many complications need to be considered, such as cloud coverage, time variance in scene content, and invariance to absolute brightness variability.

There has been significant progress in Single-image SR as deep learning methods and deep neural networks were brought into use, allowing a better efficient generation of non-linear maps to deal with complex degradation models. However, there has not been any similar progress in MISR.

In this paper, building over the latest breakthroughs in SISR [4–8], we propose a deep learning MISR solution for remote-sensing applications that exploits both spatial and temporal correlations to combine multiple low-resolution acquisitions smartly. Indeed, our model provides a real end-to-end efficient solution to recover high-resolution images, overcoming limitations of previous similar methodologies, and providing enhanced reconstruction results. Therefore, the presented research constitutes an exceptional opportunity, easily replicable, to access better quality and more useful information for the remote-sensing community. In particular, the main contribution of our work lies in:

1. The use of 3D convolutions to efficiently extract, directly from the stack of multiple low-resolution images, high-level representations, simultaneously exploiting spatial and temporal correlations.
2. The introduction of a novel feature attention mechanism for 3D convolutions that lets the network focus on most promising high-frequency information largely overcoming main locality limitations of convolutional operations. Moreover, the concurrent use of multiple nested residuals, inside the network, let low-frequency components flow directly to the output of the model.
3. The conceptualization and development of an efficient, highly replicable, deep learning neural network for MISR that makes use of 2D and 3D convolutions exclusively in the low-resolution space. It has been extensively evaluated on a major multi-frame open-source remote-sensing dataset proving state-of-the-art results with a considerable margin. Therefore, it constitutes an exceptional tool and opportunity for the remote-sensing research community.

The remainder of this paper is structured as follows. Section 2 covers the related work on SR and its developments in techniques for both SISR and MISR. Section 3 explains the overall methodology, network architecture and its subsequent blocks, and training process. Section 4 discusses the experimentation, the Proba-V dataset, data pre-processing, and results. Section 5 draws some conclusions and future directions.

2. Related Work

Related literature is organized as follows. Firstly, a wide range of studies related to SISR are discussed which involve state-of-the-art methods and recent developments in SISR techniques, which is the basis of every SR method. Secondly, studies performed for SR in remote sensing domain are

discussed. Lastly, MISR related studies, which are rarely addressed, are discussed including latest developments.

2.1. Single-Image Super-Resolution

Ever since the late 1980s and the early 1990s, there has been an eager interest in SR, comprehensively reviewed by Borman and Stevenson [9]. Following forth in the works of Tsai and Huang [10] and afterward, Kim et al. [11], the new approaches considered processing images in the frequency domain to recover lost information of higher-frequency. These first works had certain drawbacks, such as the level of difficulty observed in successfully incorporating the prior available spatial information. Several studies performed by Irani and Peleg [12–14] focused over the spatial domain, proposing methods for SR reconstruction.

Learning-based methods build upon the relation between LR-HR images, and there have been many recent advancements in this approach, mostly due to deep convolutional neural networks (CNNs) [4,15,16]. The leading force for this was Dong et al. [17], who achieved superior results by proposing a Super-resolution CNN (SRCNN) framework. Kim et al. introduced residual learning and suggested very deep SR (VDSR) [16] and deeply recursive CN (DRCN) [18] with 20 layers. Later, Tai et al. pioneered deep recursive residual network (DRRN) [19] and memory blocks in MemNet [20]. There is some inevitable loss of details; however, the increase in computation is significant. So going forth, particular emphasis has been placed on proper upscaling of spatial resolutions at network tail-ends, as well as extracting information of the original scale LR inputs. To that end, some enhancements were proposed for accelerating the testing and training needed for SRCNN, a faster network structure FSRCNN [15]. Ledig et al. [21] proposed SRGAN, a generative adversarial network (GAN) for photo-realistic SR with perceptual losses [22], and K. He et al. introduced ResNet [23] for image SR and to make a deeper network SRResNet. EnhanceNet [24] also used a GAN-based model to merge perceptual loss with automated texture synthesis. Though, the predicted results can produce some artifacts and may not be a faithful reconstruction.

In recent past years, enhancements in deep networks have been proposed and showed promising results for SISR, for example, in [5], an Enhanced Deep Super-resolution (EDSR) network was developed to improve the performance by removing unnecessary modules and expanding the model size with the stable training process in conventional residual networks. Yu et al. [6] demonstrated better results in terms of accurate SR by generating models with a wide range of features before ReLU activation and training with normalized weights. Zhang et al. [7] proposed residual channel attention networks (RCAN) that exploits very deep network structure based on residual in residual (RIR) which bypass excessive low-frequency information through multiple skip connections.

2.2. SR for Remotely Sensed Imagery

With the increasing availability of recent satellite-based multispectral sensors and transmission bandwidth restrictions [25], it is possible to obtain images at different spatial resolutions with multiple spectral bands. Keen attention is being paid to developing better methods of super-resolving the lower-resolution bands but simultaneously keeping the images at a high spatial resolution. An example can be seen in [26], where-through the use of only lower resolution bands–SR of multispectral remote sensing images is applied with convolutional layers. [27] shows the integration of residual connections into a single image SR-based architecture to achieve better SR performance. The performance of image enhancement methods in computer vision can also be increased prominently through generative adversarial networks (GANs) [21,28]. Moreover, GANs were also exploited to super-resolve remote sensing images. For example, Ma et al. [29] developed a dense residual generative adversarial network (DRGAN)-based SISR method to super resolve remote sensing images. By designing a dense residual network as the generative network in GAN, their method makes full use of the hierarchical features from low-resolution (LR) images.

Dong et al. [30] proposed a novel multi-perception attention network (MPSR) for Super-resolution of low resolution remotely sensed images, which achieved better results by incorporating the proposed enhanced residual block (ERB) and residual channel attention group (RCAG). Their methodology is capable of dealing with low-resolution remote sensing images via multi-perception learning and multi-level information adaptive weighted fusion. They claimed that a pre-train and transfer learning strategy can improve the SR performance and stabilize the training procedure. Gargiulo et. al. [31] proposed a CNN-based approach to provide a 10 m super-resolved image of the original 20 m bands of remotely sensed Sentinel-2 images. In their experimental results, they claimed that the proposed solution can achieve better performance with respect to most of the state-of-the-art methods, including other deep learning based ones with a considerable saving of computational burden. Recently methods to enhance spatial resolution of remotely sensed images used Parallel Residual Network [32], Bidirectional Convolutional LSTMs [33], Deep Residual Squeez and Excitation Network [34].

2.3. Multi-Image Super-Resolution

Multi-image SR (MISR) involves the extraction of information from many LR observations of the same scene to reconstruct HR images [35]. The earliest work for MISR was proposed by Tsai and Huang [10] using a frequency-domain technique, by combining multiple images with sub-pixel displacements to improve the spatial resolution of images. Due to the some weaknesses of the first proposed method related to incorporate prior information of HR images, several spatial domain MISR techniques were considered [36]. These include projection onto convex sets (POCS) [37], non-uniform interpolation [38], regularized methods [39,40], and sparse coding [41]. With the availability of more data from the multiple observations of the scene, it is possible to obtain a more accurate reconstruction than through single-image methods. MISR techniques involve different ways of degrading the original image by following an image model, and these involve blurring, warping, noise contamination, and decimation. Then the degradation is reversed by solving an ill-posed optimization problem. To this end, Bayesian reconstruction in the gradient projection algorithm was used alongside subpixel displacement estimation [42]. An enhanced Fast and Robust SR (FRSR) [43] employs estimation of maximum likelihood analysis and simplified regulation. Another proposal in SR was for the Adaptive detail enhancement (SR-ADE) [44], which reconstructs satellite images with the use of a bilateral filter for decomposing input images while also amplifying high-frequency detail information.

Another approach Iterative Back Projection (IBP), introduced by Irani and Peleg [13], used a back-projection of the difference between the actual LR images obtained and the simulated LR images to the SR image. The forward imaging process is inverted and iteratively attempted in updates. As with MISR, there are apparent drawbacks when prior images are difficult to be included, or it is difficult to model an image's degradation process.

In the past few years, many deep learning-based approaches have been exploited to address the MISR problems in the context of enhancing video sequences [45–47]. However, MISR is rarely exploited for remotely sensed satellite imagery. Kawulok et al. [48] demonstrated the potential benefits of information fusion offered by multiple satellite images reconstruction and learning-based SISR approaches. In their work, EvoNet framework [49] based on several deep CNNs was adopted to exploit SISR in the preprocessing phase of the input data for MISR.

