Remote Sensing Image Super-Resolution Using Deep Residual Channel Attention

Juan M. Haut, *Student Member, IEEE,* Ruben Fernandez-Beltran, Mercedes E. Paoletti, *Student Member, IEEE,* Javier Plaza, *Senior Member, IEEE,* and Antonio Plaza, *Fellow, IEEE*

***Abstract*—The current trend in remote sensing image super-resolution (SR) is to use supervised deep learning models to effectively enhance the spatial resolution of airborne and satellite-based optical imagery. Nonetheless, the inherent complexity of these architectures/data often makes these methods very difficult to train. Despite these recent advances, the huge amount of network parameters that must be fine-tuned and the lack of suitable high-resolution remotely sensed imagery in actual operational scenarios still raise some important challenges that may become relevant limitations in existent Earth observation data production environments. To address these problems, we propose a new remote sensing SR approach that integrates a visual attention mechanism within a residual-based network design in order to allow the SR process to focus on those features extracted from land-cover components that require more computations to be super-resolved. As a result, the network training process is significantly improved, because it aims at learning the most relevant high-frequency information while the proposed architecture allows neglecting the low-frequency features extracted from spatially uninformative Earth surface areas by means of several levels of skip connections. Our experimental assessment, conducted using the UC Merced and GaoFen-2 remote sensing image collections, three scaling factors, and eight different SR methods, demonstrates that our newly proposed approach exhibits competitive performance in the task of super-resolving remotely sensed imagery.**

***Index Terms*—Remote sensing, single-image super-resolution, deep learning, visual attention.**

I. Introduction

Over the past years, super-resolution (SR) techniques have become very helpful owing to their great potential to overcome the physical resolution constraints of remote sensing imaging sensors [1]. As a result, many of the most important operational satellites are currently focused on providing super-resolved data products, mainly because this kind of technology is able to generate enhanced remotely sensed imagery which are very useful to deal with current and future challenges and societal needs [2], [3]. For instance, fine-grained satellite image classification [4], [5], hyperspectral remote sensing data analysis [6]–[9], remote target identification [10], [11] and detailed land cover mapping [12]–[14] are some of the most popular remote sensing applications in which SR has provided important advantages.

This work has been supported by the Spanish Ministry (FPU14/02012-FPU15/02090), Junta de Extremadura (GR15005) and Generalitat Valenciana (APOSTD/2017/007). Funding from MINECO project TIN2015-63646-C5-5-R and from European Union under Grant 734541-EXPOSURE are gratefully acknowledged. *(Corresponding author: Juan M. Haut)*

J. M. Haut, M. E. Paoletti, J. Plaza and A. Plaza are with the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, 10003 Cáceres, Spain.(e-mail: juanmariohaut@unex.es; mpaoletti@unex.es; jplaza@unex.es; aplaza@unex.es). R. Fernandez-Beltran is with the Institute of New Imaging Technologies, University Jaume I, 12071 Castellón de la Plana, Spain. (e-mail: rufernan@uji.es).

Broadly speaking, SR [15]–[17] refers to those algorithmic tools aimed at increasing the spatial resolution of remotely sensed data while providing spatial information beyond the sensor resolution, that is, spatial details not present in the lower-resolution (LR) image captured by the sensing instrument. In the literature, it is possible to distinguish between two different trends that have been successfully adopted to super-resolve remotely sensed data: multi-image SR [18] and single-image SR [19]. Whereas multi-image techniques require several multi-angular shots within a very short time span, single-image SR offers a more flexible scheme for remote sensing applications, because the SR process is conducted using a single LR image of the target scene. In particular, there are two main factors that make single-image SR methods particularly attractive in the context of remote sensing applications. On the one hand, many of the currently operational satellites can only offer a revisiting period of at least several days [20], which does not allow using the straightforward multi-image SR approach, because of the existing temporal gap among different Earth observations. On the other hand, single-image SR can be applied without the need of using any satellite constellation, which eventually results in substantial cost savings and provides a good opportunity for small platforms, with low resolution and cheap instruments [21].

When focusing on the single-image SR domain [22], it is also possible to identify two different kinds of techniques, depending on the required training data: unsupervised and supervised methods. Regarding the unsupervised category, these SR approaches estimate the high-resolution (HR) details present in the super-resolved output from the LR input image itself. One of the simplest unsupervised SR methods was developed by Irani and Peleg in [23], where several back-propagation iterations were applied to gradually enhance the gradient of the up-scaled LR image. Since then, other more advanced unsupervised SR methods have been successfully applied to remotely sensed data. This is the case of the work published in [24], where the authors present an innovative self-learning procedure based on regularized patch-search criteria across scales to generate the corresponding super-resolved result. Another relevant work is [25], where the authors adopt a generative neural network to address the SR problem from an unsupervised perspective. Despite the evident benefits of not using any external training set, the performance of the unsupervised SR approach typically becomes rather limited under the most challenging remote sensing scenarios, because of the limited spatial information present in the LR input image. Note that remote sensing images are usually fully focused multi-band shots with plenty of complex spatial details, which makes the SR process particularly challenging [19].

In this sense, supervised methods are able to provide a more robust SR scheme by learning the relationships between LR and HR image domains by means of an external training set. One of the most popular supervised SR methods was introduced by Yang *et al.* [26], and it was later adapted to remote sensing problems in [27]. These approaches take advantage of the fact that natural images tend to be sparse when they are represented as a linear combination of small patches. Therefore, it is possible to learn a SR mapping by forcing the LR and HR training images to share the same sparse codes. Alternative works, such as [28]–[30], follow a similar idea but using different image characterization spaces which are able to provide specific advantages. Nonetheless, convolutional neural networks (CNN) represent certainly one of the most important paradigms within the supervised SR field, due to their great potential to uncover high-level features from optical data. Hence, multiple authors have successfully presented different methods based on CNNs. For instance, Dong *et al.* [31]

proposed a deep learning architecture to super-resolve LR images. More specifically, this method initially upscales the input LR patches by means of a bi-cubic interpolation, and then uses a 3-layer CNN to learn the mapping between the LR and HR image domains. Other authors have introduced additional improvements over this baseline work in order to achieve superior results. For instance, a relevant extension is presented in [32], where the authors define a deeper architecture which reduces the input feature space (and also removes the initial interpolation step) by providing an actual end-to-end mapping. Another important work is the one described in [33], where Kim *et al.* propose a 20-layer CNN architecture which considers image residuals together with data augmenting and multi-scaling learning schemes. Despite the remarkable performance achieved by all these methods when considering standard images, the special complexity of airborne and space-borne optical data usually limits their SR performance in several remote sensing tasks.