Recently, a challenge was set by the European Space Agency (ESA) to super-resolved multi-temporal PROBA-V satellite imagery (https://kelvins.esa.int/proba-v-super-resolution (accessed on 2 July 2020)). In this context, a new CNN-based architecture *DeepSUM* was proposed by Molini et al. [50] to super resolve multi-temporal PROBA-V imagery. An end-to-end learning approach was established by exploiting both spatial and temporal correlations. Most recently, Deudon et al. presented *HighRes-net* based on deep learning to deal with the MISR of remotely sensed PROBA-V satellite imagery [51]. They proposed an end-to-end mechanism that learns the sub-tasks involved in MISR, which are co-registration, fusion, upsampling, and registration-at-the-loss.

3. Methodology

MISR aims at recovering an HR image I^{HR} from a set of T LR images $I^{\mathrm{LR}}_{[1,T]}$ of the same scene acquired in a certain temporal window. In contrast to SISR, MISR can simultaneously benefit from spatial and temporal correlations, being able to achieve far better reconstruction results theoretically. Either way, SR is an inherently ill-posed problem since a multiplicity of solutions exist for any given set of low-resolution images. Therefore, it is an underdetermined inverse problem, whose solution is not unique. Our proposed methodology, based on a representation learning model, aims at generating a super-resolved image I^{SR} applying a function H_{RAMS} to the set of $I^{\mathrm{LR}}_{[1,T]}$ images:

$$I^{\mathrm{SR}} = H_{\mathrm{RAMS}}(I^{\mathrm{LR}}_{[1,T]}, \Theta) \tag{1}$$

where Θ are model parameters learned with an iterative optimization process.

In other words, we propose a fully convolutional Residual Attention Multi-image Super-resolution network (RAMS) that can efficiently extract high-level features concurrently from T LR images and fuse them exploiting a built-in visual attention mechanism. Attention directs the focus of the model only on most promising extracted features, reducing the importance of less relevant ones and mostly transcending limitations of the local region of convolutional operations. Moreover, extensive use of nested residual connections lets all the redundant low-frequency information, present in the set $I^{\mathrm{LR}}_{[1,T]}$ of LR images, flow through the network, leaving the model focusing its computation only on high-frequency components. Indeed, high-frequency features are more informative for HR reconstruction, while LR images contain abundant low-frequency information that can directly be forwarded to the final output [7]. Finally, as the majority of the model for Single-image Super-resolution [5,6,8,15], all computations in our network are efficiently performed in the LR feature space requiring only an upsample operation at the final stage of the model.

In the following paragraphs, we present the overall architecture of the network with a detailed overview of the main blocks. Finally, we conclude the methodology section with precise details of the optimization process for training the network.

3.1. Network Architecture

An overview of the RAMS network, with its main three blocks and two branches, is depicted in Figure 1. As a high-level description, the model takes as input a single set of T low-resolution images $I^{\mathrm{LR}}_{[1,T]}$ that can be represented as a tensor $\mathbf{X}^{(i)}$ with shape $H \times W \times T \times C$ where H, W and C are the height, width, and channels of the single low-resolution images, respectively. The upper global residual path proposes a simple SR solution, making an aware fusion of the T input images. On the other hand, the central branch exploits 3D convolutions residual-based blocks in order to extract spatial and temporal correlations from the same set of T LR images and provide a refinement to the residual simple SR image.

More in detail, in the first part of the main path of the model, we use a simple 3D convolutional layer, with each filter of size $f_h \times f_w \times f_t$, to extract F shallow features from the input set $I^{\mathrm{LR}}_{[1,T]}$ of LR images. Then, we apply a cascade of N residual feature attention blocks that increasingly extract higher-level features, exploiting local and non-local, temporal, and spatial correlations. Moreover, we make use of a long skip connection for the shallow features and several short skip connections inside each feature attention block to let flow all redundant low-frequency signals and let the network focus on more valuable high-frequency components. The three dimensions H, W and T are always preserved through reflecting padding. The first part of the main branch can be modeled as a single function H_I that maps each tensor $\mathbf{X}^{(i)}$ to a new higher dimensional one $\mathbf{X}^{(i)}_I$ with shape $H \times W \times T \times F$:

$$\mathbf{X}^{(i)}_I = H_I(\mathbf{X}^{(i)}) \tag{2}$$

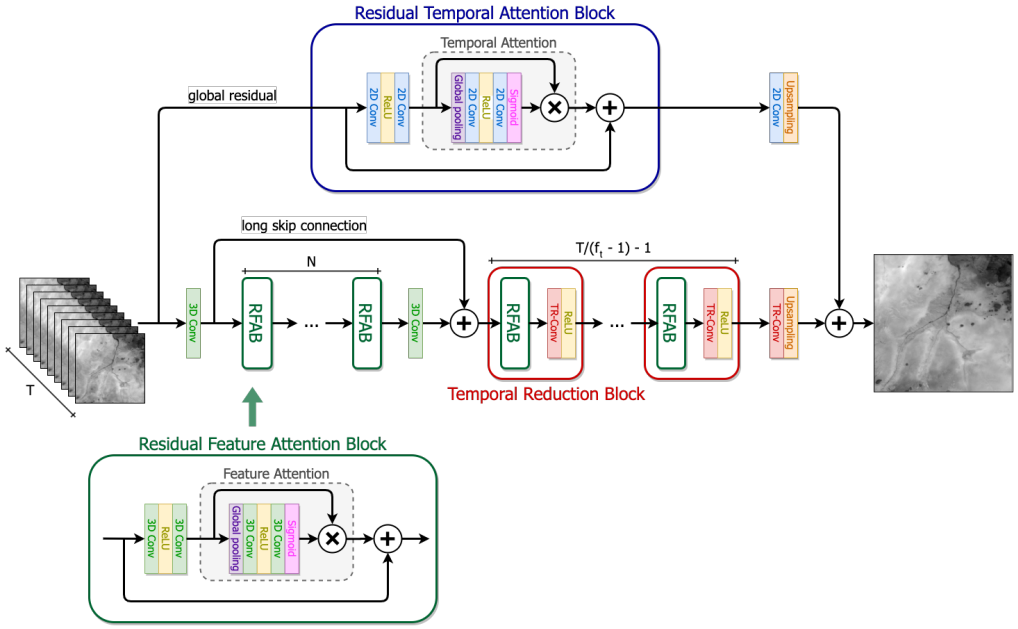

Figure 1. Overview of the Residual Attention Multi-image Super-resolution Network (RAMS), assuming to work with single-channel LR images ($C = 1$) to simplify the discussion. A tensor of T single-channel LR images constitutes the input of the proposed model. The main branch extracts features, with 3D convolutions, in a hierarchical fashion, while a feature attention mechanism allows the network to select and focus on most promising inner representations. Concurrently, a global residual path exploits a similar attention operation in order to make an aware fusion of the T distinct LR images. All computations are efficiently performed in the LR feature space and only at the last stage of the model an upsampling operation is performed in both branches.

In the second part of the main branch, we further process the output tensor $\mathbf{X}_I^{(i)}$ with $\lfloor T/(f_t - 1) \rfloor - 1$ temporal reduction blocks. In each block, we intersperse a residual feature attention block with 3D convolutional layer without padding on the temporal T dimension (TR-Conv). Therefore, H, W and F remain invariant and only the temporal dimension is reduced. The output of this second block is a new tensor $\mathbf{X}_{II}^{(i)}$ with shape $H \times W \times f_t \times F$, where the temporal dimension T is reduced to f_t:

$$\mathbf{X}_{II}^{(i)} = H_{II}(\mathbf{X}_I^{(i)}) \tag{3}$$

Finally, the output tensor $\mathbf{X}_{II}^{(i)}$ is processed by a further TR-Conv layer that reduces T to one and an upscale function $H_{\text{UP}|_{3D}}$ that generates a tensor $\mathbf{X}_{\text{UP}|_{3D}}^{(i)}$ of shape $sH \times sW \times C$ where s is the scaling factor.