Consequently, other CNN-based SR methods have been designed to specifically manage remotely sensed imagery. For instance, Lei *et al.* [34] define a multi-level CNN architecture able to capture multi-scale features, which allow the network to simultaneously take into account local and global image features when introducing new spatial details in the SR process. Another relevant work was presented in [35], where the authors introduce several improvements on the network design in order to effectively super-resolve remotely sensed data. Specifically, residual units and skip connections were adopted to uncover more relevant features on both local and global image areas. Additionally, the super-resolved image reconstruction process was conducted using a network-in-network architecture [36], which improved model discriminability for different image features. In spite of all the efforts directed to designing highly accurate CNN-based SR models for remotely sensed data, many of the existing approaches still face challenges related to the convergence of network parameters, which eventually leaves room for improvement, especially when dealing with challenging remotely sensed data. Note that the most advanced deep learning SR models are very difficult to train because of their own complexity, and also because of the lack of significant training data [37], which may become an important limitation in some pre-operational airborne and space-borne optical acquisition scenarios.

With all these considerations in mind, this paper presents a new supervised SR network architecture that is especially designed to effectively super-resolve remotely sensed imagery. Some of the most recent CNN-based SR methods used in remote sensing applications assume that all the features extracted from the LR input image are equally important [34], [35]. This fact is fundamentally due to the behaviour of the convolutional kernel itself, where a sliding weight window (defined by the receptive field) is equally applied to the entire volume of data. However, this assumption may result in a lack of flexibility when analyzing the different kinds of features that are typically present in aerial shots. While the features extracted from smoother areas in the surface of the Earth are not expected to incorporate many HR spatial details, the SR process itself is mainly focused on enhancing the most textured areas, where the corresponding features are expected to introduce new high-frequency information. In this scenario, our newly proposed SR approach adopts a visual attention mechanism [38]–[40] that guides the network training process towards the most informative features, thus focusing the attention of the model on those Earth surface features related to structural components which require finer HR details. Note that convolutional kernels are able to capture specific land-cover features from the input data. As a result, the considered attention mechanism can provide competitive advantages to super-resolve remote sensing data, since the network filters are able to inherently involve multiple related spatial locations over the surface of the Earth. Additionally, in order to take full advantage of the information contained in the hierarchical features obtained from the LR-image, our newly proposed approach incorporates several residual units associated with multiple levels of skip connections that allow the network architecture to neglect low-frequency features, which correspond to spatially irrelevant areas on the surface of the Earth. Our experimental assessment, conducted using the UC Merced and GaoFen-2 remote sensing image collections, three scaling factors, and eight different single-image SR methods, reveals that the proposed approach exhibits competitive advantages when compared to other state-of-the-art SR methods.

The rest of this paper is organized as follows. Section II describes some related works. Section III and describes our newly proposed architecture to super-resolve remotely sensed data. Section IV describes our experimental assessment, where eight different SR methods were tested using two different remote sensing image collections in order to thoroughly discuss and validate the performance of our newly developed approach. Finally, Section V concludes the paper with some remarks and hints at plausible future research lines.

II. Related Works

A. CNNs as Feature Extractors

The CNN model has been widely used in a large range of remote sensing applications (including object detection [41], image classification [42], [43], segmentation [44] and SR [45]) due to its great potential for extracting highly discriminative mid- and high-level abstract features from raw remote sensing data, without involving the hand-crafted selection of these features [46], [47].

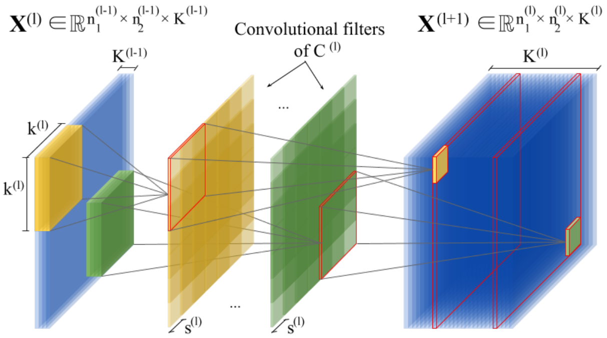

$\mathbf{X}^{(l)} \in \mathbb{R}^{n_1^{(l-1)} \times n_2^{(l-1)} \times K^{(l-1)}}$
Convolutional filters of C $^{(l)}$
$\mathbf{X}^{(l+1)} \in \mathbb{R}^{n_1^{(l)} \times n_2^{(l)} \times K^{(l)}}$
K$^{(l-1)}$
K$^{(l)}$
k$^{(l)}$
k$^{(l)}$
...
...
s$^{(l)}$
s$^{(l)}$

Fig. 1. Graphical visualization of the l-th 2D convolutional layer of our model, denoted by $C^{(l)}$ and composed by $K^{(l)}$ filters. Each filter is defined by the receptive field of the layer, with dimensions $k^{(l)} \times k^{(l)}$, creating a small window that slides over the input volume $\mathbf{X}^{(l)}$ with stride $s^{(l)}$. In each filter, the convolution of the window over the input patches generates a feature, and the collection of features extracted by a filter comprise its feature map. Finally, after applying the activation function and the pooling (omitted for clarity), the resulting collection of feature maps generate the output volume of the layer, $\mathbf{X}^{(l+1)}$.

As any deep neural network (DNN), the CNN's goal is to approximate a function of the form $f : \mathcal{X} \rightarrow \mathcal{Y}$ through the hierarchical concatenation of transformation blocks. In this way, and focusing on the SR problem, the basic performance of a CNN used for SR purposes relies on the sequential and successive transformation of the input LR data, which can be denoted as $\mathcal{X} = \mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, obtaining a highly abstract and discriminative output representation composed by neuron activation values, to which a final mapping is applied in order to obtain the desired HR image $\mathcal{Y} = \mathbf{Y} \in \mathbb{R}^{m_1 \times m_2 \times m_3}$, being $n_1 < m_1$, $n_2 < m_2$ and $n_3 = m_3$.

However, instead of adopting a standard fully-connected (FC) architecture, the CNN model applies the concept of *local receptive field* to connect the neurons in the l-th layer to local smaller windows of each input data $\mathbf{X}^{(l)}$. This idea is inspired by the working mechanism of visual cortex neurons, which are excited by certain stimuli in their receptive area, producing a neural response of higher or lower intensity depending on the stimulus, an effect know as *neuronal tuning* [48]. That is, these neuronal cells are able to look for specific characteristics. Moreover, the neural tuning process becomes more complex as we delve higher into visual areas. In fact, it can be seen as a hierarchical structure, where the visual information is stored in sequences of increasingly complex patterns (feature representations) in sequential order along the visual stream [49]. For instance, neurons in the primary visual cortex (V1) usually respond to simple stimuli, such as edges and shapes, while neurons in higher-level visual respond to more complex stimuli, such as familiar faces.

Two fundamental aspects are drawn from the way the visual cortex operates: i) the hierarchical extraction of higher-level abstraction features in a stacked-layer architecture, and ii) the local connectivity of neurons to small areas of the input data. Regarding the first aspect, the CNN model simulates the hierarchical transformations of the visual stream by applying a deep architecture, composed by a stack of trainable feature extraction stages, while the application of these stages to the data is performed by following a local connectivity design. Usually, each feature extraction stage is composed of three main steps, indicated by Eq. (1): i) the *convolutional layer*, ii) the *non-linear layer* and iii) the *down-sampling* or *pooling layer*. The first one is the basic feature extractor of the CNN model. It is defined by a kernel of weights, whose dimensions determine the receptive field of the layer. In this sense, the convolutional layer acts as a traditional sliding-window algorithm, where the linear kernel convolves ($*$) its weights $\mathbf{W}^{(l)}$ and bias $b^{(l)}$ on local patches of the input data by sliding and overlapping the filter over the input. At every location, the convolutional layer applies an affine transformation between the kernel's weights and the current input data location, obtaining an output volume (set of feature maps), as shown by the first part of Eq. (1):

$$\begin{aligned}\mathbf{O}^{(l+1)} &= \mathbf{W}^{(l)} * \mathbf{X}^{(l)} + b^{(l)}\\ \hat{\mathbf{O}}^{(l+1)} &= \mathcal{H}\left(\mathbf{O}^{(l+1)}\right)\\ \mathbf{X}^{(l+1)} &= \mathcal{P}_{k\times k}\left(\hat{\mathbf{O}}^{(l+1)}\right)\end{aligned} \tag{1}$$

Fig. 1 graphically illustrates the performance of a 2-dimensional convolutional layer, denoted as $C^{(l)}$. This layer receives $\mathbf{X}^{(l)} \in \mathbb{R}^{n_1^{(l-1)}\times n_2^{(l-1)}\times K^{(l-1)}}$ as the input volume. Such volume is characterized by two spatial dimensions, i.e., the volume's height and width $n_1^{(l-1)} \times n_2^{(l-1)}$, and by one spectral dimension, given by the number of filters computed by the previous layer $K^{(l-1)}$. It must be noted that, for $C^{(1)}$ (i.e. the first layer), the number of channels of the input image is given by $K^{(0)} = n_3$. The convolutional layer applies its $K^{(l)}$ filters on the input volume $\mathbf{X}^{(l)}$, with the receptive field defined by $k^{(l)} \times k^{(l)}$. As it can be observed, those kernels are slid over the input, using a stride value $s^{(l)}$ (which usually performs a sub-sampling of the input volume). Each application of those kernels performs a linear element-wise multiplication between the kernel's weights and the current input data location, summing up the obtained results in order to obtain the final feature, which is allocated into the corresponding filter position of the output volume. Eq. (2) gives the mathematical expression of the obtained feature $a_{i,j}^{(l)_z}$ at the (i,j)-th position of the z-th filter in the l-th convolutional layer.

$$\begin{aligned}a_{i,j}^{(l)_z} &= \left(\mathbf{W}^{(l)} * \mathbf{X}^{(l)} + b^{(l)}\right)_{i,j}\\ a_{i,j}^{(l)_z} &= \sum_{\hat{i}=1}^{k^{(l)}}\sum_{\hat{j}=1}^{k^{(l)}} x^{(l)}_{(i\cdot s^{(l)}+\hat{i}),(j\cdot s^{(l)}+\hat{j})} \cdot w^{(l)}_{\hat{i},\hat{j}} + b^{(l)}\end{aligned} \tag{2}$$

Following the second part of Eq. (1), the obtained output volume is passed through the non-linear layer, which applies an element-wise non-linear activation function $\mathcal{H}(\cdot)$ in order to obtain an activity volume that encodes non-linear internal structures and relationships that are hidden in the data. Usually $\mathcal{H}(\cdot)$ is implemented as the rectified linear unit (ReLU) [50].

Finally, at the end of the feature extraction stage, a down-sampling step, performed by a pooling layer $\mathcal{P}_{k\times k}(\cdot)$ with a kernel of dimensions $k \times k$, is added in order to comprise the obtained features in the output volume $\mathbf{X}^{(l+1)} \in \mathbb{R}^{n_1^{(l)}\times n_2^{(l)}\times K^{(l)}}$, and to provide some kind of invariance to small translations of the data.

B. Limitations of CNNs in Remote Sensing Image SR

The application of CNNs to feature extraction from remotely sensed data, in general, and to single-image SR, in particular, has been explored by plenty of works [51], demonstrating very good performance. However, CNN models still face two main limitations in this context. The first one is the fact that there is a direct relationship between the model's depth (i.e., the level of abstraction of data representations), and the quality of the SR method [52], [53]. In this sense, (very) deep CNNs are difficult to train due to the vanishing gradient [54] and data degradation [55] problems, not to mention the intrinsic complexity associated to the task of optimizing a non-convex problem by means of fine-tuning the model's parameters, which can be hampered by the presence of multiple local minima.