The overall output $\mathbf{X}_{\text{UP}|_{3D}}^{(i)}$ of the main branch sums with the trivial solution provided by the global residual. Indeed, the global path simply weights the T LR images of the input tensor $\mathbf{X}^{(i)}$ with a residual temporal attention block with filters of size $f_h \times f_w$. Then it produces an output tensor $\mathbf{X}_{\text{UP}|_{2D}}^{(i)}$ of shape $sH \times sW \times C$ that is added to the one of the main branch. Therefore, the final SR prediction of the network $\hat{\mathbf{Y}}^{(i)} = I^{\text{SR}}$ is the sum of the two contributions:

$$\hat{\mathbf{Y}}^{(i)} = H_{\text{RAMS}}(\mathbf{X}^{(i)}) = (\mathbf{X}_{\text{UP}|_{3D}}^{(i)} + \mathbf{X}_{\text{UP}|_{2D}}^{(i)}) \tag{4}$$

The upscaling procedure is identical for both branches; after several trials with different methodologies, such as transposed convolutions [51], bi-linear resizing and nearest-neighbor upsampling [52], we adopted a sub-pixel convolution layer as explained in detail in [53]. Therefore, for either branch, the last 2D or 3D convolution generates $s^2 \cdot C$ features in order to produce the final tensors of shape $sH \times sW \times C$ for the residual sum.

In conclusion, the overall model takes as input a tensor $\mathbf{X}^{(i)}$ with shape $H \times W \times T \times C$, works always efficiently in the LR space and generates only at the final stage an output tensor $\hat{\mathbf{Y}}^{(i)}$ with shape $sH \times sW \times C$.

In the following sub-paragraphs, the three major functional blocks, residual feature attention, residual temporal attention, and temporal reduction blocks are further explained and analyzed.

3.2. Residual Attention Blocks

Residual attention blocks are at the core of the RAMS model; their specific architecture allows it to focus on relevant high-frequency components and let redundant, low-frequency information flow through the residual connections of the network. Inter-dependencies among features, in the case of feature attention blocks, or temporal steps, in the case of temporal attention blocks, are taken into account computing for each of them, relevant statistics that take into account local and non-local, temporal and spatial correlations. Indeed, either 3D or 2D convolution filters operate with local receptive fields loosing the possibility to exploit contextual information outside of their limited region of view.

3.2.1. Residual Feature Attention

Except for the global residual path, all residual attention blocks are residual feature attention blocks, as shown in Figure 1. Indeed, each block of features is weighted up in order to trace most promising high-frequency components, and a residual connection lets low-frequency information flow through the network.

More formally, the output of a residual feature attention block with a generic tensor, $\mathbf{X}_n^{(i)}$, is equal to:

$$F_{RFA}(\mathbf{X}_n^{(i)}) = \mathbf{X}_n^{(i)} + H_{FA}(\mathbf{X}_*^{(i)}) \cdot \mathbf{X}_*^{(i)} \tag{5}$$

where H_{FA} is the feature attention function and $\mathbf{X}_*^{(i)}$ is the output of two stacked 3D convolutional layers.

$$\mathbf{X}_*^{(i)} = W_2 * max(0, W_1 * \mathbf{X}_n^{(i)} + B_1) + B_2 \tag{6}$$

where W_1,W_2 and B_1, B_2 represent the filters with size $f_h \times f_w \times f_t$ and biases respectively and, '$*$' denotes the 3D convolution operation. The number of filters is always equal to F as the ones extracted by the first 3D convolutional layer.

Therefore, all low-frequency components in $\mathbf{X}_n^{(i)}$ can flow through the residual connection and H_{FA} can focus the attention of the network to more valuable high-frequency signals. To this end, the feature attention block takes the feature-wise global spatial and temporal information into a feature descriptor by using a global average pooling. Therefore, from the tensor $\mathbf{X}_*^{(i)}$ with shape $H \times W \times T \times F$ we extract $z_F \in \mathbb{R}^F$ feature statistics using the following equation:

$$z_F = \frac{1}{H \times W \times T} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{k=1}^{T} \mathbf{X}_*^{(i)}(i,j,k) \tag{7}$$

Statistics of the feature z_F can be viewed as a collection of descriptors, whose values contribute to express the whole stack of temporal images [54].

In Figure 2, it is possible to observe the global pooling operation which output is a tensor $\mathbf{Z}_F^{(i)}$ with shape $1 \times 1 \times 1 \times F$ and last dimension equal to z_F. In addition, the output tensor $\mathbf{Z}_F^{(i)}$ is further processed by a stack of two 3D convolutional layers with a ReLU [55] and sigmoid activation function, respectively. Indeed, as discussed in [54], the stack of two convolutional layers with the filter of size $1 \times 1 \times 1$ allows creating a non-linear mapping function which is able to deeply capture feature-wise dependencies from the aggregated information extracted by the global pooling operation. The first 3D convolutional layer reduces the feature size by a factor of r, and then the second layer restores

the original dimension and constraints its values from zero to one with a sigmoid function in a non-mutually exclusive relationship.

Finally, the original tensor $\mathbf{X}_*^{(i)}$ is weighted up by the processed attention statistics as shown in Equation (5).

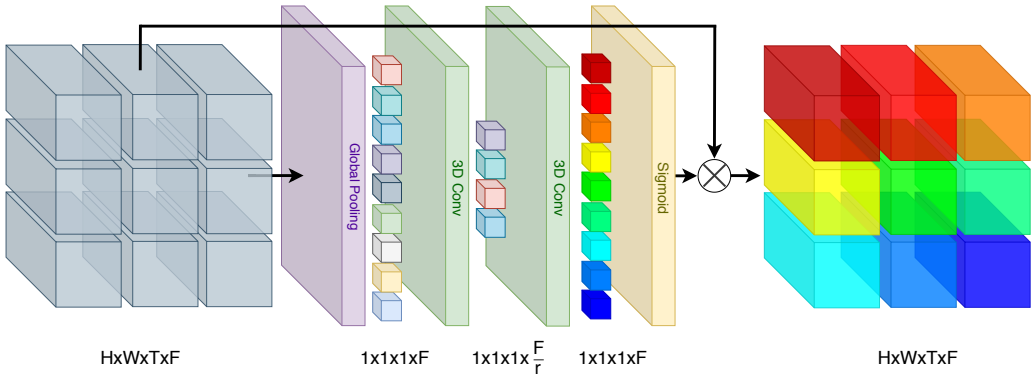

Figure 2. Reference architecture of a feature attention block. A series of convolutional operations and non-linear activations are applied to the input tensor with shape $H \times W \times T \times F$ in order to generate different attention statistics for each feature F that concurrently take advantage of local and non-local correlations. Consequently, each tensor's feature is properly re-scaled, enabling the network to focus on most promising components and letting residual connections heed of all redundant low-frequency signals.

3.2.2. Residual Temporal Attention

The primary purpose of the global residual path is to generate a starting trivial solution for the upsampling problem. More accurate is this starting prediction, and more simplified is the role of the main branch of the network, leading to a lower reconstruction error. However, the input of the model $\mathbf{X}^{(i)}$ has T different LR images that have to be merged. Intuitively, for each input sample $I^{\mathrm{LR}}_{[1,T]}$, there are some LR images more similar to each other. Therefore, giving them more relevance when merging the T LR images would most probably lead to higher quality predictions. In this context, the aim of the residual temporal attention block is to make an aware weighing of the different input temporal images, letting the network to make an upsample solution with primarily the most similar temporal steps. That is accomplished with an asymmetrical mechanism to the one employed in the residual feature attention blocks and can be summarized by the following formula:

$$F_{RTA}(\mathbf{X}^{(i)}) = \mathbf{X}^{(i)} + H_{TA}(\mathbf{X}_*^{(i)}) \cdot \mathbf{X}_*^{(i)} \tag{8}$$

where H_{TA} is the temporal attention function and $\mathbf{X}_*^{(i)}$ is the product of a stack of two 2D convolutional operations as depicted in Figure 3 with $f_h \times f_w$ and $T \cdot C$ as filter size and number of features, respectively. Then, as already introduced with the feature attention blocks, the temporal block takes the temporal-wise global spatial information into a feature descriptor by using a global average pooling operation. Finally, those statistical descriptors are processed by a stack of 2D convolutional layers with ReLU and sigmoid as activation function, respectively, scaling the $T \cdot C$ channels of the input tensor, as shown in Equation (8). As for feature attention blocks, the first convolutional layer reduces the number of the last dimension by a factor of r, giving the network the possibility to fully capture temporal-wise dependencies from the aggregated output information of the global average pooling operation.

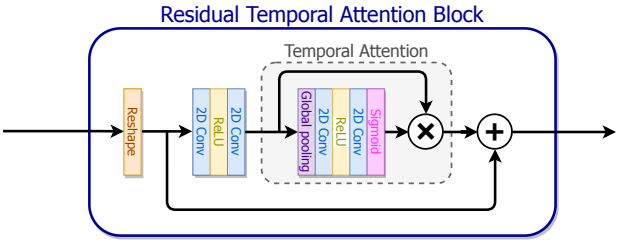

Figure 3. Reference architecture of a residual temporal attention block. If the number of channels $C \neq 1$ the input tensor $\mathbf{X}^{(i)}$ is reshaped in $H \times W \times (T \cdot C)$. Consequently, all temporal channels are weighted with some statistics computed by the layers of the temporal attention block.