Residual learning [55] represents an important evolution in state-of-the-art DNNs, as it introduces an identity mapping between groups of feature extraction stages, denoted as residual units, whose operation on the input data is indicated as $\mathcal{F}(\cdot)$ and is affected by the weights $\mathcal{W}$ and biases $\mathcal{B}$ of those convolutional layers that compose the unit, as Eq. (3) shows:

$$\begin{aligned}\mathbf{O}^{(l+1)} &= \mathcal{F}\left(\mathcal{W}, \mathbf{X}^{(l)}, \mathcal{B}\right) + \mathbf{X}^{(l)}\\ \mathbf{X}^{(l+1)} &= \mathcal{H}\left(\mathbf{O}^{(l+1)}\right)\end{aligned} \tag{3}$$

Direct data propagation through residual and skip connections can alleviate the data degradation problem, thus leading to the development of very deep models for single-image SR [52], [56]. However, despite the fact that connection mechanisms improve information propagation across layers, these CNNs still suffer from a second limitation, related to the intrinsic characteristics of remote sensing images and the internal operation of the convolutional kernel. In particular, remotely sensed image data suffer from certain degradations during their acquisition process, due to atmospheric interferers and sensor noise (among other factors). This often introduces an important amount of noise and variability in the data (in addition to abundant low-frequency information [57]). An optimal feature extractor would be able to discard such irrelevant (or even damaging) information in order to enhance the system's performance. However, the convolutional kernel treats all image content equally, without making any distinctions between relevant and/or useless information, which in the end can hinder the whole SR procedure.

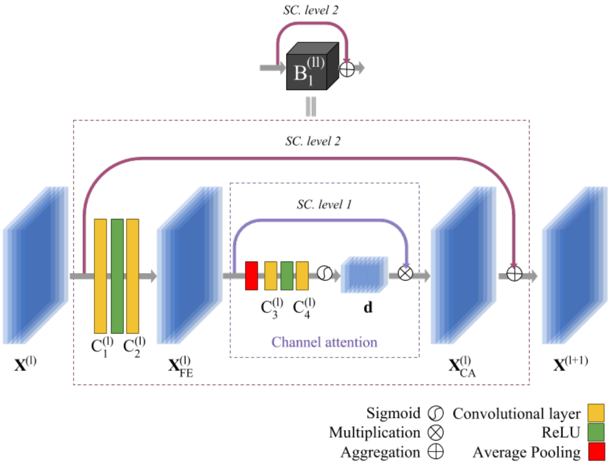


Fig. 2. Graphical visualization of the l-th channel attention block of our model, allocated in the ll-th residual group, $B_l^{(ll)}$. It comprises the first and second SC levels.

In order to overcome the aforementioned shortcomings, certain efforts have been made to equip deep neural network models with visual attention (VA) mechanisms,

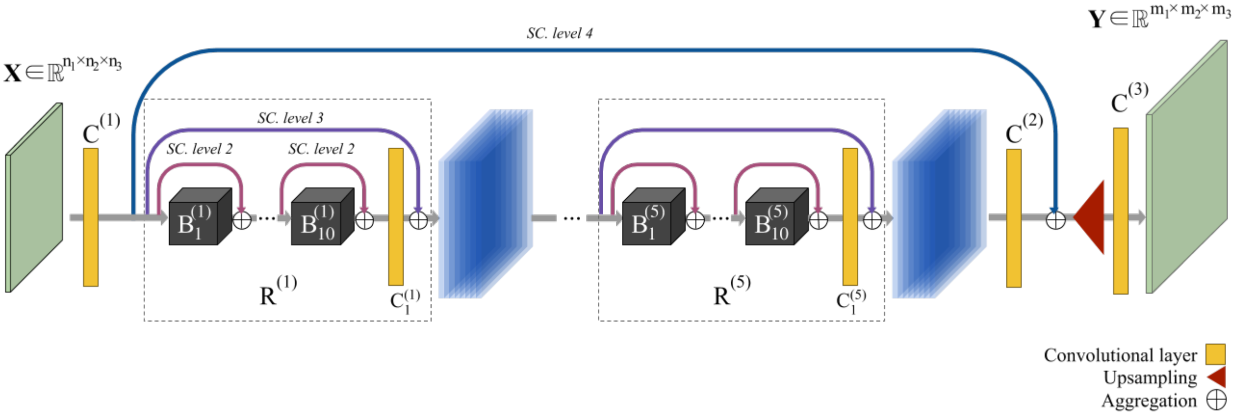


Fig. 3. Architecture of the proposed residual channel attention-based neural network model for remotely sensed image SR. Three different parts of the implemented network can be clearly differentiated in the figure: i) a first convolutional layer $C^{(1)}$ that processes the original LR image $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ to extract an initial output volume composed of $K^{(1)}$ feature maps, which feeds ii) five residual groups ($R^{(ll)}$ with $ll = 1, \cdots, 5$) that are composed by ten channel attention blocks ($B_l^{(ll)}$ with $l = 1, \cdots, 10$) and one final convolution layer, $C_1^{(ll)}$. Shortcut connections of second and third level are applied in order to exploit low- and high-level features that enhance the network's performance. The output volume of $R^{(5)}$ is processed by the $C^{(2)}$ convolutional layer to perform feature extraction before adding the fourth skip connection level, complementing the information extracted by the previous residual groups with the original features. This information is finally iii) up-sampled and processed by $C^{(3)}$ to obtain the HR image $\mathbf{Y} \in \mathbb{R}^{m_1 \times m_2 \times m_3}$, with $n_1 < m_1$, $n_2 < m_2$ and $n_3 = m_3$.

allowing them to focus selectively on the most relevant features [38], [39], [58], [59]. These mechanisms are again inspired by the human visual cortex, where the eye tracks those objects or regions in the scene that stand out from the visual field, following two main components [60]: i) *bottom-up* components, which are stimulus-driven features extracted from raw data in an automatical and involuntary way, i.e., without the understanding of the scene's context information, and ii) *top-down* components, which are task-driven or goal-oriented features extracted through voluntary attention to some scene characteristics, which implies the explicit understanding of the scene's context.

Usually, VA has been included into DNNs by adding a mask or *gating* mechanism, computed from the original data and applied to the features obtained by the network in order to single-out the most relevant ones. In fact, VA mechanisms allow to re-calibrate and refine the feature maps obtained by the CNN model, leading to a more effective training stage. The use of VA mechanisms directed to spatial components of the image has been extensively studied, along with other mechanisms to improve the spatial encoding of data [61]. However, no significant attention has been given as of yet to the spectral component of the data, resulting in the fact that there is currently a lack of methods able to exploit channel relationships [62]. This greatly limits the convolution's flexibility and its representational capacity [63].

III. Deep Residual Channel Attention Model for Remote Sensing Image Super-Resolution

In this section, we introduce a new convolutional-based neural network for remote sensing image SR that employs residual and skip connections to devise a very deep architecture, transferring the information processed at different levels of abstraction and alleviating data degradation problems. At the same time, the internal feature extraction stages in our network have been equipped with VA mechanisms in order to efficiently take advantage of this kind of techniques, which have demonstrated to be very useful in many different high-level tasks related to a wide rage of application domains [64], such as natural image classification tasks [63]. Inspired by the squeeze-and-excitation (SE) building blocks of [63], our proposal integrates the attention technique into a deep-learning-based architecture, adapting it to perform the SR of remote-sensing images. In particular, channel attention blocks (see Fig. 2) have been developed in order to learn and recover high-frequency information, paying attention to channel-wise feature responses and reducing the computations related to low-frequency information. Specifically, our newly developed network relies on improving the obtained deep data representations by modelling the relationships between the channels of the convolved feature maps for each layer, applying a VA gating mechanism over them to extract relevant and high-frequency information. In this context, four levels of skip connections have been included into the proposed architecture in order to enforce different feature levels across different groups of attention blocks, reinforcing also the spatial details not captured in the LR domain by the remote imaging sensor, as it is possible to see in Figs. 2 and 3. Following other deep learning-based SR approaches [25], [35], our newly developed network is composed by three main parts: i) the network's *head*, ii) the *feature processing body*, and iii) the network's *tail*.

1) Network's head: The architecture of the proposed network starts with a first convolutional layer $C^{(1)}$ that transforms the original LR input data $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ into a first-level feature representation, in order to prepare the information that will be fed to the subsequent parts of the network.

2) Network's feature processing body: This can be considered as the main architectural body of the network, as it encodes its main structure. In this part, the data flows through different levels of shortcut connections (SCs) in order to facilitate the forward propagation of the feature maps resulting from each block. These SCs provide a direct and effective way to combine low- and high-level features, improving the network's performance and enhancing its computational efficiency [65]. In fact, the network's body has been developed under a residual paradigm, following a so-called ResNet of ResNets (RoR) architecture [66] that creates four levels of SCs to improve the optimization ability of residual units. In this sense, each SC level groups the network's layers into different structural blocks, creating an architecture of relatively simple blocks embedded into more complex ones, as depicted in Figs. 2 and 3.

From the "internal" or simple structures to the more "external" and complex ones, the network implements the first and second SC levels, circumscribing the basic building block of our SR model, denoted as *channel attention block*. This block performs two steps: i) a feature extraction step, and ii) a channel attention step. In mathematical terms, we denote the input volume of the feature extraction step in the l-th channel attention block as $\mathbf{X}^{(l)}$. Similarly, we denote the input and output volumes of the channel attention step as $\mathbf{X}_{\text{FE}}^{(l)}$ and $\mathbf{X}_{\text{CA}}^{(l)}$, respectively. Finally, we denote the output of the channel attention block as $\mathbf{X}^{(l+1)}$. Fig. 2 provides a graphical overview of these blocks. It can be observed that the feature extraction step is performed by two convolutional layers, $C_1^{(l)}$ and $C_2^{(l)}$, which are connected through a non-linear layer. Each layer extracts more refined features from $\mathbf{X}^{(l)}$, while the non-linear layer applies the ReLU function to obtain the activation values. As a result, the $\mathbf{X}_{\text{FE}}^{(l)} \in \mathbb{R}^{n_1^{(l)} \times n_2^{(l)} \times K^{(l)}}$ contains the $K^{(l)}$ feature maps extracted by $C_2^{(l)}$. These feature maps are sent to the channel attention step to be re-calibrated. A first average pooling layer is then applied over the spatial dimensions in order to *squeeze* the spatial information, reducing the collection of feature maps to a channel descriptor denoted as $\hat{\mathbf{d}} \in \mathbb{R}^{K^{(l)}}$ [63]. This vector collects the global spatial information, where each vector element d_z is obtained by Eq. (4) as follows:

$$\hat{d}_z = \frac{1}{n_1^{(l)} \cdot n_2^{(l)}} \cdot \sum_{i=1}^{n_1^{(l)}} \sum_{j=1}^{n_2^{(l)}} x_{\text{FE}(i,j)}^{(l)}, \text{ with } z = 1, \cdots, K^{(l)} \tag{4}$$

After the squeeze step, an *excitation* process is adopted to fully capture the internal relationships and dependencies between the feature channels. In this way, a gating mechanism is implemented by a spectral encoder-decoder architecture, where the encoder layer, $C_3^{(l)}$, performs a channel down-scaling step followed by a ReLU function, and the decoder layer, $C_4^{(l)}$, recovers the spectral dimension, performing a channel up-scaling step (both layers comprise 1×1 kernels). Finally, the sigmoid function is employed to obtain a *scaled* channel descriptor $\mathbf{d}$ whose values lie in the interval $[0, 1]$. Eq. (5) gives a mathematical expression for the aforementioned excitation and scaling procedures:

$$\mathbf{d} = \mathcal{H}_\sigma \left(C_4^{(l)} \left(\text{ReLU} \left(C_3^{(l)} \left(\hat{\mathbf{d}} \right) \right) \right) \right) \tag{5}$$

The channel-wise statistics contained in $\mathbf{d}$ act as a traditional VA mask, scaling the feature maps that comprise the volume $\mathbf{X}_{\text{FE}}^{(l)}$. As a result, the channel-attention output volume $\mathbf{X}_{\text{CA}}^{(l)}$ is obtained by performing the first SC level, as Eq. (6) indicates:

$$\mathbf{X}_{\text{CA}}^{(l)} = \mathbf{d} \cdot \mathbf{X}_{\text{FE}}^{(l)} \tag{6}$$