3.3. Temporal Reduction Blocks

The aim of the last block of the main branch is to slowly reduce the number of temporal steps so that the temporal depth eventually reduces to one. Indeed, the output tensor $X_I^{(i)}$ of the N residual feature attention blocks has T temporal dimensions that need to be merged. To this end, we further process the incoming tensors with $\lfloor T/(f_t - 1) \rfloor - 1$ temporal reduction blocks. Each one is composed of a residual feature attention block and a 3D convolutional layer without any reflecting padding in the temporal dimension, denoted TR-Conv. Therefore, at each TR-Conv layer we reduce T of $f_t - 1$. The attention blocks allow the network to learn the best space to decouple image features, "highlighting" more promising features to maintain when reducing the temporal dimension. The output of the last temporal reduction block is a tensor $X_{II}^{(i)}$ with shape $H \times W \times f_t \times F$ where the temporal dimension T is reduced to f_t. The last TR-Conv, before the upsampling function $H_{\text{UP}|_{3D}}$, reduces to one the number of temporal steps and generates $s^2 \cdot C$ features for the sub-pixel convolutional layer.

3.4. Training Process

Learning the end-to-end mapping function H_{RAMS} requires the estimation of model parameters Θ. That is achieved by minimizing a loss $\mathcal{L}$ between the reconstructed super-resolved images I^{SR} and the corresponding ground truth high-resolution images I^{HR}.

Several loss functions were proposed and investigated for the SISR problem, such as L_1 [5,6,56,57], L_2 [4,11,20,50] and perceptual and adversarial losses [21,22]. However, in typical MISR remote-sensing problems, LR images are taken within a certain time window and they could have an undefined spatial misalignment one to each other. Therefore, we must take into account that the super-resolved output of the model I^{SR} will be inherently not registered with the target image I^{HR}. Moreover, since we can have very different conditions among the images part of the same scene, it is important to make the loss function independent from possible intensity biases between the super-resolved I^{SR} and the target I^{HR}. Indeed, if we get a super-resolved image $I^{\text{SR}} = I^{\text{HR}} + \epsilon$, with ϵ constant and low enough to avoid numerical saturation, we can consider its reconstruction perfect, since it represents the scene with the same level of detail of the ground truth.

With these premises, inspired by the metric proposed in [58], we defined $I^{\text{SR}}_{\text{crop}}$ as the super-resolved output cropped of d pixels on each border and we consider each possible patch $I^{\text{HR}}_{u,v}$, $u, v \in [0, 2d]$ of size $(sH - 2d) \times (sW - 2d)$ extracted from the ground truth I^{HR}. We compute the mean biases between the cropped $I^{\text{SR}}_{\text{crop}}$ and the patches $I^{\text{HR}}_{u,v}$ as follows:

$$\mathrm{b}_{u,v} = \frac{\sum_{i=1}^{sH-2d} \sum_{j=1}^{sW-2d} \left[I^{\text{HR}}_{u,v} - I^{\text{SR}}_{u,v} \right](i,j)}{(sH - 2d)(sW - 2d)} \tag{9}$$

The loss is then defined as the minimum mean absolute error (L_1 loss) between $I^{\text{SR}}_{\text{crop}}$ and each possible alignment patch $I^{\text{HR}}_{u,v}$. We use the MAE instead of the most used MSE since we experimentally find that provides better results for image restoration problems, as proved by the previous works [5,7,59].

$$\mathcal{L} = \min_{u,v \in [0,2d]} \frac{\| I^{HR}_{u,v} - (I^{SR}_{u,v} + b_{u,v}) \|_1}{(sH - 2d)(sW - 2d)} \tag{10}$$

where $\| \cdot \|_1$ represents the L_1 norm of a matrix, i.e., the sum of its absolute values.

4. Experiments and Discussion

In this section, we test the proposed methodology in an experimental context, training it on a dataset of real-world satellite images and evaluating its performance in comparison with other approaches, including a state-of-the-art SISR algorithm, to demonstrate the superiority of Multi-image models. We first present the dataset and the preprocessing stages, we define all the parameters used during the experimentation, and then we propose quantitative and qualitative results. We also perform an ablation study to demonstrate the contribution of the global residual branch that implements a temporal attention mechanism. To implement our network, we use the TensorFlow framework. The complete code with a pre-trained version of our model is available online (https://github.com/EscVM/RAMS).

4.1. The Proba-V Dataset

To train our model, we exploit the dataset released by the Advanced Concept Team of the European Space Agency (ESA) [58]. This dataset has been specifically conceived for MISR problems, and it is composed of several images taken by the Proba-V satellite (https://esa.int/Applications/Observing_the_Earth/Proba-V (accessed on 2 July 2020)) in the two different spectral bands RED and NIR (near-infrared). Proba-V satellite has been launched by ESA in 2013 and is specifically designed for land covering and vegetation growth monitoring across almost the entire globe. The satellite provides images in two resolutions with different revisit frequency. HR images have a 100 m per pixel spatial resolution and are released roughly every five days, while LR images have 300 m per pixel resolution and are available almost daily. The characteristics of the Proba-V imagery make it particularly suitable for MISR algorithms since it provides both resolutions natively, allowing for the application of the SR process without the need for artificially degrading and downsampling the HR images.

The dataset has been released for the Proba-V Super Resolution challenge (https://kelvins.esa.int/proba-v-super-resolution (accessed on 2 July 2020)) and is composed of two main parts: the train part provides both LR and HR images, while the test part LR images, only. In order to verify the effectiveness of our approach, we consider the train part and not the test part, since it has been conceived for the challenge evaluation only and it does not include the ground truths. Thus, we subdivide the train part in training and validation sets. To ease the comparison with previous methods, we use the same validation images used in [50]. In total, we have 415 scenes for training and 176 for validation for the RED band and 396 for training and 170 for validation for NIR.

Each scene is composed of several LR images (from 9 to 35, depending on the scene) with a dimension of 128 $\times$ 128 pixels and a single HR ground truth with a dimension of 384 $\times$ 384 pixels. The images are encoded as 16-bit png files, even if the actual signal bit-depth is 14 bits. Additionally, each image features a binary mask that distinguishes reliable pixels from unreliable ones (e.g., due to cloud coverage). This information is vital since the images are not taken in the same weather and temporal conditions, but a maximum period of 30 days can be covered in a single scene. For this reason, non-trivial changes in the landscape can occur between different LR images and their HR counterpart and are essential to understand which pixels carry meaningful information and which do not. Trying to infer the value of pixels that are concealed by clouds would mean being able to predict the weather in an unknown time by merely looking at the condition in other unspecified moments. For this reason, it is essential to train the network so that unreliable pixels do not influence the SR process. To assess the quality of each image, we define c as the clearance of the image, i.e., the fraction of reliable pixels in the correspondent binary mask.

4.2. Data Pre-Processing

Before training the model, we pre-process the dataset with the following steps:

- register each LR image using as reference the one with maximum clearance c
- select the clearest T images from each scene that are above a certain clearance threshold c_{min}
- pre-augment the training dataset with n_p temporal permutations of the LR input images
- normalize the images by subtracting the dataset mean intensity value and dividing by the standard deviation

Since each LR image is taken at a different time and with some intrinsic spatial misalignment with respect to the others, it is important to resample each pixel value in order to have a coherent reference frame. For each scene of the dataset, we consider as a reference image the one with the maximum clearance c. During the registration process, we consider translation as transformation model, which computes the necessary shifts to register each image for both the axes. Masks are taken into consideration during this process in order to avoid bad registration caused by unreliable pixels. The registration is performed in the Fourier domain using normalized cross-correlation as in [60]. After computing the shifts, both LR images and the correspondent masks are shifted accordingly. We use a reflect padding to add pixels to LR images and a constant zero padding for masks. In this way, these extra pixels will be considered unreliable.

For each scene, we must select some LR images in order to match the temporal dimension T of the network. We set a threshold $c_{\text{min}} = 0.85$ on the clearance for an image to be accepted to avoid using awful images that can worsen the SR performance. The acceptable images are then sorted in order of clearance, and the best T are selected. In the case of a scene with less than T images, we sample randomly from the set of acceptable images until T are reached. If a scene is only composed of clearances under c_{min}, it is entirely removed from the dataset. This selection process is performed after the registration so that heavily bad registered images are also removed, even if they had an initial clearance above the threshold. Since each scene of the dataset contains at least 9 LR images, we set $T = 9$ to fully exploit all the available information for most of the scenes.