The channel attention block ends with the aggregation of the original input volume $\mathbf{X}^{(l)}$ and the channel-attention volume $\mathbf{X}_{\text{CA}}^{(l)}$ through the second SC level, given by Eq. (7). This allows to improve the block's input features, thanks to the enhancement made by Eq. (6).

$$\mathbf{X}^{(l+1)} = \mathbf{X}_{\text{CA}}^{(l)} + \mathbf{X}^{(l)} \tag{7}$$

The third SC level groups several channel attention blocks into a complex structure, denoted as residual group: $R^{(ll)}$. As we can observe in Fig. 3, each $R^{(ll)}$ is composed by ten channel attention blocks $B_l^{(ll)}$ that perform the deep feature extraction stage of the network. A final convolutional layer, $C_1^{(ll)}$, is added before conducting the aggregation between the residual group's input volume and the $C_1^{(ll)}$ output feature maps.

Finally, the network implements a fourth SC level that directly connects the output of the network's head with the input of the network's tail by means of an aggregation function, circumscribing the network's body. This allows us to reuse the low-level features extracted by the first convolutional layer, $C^{(1)}$, without any additional computational cost, and the more abstract features obtained by the five implemented residual blocks which are processed by the body's final convolutional layer, $C^{(2)}$, before the final aggregation.

3) Network's tail: After the body of the network has been executed, an output volume composed by very deep feature maps is obtained. Inside, each channel has been re-calibrated by the aforementioned channel attention mechanism, creating a volume of highly informative data. Based on this, up-sampling of the data cube is now carried out, expanding the volume's spatial dimensions to those of the target HR image, i.e. $\mathbf{Y} \in \mathbb{R}^{m_1 \times m_2 \times m_3}$. In this sense, the up-sampling procedure consists of several pairs of convolution and pixel-shuffle layers, denoted as U_i layers, whose number and associated parameters depend on the considered scaling factor. Finally, at the end of the network, the $C^{(3)}$ convolutional layer extracts the necessary information –already scaled– to generate the desired HR image $\mathbf{Y}$.

The details of the network's parameters are listed in Table I. The architectural design of the proposed SR-network has been inspired by some recent convolutional models developed for image SR and available in the remote sensing literature [34], [35], where convolutional filters of size 3×3 have proven to be large enough

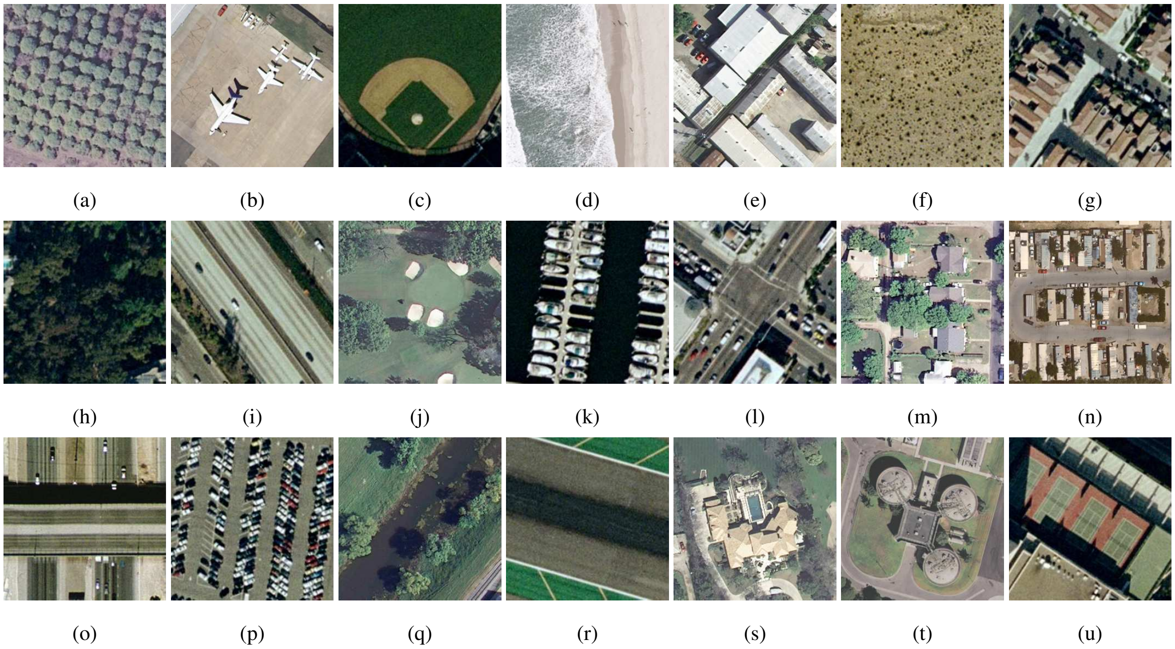


Fig. 4. Examples of the land-use classes present in the UC Merced remote sensing image collection: (a) agricultural, (b) airplane, (c) baseball-diamond, (d) beach, (e) buildings, (f) chaparral, (g) dense-residential, (h) forest, (i) freeway, (j) golf-course, (k) harbor, (l) intersection, (m) medium-residential, (n) mobile-home-park, (o) overpass, (p) parking-lot, (q) river, (r) runway, (s) sparse-residential, (t) storage-tanks and (u) tennis-court.

TABLE I

TOPOLOGY OF THE PROPOSED NETWORK ARCHITECTURE. FOR SIMPLIFICATION, ONLY ONE RESIDUAL GROUP IS SHOWN. $ll = 1, \cdots, 5$ AND $l = 1, \cdots, 10$. * UP-SAMPLING LAYERS EMPLOY A FACTOR OF 2 FOR 2X AND 4X SCALES, AND A FACTOR OF 3 FOR 3X. ONE U LAYER IS EMPLOYED FOR 2X AND 3X, WHILE TWO LAYERS, U_1 AND U_2, IS EMPLOYED FOR 4X.

Network part	**Configuration**: $K^{(l)} \times k^{(l)} \times k^{(l)} \times K^{(l-1)}$			
$Head$	$C^{(1)}$			$64 \times 3 \times 3 \times 3$
$Body$	$R^{(ll)}$	$B_l^{(ll)}$	$C_1^{(l)}$	$64 \times 3 \times 3 \times 64$
			$C_2^{(l)}$	$64 \times 3 \times 3 \times 64$
			$AVG.POOL$	64
			$C_3^{(l)}$	$16 \times 1 \times 1 \times 64$
			$C_4^{(l)}$	$64 \times 3 \times 3 \times 16$
		$C_1^{(ll)}$		$64 \times 3 \times 3 \times 16$
	$C^{(2)}$			$64 \times 3 \times 3 \times 64$
$Tail$	U_i*			$64 \times 3 \times 3 \times 64$
	$C^{(3)}$			$3 \times 3 \times 3 \times 64$

to take advantage of the spatial information contained in neighbourhood windows without losing the level of detail. Moreover, the proposed architecture has been designed to maintain the size of the input volume until reaching the upscaling step, being the layers of each channel attention block the ones that reduce and recover the spatial dimensions of the data volume. For instance, giving an input LR image $\mathbf{X} \in \mathbb{R}^{24 \times 24 \times 3}$, the network's head prepares the input, elongating the spectral information from 3 to 64 channels and keeping constant the spatial dimension by including zero-padding, obtaining a volume of $24 \times 24 \times 64$. In this regard, each channel attention block $B_l^{(ll)}$ compacts the spatial information to a single element through the average pooling layer. As Fig. 2 shows, the volume $\mathbf{X}_{FE}^{(l)}$ keeps the original size of $24 \times 24 \times 64$, and then the average pooling obtains a spectral vector of size $1 \times 1 \times 64$, which is processed by $C_3^{(l)}$ and $C_4^{(l)}$ to obtain the channel descriptor $\mathbf{d} \in \mathbb{R}^{1 \times 1 \times 64}$. The multiplication of the volume $\mathbf{X}_{FE}^{(l)}$ by the channel descriptor $\mathbf{d}$ gives as a result the feature volume $\mathbf{X}_{CA}^{(l)} \in \mathbb{R}^{24\times24\times64}$, recovering the original spatial dimensions. This process is repeated by each block $B_l^{(ll)}$. At the end, the size of the input and output volumes of the network's body is kept constant until reaching the upscaling layer U_{i*}, which scales the spatial dimensions of the feature volume depending on a scale factor. For instance, if the scale factor is $2\times$, the obtained volume will be of size $48 \times 48 \times 64$. Finally, the last convolutional layer $C^{(3)}$ reduces the spectral dimension to 3 channels, giving as a result the output volume $\mathbf{Y}' \in \mathbb{R}^{48\times48\times3}$.

In addition, the proposed network is trained to minimize the error between the desired HR image, $\mathbf{Y}$, and the obtained one, $\mathbf{Y}'$, as follows:

$$E = |\mathbf{Y} - \mathbf{Y}'| \tag{8}$$

The ADAM optimizer [67] has been adopted to minimize Eq. (8), using 100 epochs with a learning rate $lr = 2e^{-4}$ and a learning decay of 10.

In the following section, a set of experiments have been conducted to evaluate the performance of the proposed network in SR problems involving remotely sensed imagery collected from spaceborne and airborne instruments.

IV. Experimental Results

A. Datasets

The experimental part of this work has been conducted using two different remote sensing image collections:

1) UC Merced [68]: This dataset is one of the most popular image collections within the remote sensing community. It contains a total of 2100 images of the surface of the Earch, which are uniformly distributed in 21 different land-use classes [see Fig. 4: (a) agricultural, (b) airplane, (c) baseball-diamond, (d) beach, (e) buildings, (f) chaparral, (g) dense-residential, (h) forest, (i) freeway, (j) golf-course, (k) harbor, (l) intersection, (m) medium-residential, (n) mobile-home-park, (o) overpass, (p) parking-lot, (q) river, (r) runway, (s) sparse-residential, (t) storage-tanks and (u) tennis-court]. In particular, these images were originally downloaded from the United States Geological Survey (USGS) National Map of different US regions, and they consist of aerial RGB orthoimagery with size of 256×256 pixels and spatial resolution of one foot per pixel.
2) GaoFen-2 [34]: This data set consists of two different remotely sensed multi-spectral data products acquired by the GaoFen-2 satellite over a region in China. Specifically, both scenes have nominal spatial resolution of 3.2 meters/pixel and only the RGB channels from the visible spectrum have been considered for the experiments. These data have been provided by the authors of [34] and will be used for qualitative assessment purposes.

B. Experimental Settings

In order to test the performance of the proposed remote sensing SR model, two kinds of experiments have been conducted on the UC Merced and GaoFen-2 data sets, using the following supervised single-image SR methods available in the literature: SC [26], SRCNN [31], FSRCNN [32], CNN-7 [34], LGCNet [34] and DCM [35]. It should be also mentioned that the the bicubic interpolation algorithm (BC) is provided as the baseline result.

On the one hand, the UC Merced collection has been used to train and test the proposed network, taking into account the high variety of classes and samples present in this data set. Specifically, the UC Merced data collection has been randomly split into two balanced halves, to generate equitable training and test partitions with

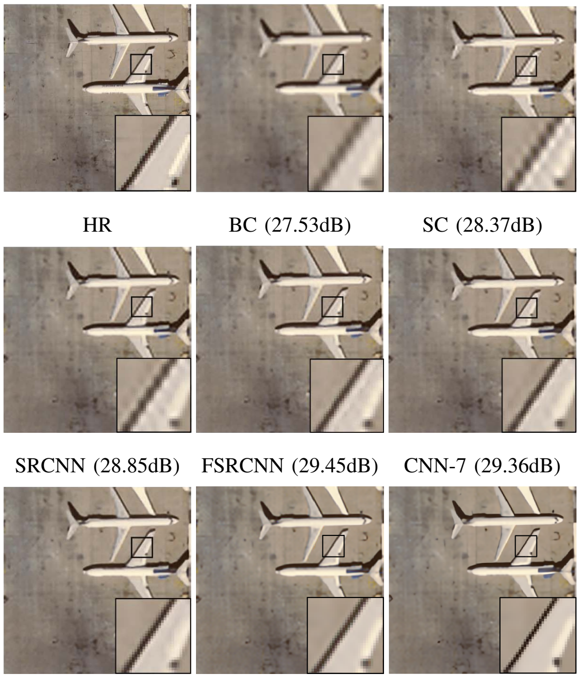
$1,050$ samples each. In addition, 20% of the available training data (i.e., 10 images per class) is used for validation purposes, to set the hyperparameters of the proposed approach and the other tested SR methods. Regarding the experimental protocol of the testing phase, the original UC Merced HR images have been down-sampled according to three different scaling factors: $2\times$, $3\times$ and $4\times$, using the bi-cubic interpolation kernel to generate the corresponding LR counterparts. Moreover, five different Monte Carlo runs have been conducted for each test image, which makes a total of $9,450$ runs per SR method.