One of the characteristics of the Proba-V dataset is that the LR images of a particular scene have no clear temporal order. Therefore, there is no reason to prefer a specific order in the T input images to another. The training dataset is, therefore, pre-augmented by performing n_p random temporal permutations of the selected T input images to help generalization. In this way, we can train the algorithm to identify the best temporal image independently on the position inside the input tensor. We set this permutation parameter to $n_p = 7$, reaching a total of 2905 training data-points for RED and 2751 for NIR.

Finally, each image is normalized by subtracting the mean pixel intensity value computed on the entire dataset and dividing by the standard deviation. After the forward pass in the network, all the tensors are then denormalized, and the subsequent evaluations are performed on the 16 bits unsigned integer arrays.

4.3. Experimental Settings

The scaling factor of the Proba-V dataset is $s = 3$. Since we have different scenes for RED and NIR data, we treat the problem for the two bands separately. For this reason, we have $C = 1$, since we consider images with a single channel. We set $F = 32$ and $f_h = f_w = f_t = 3$ as number of filters and kernel size respectively for each convolutional layer. Therefore, the number of temporal reduction blocks is $\lfloor T/(f_t - 1) \rfloor - 1 = 3$, since each block reduces the temporal dimension of 2. In all the residual attention blocks, we set $r = 8$ as the reduction factor. After testing different values with a grid search, we set $N = 12$ as the number of residual feature attention blocks in the main branch of the network. We find that decreasing this number causes a loss of performance while increasing it gives a little improvement in the results at the cost of a high increase in the number of parameters. $N = 12$ is,

therefore, the best compromise between network size and prediction results. In total, our model has slightly less than 1M parameters.

In most of the SR applications present in the literature, LR images are obtained from the artificial degradation of the target HR images. In contrast, the real-world nature of the dataset, in which LR images are obtained independently from HR images, causes an unavoidable misalignment between the super-resolved output and the ground truth. To take into account this problem, the authors of the dataset consider a maximum shift of ± 3 pixels on each axis between I^{SR} and target I^{HR}, computed on the basis of the geolocation accuracy of the Proba-V satellite [58]. When computing the loss function presented in Section 3.4, we can therefore set $d = 3$. Besides, since the Proba-V dataset also provides binary mask that marks with one reliable pixel and with 0 unreliable (e.g., concealed by clouds) ones, we adapt the loss function to use this information to refine the training process. During the loss computation, we want pixels marked as unreliable in the target binary mask M^{HR} not to influence the loss computation. Practically, we can simply multiply the cropped super-resolved image $I^{\mathrm{SR}}_{\mathrm{crop}}$, and the HR patch $I^{\mathrm{HR}}_{u,v}$ by the correspondent cropped mask $M^{\mathrm{HR}}_{u,v}$ and average all the quantities over the number of clear pixels. The bias computation is therefore adapted from Equation (9) as:

$$\mathrm{b}_{u,v} = \frac{\sum_{i,j} \left[I^{\mathrm{HR}}_{u,v} \cdot M^{\mathrm{HR}}_{u,v} - I^{\mathrm{SR}}_{u,v} \cdot M^{\mathrm{HR}}_{u,v} \right](i,j)}{\| M^{\mathrm{HR}}_{u,v} \|_1} \tag{11}$$

where $\| \cdot \|_1$ represents the L_1 norm of a matrix, i.e., the sum of its absolute values. In the same way, the loss is adapted from Equation (10) as:

$$\mathcal{L} = \min_{u,v \in [0,6]} \frac{\| I^{\mathrm{HR}}_{u,v} \cdot M^{\mathrm{HR}}_{u,v} - (I^{\mathrm{SR}}_{u,v} \cdot M^{\mathrm{HR}}_{u,v} + b_{u,v}) \|_1}{\| M^{\mathrm{HR}}_{u,v} \|_1} \tag{12}$$

To train the model, we extract from each LR image 16 patches with a size of 32×32 pixels and the corresponding HR and masks patches with a size of 96×96. We further check every single patch and remove those that have a target mask M^{HR} with less than 0.85 clearance. The total number of training data points obtained is 41,678 for RED and 40,173 for NIR. During the training process, we further perform data augmentation with random rotations of 90°, 180° and 270° and random horizontal flips.

We set the batch size to 32. Therefore, during training, we have an input tensor with shape $32 \times 32 \times 32 \times 9 \times 1$ and an output tensor with shape $32 \times 96 \times 96 \times 1$. We optimize the loss function with Adam algorithm [61] with default parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-7}$. We set an initial learning rate $\eta_{\mathrm{i}} = 5 \times 10^{-4}$ and we reduce it with a linear decay down to $\eta_{\mathrm{f}} = 5 \times 10^{-7}$. We train two different networks for RED and NIR spectral bands on a workstation with an Nvidia RTX 2080Ti GPU with 11GB of memory and 64GB of DDR4 SDRAM. We use the TensorFlow 2.0 framework with CUDA 10. In total, we train the models for 100 training epochs for about 16 hours.

4.4. Quantitative Results

To evaluate the obtained results, we need to use a slightly modified version of PSNR and SSIM [62] criteria to take into consideration all the aspects we considered in the previous section to obtain a proper loss function. Martens et al. [58] propose a corrected version of the PSNR, called cPSNS, which is obtained from a corrected mean squared error (cMSE). The computation of the cMSE is performed in the same way as we did for the loss in Equation (12): it is the minimum MSE between $I^{\mathrm{SR}}_{\mathrm{crop}} + b_{u,v}$ and the HR patches $I^{\mathrm{HR}}_{u,v}$:

$$\mathrm{cMSE} = \min_{u,v \in [0,6]} \underset{\mathrm{clear}}{\mathrm{MSE}} \left(I^{\mathrm{HR}}_{u,v}, I^{\mathrm{SR}}_{\mathrm{crop}} + b_{u,v} \right) \tag{13}$$

where $\underset{\mathrm{clear}}{\mathrm{MSE}}$ represents the mean squared error computed only on pixels marked as clear in the binary mask $M^{\mathrm{HR}}_{u,v}$. Again, we can simply multiply the matrices by the mask to make unreliable pixels irrelevant:

$$\underset{\text{clear}}{\text{MSE}} = \frac{\| I^{\text{HR}}_{u,v} \cdot M^{\text{HR}}_{u,v} - (I^{\text{SR}}_{u,v} \cdot M^{\text{HR}}_{u,v} + b_{u,v}) \|_2^2}{\| M^{\text{HR}}_{u,v} \|_1} \tag{14}$$

where $\| \cdot \|_2$ represents the Frobenius (L_2) norm of a matrix, i.e., the square root of the sum of its squared values. We can then compute the cPSNR as:

$$\begin{aligned} \text{cPSNR} &= 10 \log_{10} \frac{(2^{16}-1)^2}{\text{cMSE}} \\ &= \max_{u,v \in [0,6]} 10 \log_{10} \frac{(2^{16}-1)^2}{\underset{\text{clear}}{\text{MSE}}(I^{\text{HR}}_{u,v}, I^{\text{SR}}_{\text{crop}} + b_{u,v})} \end{aligned} \tag{15}$$

where $2^{16}-1$ is the maximum pixel intensity for an image encoded on 16 bits.

In the same way, we can define a corrected version of the SSIM metric: cSSIM is the maximum SSIM between $I^{\text{SR}}_{\text{crop}} + b_{u,v}$ and the HR patches $I^{\text{HR}}_{u,v}$, again multiplied for the mask $M^{\text{HR}}_{u,v}$.

$$\text{cSSIM} = \max_{u,v \in [0,6]} \text{SSIM}\left(I^{\text{HR}}_{u,v} \cdot M^{\text{HR}}_{u,v}, I^{\text{SR}}_{\text{crop}} \cdot M^{\text{HR}}_{u,v} + b_{u,v}\right) \tag{16}$$

4.4.1. Temporal Self-Ensemble (RAMS+)

As in Section 4.2, during the training process images are augmented with random permutation in the temporal axis. For this reason, it is possible to maximize the performance of the model, by adopting a self-ensemble mechanism during inference, similarly to what done in previous super-resolution works [5,7,63]. For each input scene, we consider a certain number P of random permutations on the temporal axis and we denote as $\left\{ I^{\text{LR}}_{[1,T],0}, \cdots, I^{\text{LR}}_{[1,T],P} \right\}$ the resulting set of inputs. The output of the inference process is therefore the average of the predictions on the whole set. We call this methodology RAMS+$_P$, where P is the number of random permutations performed:

$$I^{\text{SR}} = \frac{1}{P} \sum_{i=1}^{P} H_{\text{RAMS}}\left(I^{\text{LR}}_{[1,T],i}\right) \tag{17}$$

Figure 4 shows cPSNR results on the testing dataset for a different number of permutated predictions. The trend clearly shows how increasing P results in better performance on both the spectral bands, with an effect that tends to saturate for $P \geq 25$. For the following evaluation, we select $P = 20$ to present the results for RAMS+. It is worth noting that even if this method allows increasing the performance of the network sharply, inference time grows linearly with P, with RAMS+$_{20}$ taking roughly 20 times as long as RAMS. Another aspect to highlight is that the permutations are performed randomly, so different results can be achieved even with the same value of P.

4.4.2. Comparison with State-of-The-Art Methods

Table 1 shows the comparison of cPSNR and cSSIM metrics with several methods on the validation set. We consider as the baseline the bicubic interpolation of the best image of the scene selected considering the clearance, i.e., the number of clear pixels as marked by the binary masks.

IBP [13] and BTV [43] methods are tested with the same methodology presented in Molini et al. [50]. They achieve slightly better results than the baseline with both the metrics.

RCAN [7] is currently one of the Single-image Super-resolution state-of-the-art networks. We trained from scratch two models, one for each spectral band, setting $G = 5$ and $B = 5$, as the number of residual groups and residual channel attention blocks respectively, for a total of about 2 million parameters. We train the two models from scratch on the Proba-V dataset, selecting the best image per scene as input. RCAN shows better performance with respect to classical methods but is far

beyond the other MISR networks, showing how the additional information coming from both spatial and temporal correlations is vital to boost the super-resolution process.

VSR-DUF [47] has been developed to upsample video signals using a temporal window of several frames. We train two models from scratch, one for each spectral bands, using 9 LR images as input as in our methodology. The authors consider three different architectures depending on the number of convolutional layers and find better results, increasing the depth of the model. We select the baseline 16 layers deep architecture that already has more than double parameters with respect to RAMS, with the same number of input images.

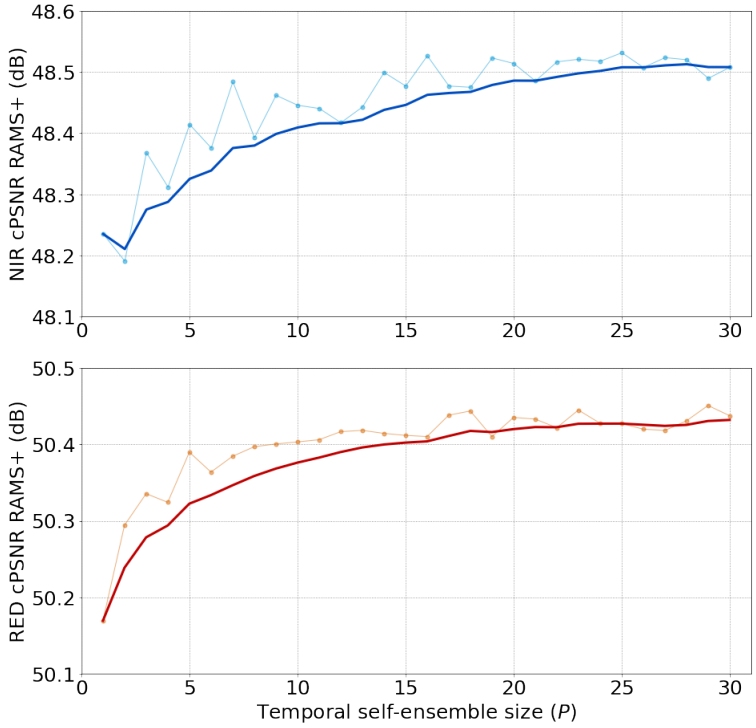

Figure 4. Results with a temporal self-ensemble of size P. The highlighted curves represent an exponential moving average of the results to clearly show the trend. The values for $P = 1$ are equivalent to RAMS.

HighRes-net [51] algorithm got the second place in the Proba-V challenge and featured a single network for both spectral bands that recursively reduce the temporal dimension to fuse the input LR images. We train the model on our training dataset with default architectures. Since the authors designed the architecture to have an input temporal dimension multiple of 2, we set it to 16, as it is closest to 9.

DeepSUM [50] is the algorithm winner of the original Proba-V challenge, and the authors have further developed it with DeepSUM++ [64]. We train our RAMS on the same training dataset as these two works.

Results clearly show how the proposed methodology can obtain the best results with the two metrics on both the spectral bands and thus represents the current state-of-the-art for Multi-image Super-resolution for remote sensing applications. Using temporal self-ensemble, RAMS+ is able to achieve even higher performance. We show the value for RAMS+, setting $P = 20$ as the size of the ensemble, which is the value at which we experimentally find that the resulting gain starts to saturate. However, further increasing the ensemble size can result in even better performance, though at the cost of a significant inference speed drop.

It is worth mentioning that our methodology reaches a result of 0.9336790819983855 on the test set of the Proba-V challenge as provided by the official site and places at the top of the leaderboard

available after the end of the official challenge (https://kelvins.esa.int/proba-v-super-resolution-post-mortem/leaderboard (accessed on 2 July 2020)). This score is computed as the mean ratio between the cPSNR values of the challenge baseline on each testing scene, and the correspondent submitted cPSNR for both the spectral bands. This result has been obtained by retraining the two networks with both training and validation datasets together.

Figure 5 shows a direct comparison between the cPSNR results of RAMS and the bicubic interpolation baseline and RCAN (SISR state-of-the-art). Each cross represents a scene of the validation dataset of the corresponding spectral band. The graphs on the left show how our method strongly beats the bicubic upsampling on almost all the scenes, 98% for RED and 91% for NIR. That is coherent with a general worse behavior of all the methods on the NIR images, probably due to an intrinsic worse information quality of the NIR dataset. The graphs on the right show, on the other hand, the enormous potential of MISR with respect to SISR methods. It can be observed how again RAMS outperforms RCAN an almost all the scenes, with results only slightly worse than to bicubic interpolation, 92% for RED, and 91% for NIR. That is reasonable since RCAN results are someway in the middle between bicubic and RAMS.

Table 1. Average cPSNR (dB) and cSSIM over the validation dataset for different methods. Our solution is highlighted in bold at the end of the table.

Band	**NIR**		**RED**	
Metric	**cPSNR**	**cSSIM**	**cPSNR**	**cSSIM**
Bicubic	45.12	0.9767	47.63	0.9846
IBP [13]	45.96	0.9796	48.21	0.9865
BTV [43]	45.93	0.9794	48.12	0.9861
RCAN [7]	45.66	0.9798	48.22	0.9870
VSR-DUF [47]	47.20	0.9850	49.59	0.9902
HighRes-net [51]	47.55	0.9855	49.75	0.9904
DeepSUM [50]	47.84	0.9858	50.00	0.9908
DeepSUM++ [64]	47.93	0.9862	50.08	0.9912
RAMS (ours)	**48.23**	**0.9875**	**50.17**	**0.9913**
RAMS+$_{20}$ (ours)	**48.51**	**0.9880**	**50.44**	**0.9917**

4.4.3. Importance of the Residual Temporal Attention Branch

As a final analysis, we perform an ablation study to demonstrate the importance of the global residual branch that implements a temporal attention mechanism. We train two alternative networks, one for each spectral band, that have the same architecture of RAMS, except that we delete the residual temporal attention (RTA) branch. These reduced networks are trained from scratch independently from the complete ones.

Table 2 shows a significant drop in the results obtained without the global residual branch and demonstrates the importance of selecting the best temporal views to ease the super-resolution process of the main branch. We find this difference particularly relevant for the RED band, since the training repeatedly failed without the RTA branch, with a diverging behavior after some epochs. The result reported in the table is computed with the last valuable parameters before the divergence starts.

Table 2. RAMS results with and without RTA (residual temporal attention) branch. Values for RED without RTA, highlighted with the asterisks, are computed with the last valuable parameters before training diverges.

	without RTA		**with RTA**	
	cPSNR	**cSSIM**	**cPSNR**	**cSSIM**
NIR	47.96	0.9869	48.23	0.9875
RED	47.98 *	0.9863 *	50.17	0.9913

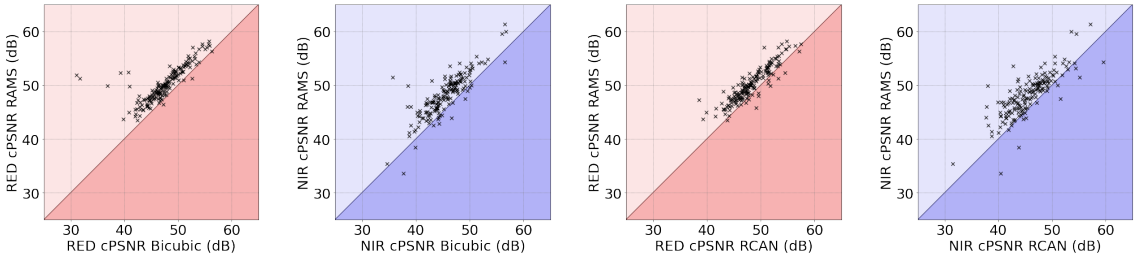

Figure 5. cPSNR comparison between RAMS and bicubic inteprolation and RAMS and RCAN(SISR) on the validation set. Each data point represents a scene of the dataset: when a cross is above the line, the correspondent scene is reconstructed better by RAMS.

4.5. Qualitative Results

A visual comparison between some of the methods taken in the exam is shown in Figures 6 and 7 for a RED and NIR image respectively. We provide a zoomed patch of the best LR input image of the scene, its bicubic interpolation, and the inference output of RCAN, VSR-DUF, DeepSUM, RAMS and RAMS+$_{20}$, together with the target HR ground truth.

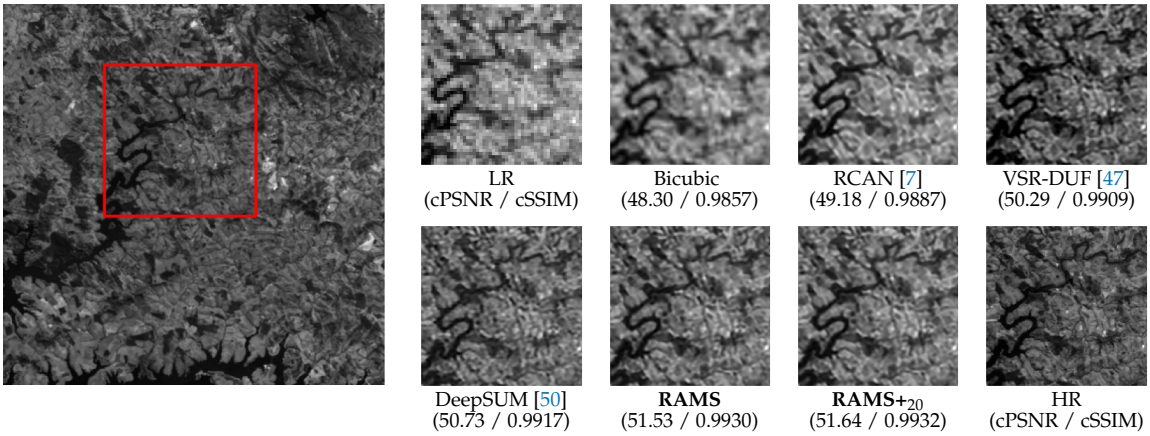

Figure 6. Qualitative comparison between different methods on RED imgset0302.

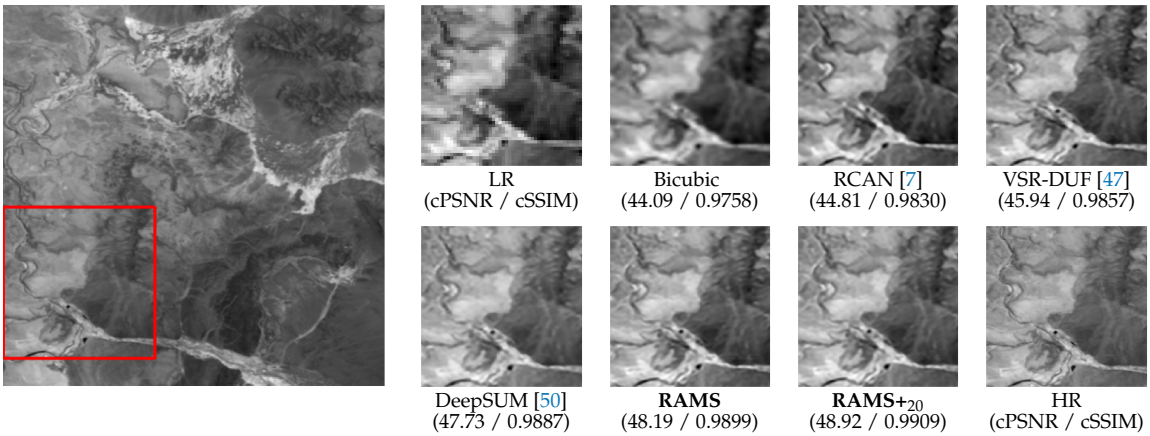

Figure 7. Qualitative comparison between different methods on NIR imgset0596.

cPSNR and cSSIM scores for the image under analysis are also provided. From this comparison, MISR methods clearly show a better performance with respect to bicubic and SISR (RCAN). However, it is not trivial to understand which method is the better among MISR algorithms with a visual inspection of the results, only. As found by Ledig et al. [21], the task of achieving pleasant-looking results is a different optimization problem from maximizing the fidelity of the

reconstructed information. Therefore, results with high content-related metrics as PSNR and SSIM frequently appear less photo-realistic to a human eye. However, in the context of remote sensing, the fidelity of the pixels content is vital to ensure that the super-resolved image are meaningful, thus the quality of results should be inferred by using content-related metrics, rather than by visual inspection.

5. Conclusions

In this paper, we proposed a novel representation learning model to super-resolve remotely sensed multi-temporal LR images by exploiting concurrently spatial and temporal correlations. We introduced feature and temporal attention mechanisms with 3D convolutions that, coupled with nestled residual connections, let the network focus on high-frequency components, flow redundant low-frequency information and transcend the local region of convolutional operations. Extensive experiments on the open-source Proba-V MISR dataset, either with single image and multi-image SR methodologies, demonstrated the effectiveness of our proposed methodology. In both NIR and RED spectral bands, our efficient and straightforward solution achieved considerably better results than other literature methodologies obtaining 48.51 dB and 50.44 dB of cPSNR, respectively for the two channels. That is further proved by the score of the official post-mortem Prova-V challenge where RAMS claimed the first place in the leaderboard. Future work may investigate the performance of the RAMS architecture on hyperspectral remote sensing imaging.

Author Contributions: Conceptualization, V.M., M.C.; Methodology, V.M.; Software, F.S. and V.M.; Validation, F.S. and V.M.; Data curation, F.S. and V.M.; Writing-original draft, F.S., V.M., and A.K.; Writing-review and editing, F.S., V.M., A.K. and M.C.; Project administration, V.M. and M.C.; Funding acquisition, M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: This work has been developed with the contribution of the Politecnico di Torino Interdepartmental Centre for Service Robotics PIC4SeR (https://pic4ser.polito.it) and SmartData@Polito (https://smartdata.polito.it).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Valsesia, D.; Magli, E. A novel rate control algorithm for onboard predictive coding of multispectral and hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 6341–6355.
2. Benecki, P.; Kawulok, M.; Kostrzewa, D.; Skonieczny, L. Evaluating super-resolution reconstruction of satellite images. *Acta Astronaut.* **2018**, *153*, 15–25.
3. Valsesia, D.; Boufounos, P.T. Universal encoding of multispectral images. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 4453–4457.
4. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295–307.
5. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 136–144.
6. Yu, J.; Fan, Y.; Yang, J.; Xu, N.; Wang, Z.; Wang, X.; Huang, T. Wide activation for efficient and accurate image super-resolution. *arXiv* **2018**, arXiv:1808.08718.
7. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 286–301.
8. Dai, T.; Cai, J.; Zhang, Y.; Xia, S.T.; Zhang, L. Second-order attention network for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, California, CA USA, 16–20 June 2019; pp. 11065–11074.

9. Borman, S.; Stevenson, R.L. Super-resolution from image sequences-a review. In Proceedings of the 1998 Midwest Symposium on Circuits and Systems (Cat. No. 98CB36268), Notre Dame, Indiana, 9–12 August 1998; pp. 374–378.
10. Tsai, R. Multiframe image restoration and registration. *Adv. Comput. Vis. Image Process.* **1984**, *1*, 317–339.
11. Kim, S.; Bose, N.K.; Valenzuela, H.M. Recursive reconstruction of high resolution image from noisy undersampled multiframes. *IEEE Trans. Acoust. Speech Signal Process.* **1990**, *38*, 1013–1027.
12. Irani, M.; Peleg, S. Super resolution from image sequences. In Proceedings of the 10th International Conference on Pattern Recognition, Atlantic City, NJ USA, 16–21 June 1990; Volume 2, pp. 115–120.
13. Irani, M.; Peleg, S. Improving resolution by image registration. *CVGIP* **1991**, *53*, 231–239.
14. Irani, M.; Peleg, S.; others. Motion analysis for image enhancement: Resolution, occlusion, and transparency. *J. Vis. Commun. Image Represent.* **1993**, *4*, 324–335.
15. Dong, C.; Loy, C.C.; Tang, X. Accelerating the super-resolution convolutional neural network. In *European Conference on Computer Vision*; Springer: New York, NY, USA, 2016; pp. 391–407.
16. Kim, J.; Kwon Lee, J.; Mu Lee, K. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 27-30 June, 2016; pp. 1646–1654.
17. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*; Springer: New York, NY, USA, 2014; pp. 184–199.
18. Kim, J.; Kwon Lee, J.; Mu Lee, K. Deeply-recursive convolutional network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 27–30 June 2016; pp. 1637–1645.
19. Tai, Y.; Yang, J.; Liu, X. Image super-resolution via deep recursive residual network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3147–3155.
20. Tai, Y.; Yang, J.; Liu, X.; Xu, C. Memnet: A persistent memory network for image restoration. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 4539–4547.
21. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
22. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*; Springer: New York, NY, USA, 2016; pp. 694–711.
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 27–30 June 2016; pp. 770–778.
24. Sajjadi, M.S.; Scholkopf, B.; Hirsch, M. Enhancenet: Single image super-resolution through automated texture synthesis. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 4491–4500.
25. Lanaras, C.; Bioucas-Dias, J.; Galliani, S.; Baltsavias, E.; Schindler, K. Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 305–319.
26. Liebel, L.; Körner, M. Single-image super resolution for multispectral remote sensing data using convolutional neural networks. *ISPRS-Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *41*, 883–890.
27. Lei, S.; Shi, Z.; Zou, Z. Super-resolution for remote sensing images via local–global combined network. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1243–1247.
28. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Change Loy, C. Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
29. Ma, W.; Pan, Z.; Yuan, F.; Lei, B. Super-Resolution of Remote Sensing Images via a Dense Residual Generative Adversarial Network. *Remote Sens.* **2019**, *11*, 2578.
30. Dong, X.; Xi, Z.; Sun, X.; Gao, L. Transferred Multi-Perception Attention Networks for Remote Sensing Image Super-Resolution. *Remote Sens.* **2019**, *11*, 2857.
31. Gargiulo, M.; Mazza, A.; Gaetano, R.; Ruello, G.; Scarpa, G. Fast Super-Resolution of 20 m Sentinel-2 Bands Using Convolutional Neural Networks. *Remote Sens.* **2019**, *11*, 2635.

32. Wu, J.; He, Z.; Hu, J. Sentinel-2 Sharpening via Parallel Residual Network. *Remote Sens.* **2020**, *12*, 279.
33. Chang, Y.; Luo, B. Bidirectional Convolutional LSTM Neural Network for Remote Sensing Image Super-Resolution. *Remote Sens.* **2019**, *11*, 2333.
34. Gu, J.; Sun, X.; Zhang, Y.; Fu, K.; Wang, L. Deep Residual Squeeze and Excitation Network for Remote Sensing Image Super-Resolution. *Remote Sens.* **2019**, *11*, 1817.
35. Yue, L.; Shen, H.; Li, J.; Yuan, Q.; Zhang, H.; Zhang, L. Image super-resolution: The techniques, applications, and future. *Signal Process.* **2016**, *128*, 389–408.
36. Elad, M.; Hel-Or, Y. A fast super-resolution reconstruction algorithm for pure translational motion and common space-invariant blur. *IEEE Trans. Image Process.* **2001**, *10*, 1187–1193.
37. Stark, H.; Oskoui, P. High-resolution image recovery from image-plane arrays, using convex projections. *JOSA A* **1989**, *6*, 1715–1726.
38. Lertrattanapanich, S.; Bose, N.K. High resolution image formation from low resolution frames using Delaunay triangulation. *IEEE Trans. Image Process.* **2002**, *11*, 1427–1441.
39. Takeda, H.; Farsiu, S.; Milanfar, P. Kernel regression for image processing and reconstruction. *IEEE Trans. Image Process.* **2007**, *16*, 349–366.
40. Shen, H.; Ng, M.K.; Li, P.; Zhang, L. Super-resolution reconstruction algorithm to MODIS remote sensing images. *The Comput. J.* **2009**, *52*, 90–100.
41. Kato, T.; Hino, H.; Murata, N. Double sparsity for multi-frame super resolution. *Neurocomputing* **2017**, *240*, 115–126.
42. Schultz, R.R.; Stevenson, R.L. Extraction of high-resolution frames from video sequences. *IEEE Trans. Image Process.* **1996**, *5*, 996–1011.
43. Farsiu, S.; Robinson, M.D.; Elad, M.; Milanfar, P. Fast and robust multiframe super resolution. *IEEE Trans. Image Process.* **2004**, *13*, 1327–1344.
44. Arbelaez, P.; Maire, M.; Fowlkes, C.; Malik, J. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 898–916.
45. Kappeler, A.; Yoo, S.; Dai, Q.; Katsaggelos, A.K. Video super-resolution with convolutional neural networks. *IEEE Trans. Comput. Imaging* **2016**, *2*, 109–122.
46. Caballero, J.; Ledig, C.; Aitken, A.; Acosta, A.; Totz, J.; Wang, Z.; Shi, W. Real-time video super-resolution with spatio-temporal networks and motion compensation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4778–4787.
47. Jo, Y.; Wug Oh, S.; Kang, J.; Joo Kim, S. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Utah, USA, 18–22 June 2018; pp. 3224–3232.
48. Kawulok, M.; Benecki, P.; Piechaczek, S.; Hrynczenko, K.; Kostrzewa, D.; Nalepa, J. Deep learning for multiple-image super-resolution. *IEEE Geosci. Remote Sens. Lett.* **2019**.
49. Kawulok, M.; Benecki, P.; Kostrzewa, D.; Skonieczny, L. Evolving imaging model for super-resolution reconstruction. In Proceedings of the Genetic and Evolutionary Computation Conference Companion, Koyoto, Japan, 15–19 July 2018; pp. 284–285.
50. Molini, A.B.; Valsesia, D.; Fracastoro, G.; Magli, E. DeepSUM: Deep neural network for Super-resolution of Unregistered Multitemporal images. *IEEE Trans. Geosci. Remote Sens.* **2019**, 58(5), 3644-3656.
51. Deudon, M.; Kalaitzis, A.; Goytom, I.; Arefin, M.R.; Lin, Z.; Sankaran, K.; Michalski, V.; Kahou, S.E.; Cornebise, J.; Bengio, Y. HighRes-net: Recursive Fusion for Multi-Frame Super-Resolution of Satellite Imagery. *arXiv* **2020**, arXiv:2002.06460.
52. Van Noord, N.; Postma, E. A learned representation of artist-specific colourisation. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 2907–2915.
53. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1874–1883.
54. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Utah, USA, 18–22 June 2018; pp. 7132–7141.

55. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010; pp. 807–814.
56. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep laplacian pyramid networks for fast and accurate super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 624–632.
57. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2599–2613.
58. Märtens, M.; Izzo, D.; Krzic, A.; Cox, D. Super-resolution of PROBA-V images using convolutional neural networks. *Astrodynamics* **2019**, *3*, 387–402.
59. Zhao, H.; Gallo, O.; Frosio, I.; Kautz, J. Loss functions for image restoration with neural networks. *IEEE Trans. Comput. Imaging* **2016**, *3*, 47–57.
60. Padfield, D. Masked object registration in the Fourier domain. *IEEE Trans. Image Process.* **2011**, *21*, 2706–2718.
61. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
62. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612.
63. Timofte, R.; Rothe, R.; Van Gool, L. Seven ways to improve example-based single image super resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 27–30 June 2016; pp. 1865–1873.
64. Molini, A.B.; Valsesia, D.; Fracastoro, G.; Magli, E. DeepSUM++: Non-local Deep Neural Network for Super-Resolution of Unregistered Multitemporal Images. *arXiv* **2020**, arXiv:2001.06342.