On the other hand, the GaoFen-2 dataset has been employed to validate the generalization ability of the proposed approach when considering a completely external test image collection. In particular, the $3\times$ and $4\times$ scaling factors used for UC Merced training have been adopted to super-resolve the two additional GaoFen-2 data products, with the ultimate goal of assessing the performance of SR methods in the task of transferring the knowledge learned from the UC Merced data set to a different remote sensing image collection.

Regarding the assessment protocol, two different full-reference image metrics have been used to quantitatively evaluate the obtained SR results: the peak signal-to-noise ratio (PSNR) [69], and the structural similarity index (SSIM) [70]. Finally, the hardware and software environments used for the experiments are made up of the following components: an Intel Core i7-6700K processor, a GPU NVIDIA GeForce GTX 1080, 40 GB of DDR4 RAM, a 2 TB Toshiba DT01ACA HDD, an ASUS Z170 motherboard, Ubuntu 18.04.1 x64 as operating system and Pytorch 0.4.1 with CUDA 9.

C. Results

Table II presents the PSNR (dB) and SSIM quantitative assessment for the SR experiments carried out over the test set of the UC Merced collection. In particular, the considered scaling factors are presented in rows, whereas the different SR methods and metrics are provided in columns. Additionally, Table III details the average PSNR (dB) metric results per class, when considering a $3\times$ up-scaling factor. We emphasize that that each table contains the average values after five Monte Carlo runs of the corresponding SR methods, and the best metric results are highlighted using bold font.



HR BC (27.53dB) SC (28.37dB)

SRCNN (28.85dB) FSRCNN (29.45dB) CNN-7 (29.36dB)

LGCNet (29.70dB) DCM (30.44dB) **Proposed (30.78dB)**

Fig. 5. Qualitative assessment of the UC Merced airplane test image considering a $3\times$ scaling factor.

For qualitative purposes, Figs. 5 and 6 display the corresponding super-resolved outputs of the considered SR methods when considering two test images of the UC Merced airplane and road classes, and $3\times$ and $4\times$ scaling factors, respectively. In addition, Figs. 7 and 8 show the output results when super-resolving the GaoFen-2 airport and factory test images, respectively using the UC Merced training information for the $3\times$ and $4\times$ scaling factors.

TABLE II

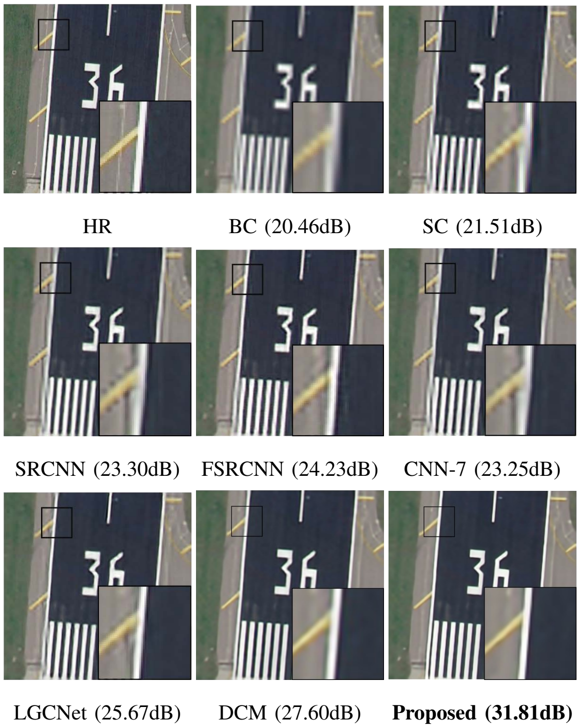
PSNR (dB) and SSIM assessment for the considered SR methods (in columns) using three different scaling factors (in rows).

	Bicubic	SC [26]	SRCNN [31]	FSRCNN [32]	CNN-7 [34]	LGCNet [34]	DCM [35]	Proposed
Scale	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM
2	30.76 / 0.8789	32.77 / 0.9166	32.84 / 0.9152	33.18 / 0.9196	33.15 / 0.9191	33.48 / 0.9235	33.65 / 0.9274	**34.37±0.930 / 0.9296±7.03e-5**
3	27.46 / 0.7631	28.26 / 0.7971	28.66 / 0.8038	29.09 / 0.8167	29.02 / 0.8155	29.28 / 0.8238	29.52 / 0.8394	**30.26±1.07e-2 / 0.8507±1.47e-3**
4	25.65 / 0.6725	26.51 / 0.7152	26.78 / 0.7219	26.93 / 0.7267	26.86 / 0.7264	27.02 / 0.7333	27.22 / 0.7528	**27.88±1.30e-3 / 0.7707±3.81e-4**

TABLE III

Class-based UC Merced quantitative SR assessment considering a 3× scaling factor.

Class	Bicubic	SC [26]	SRCNN [31]	FSRCNN [32]	CNN-7 [34]	LGCNet [34]	DCM [35]	**Proposed**
1	26.86	27.23	27.47	27.61	27.59	27.66	29.06	**30.17**
2	26.71	27.67	28.24	28.98	28.81	29.12	30.77	**31.18**
3	33.33	34.06	34.33	34.64	34.59	**34.72**	33.76	34.36
4	36.14	36.87	37.00	37.21	37.22	**37.37**	36.38	36.15
5	25.09	26.11	26.84	27.50	27.39	27.81	28.51	**29.75**
6	25.21	25.82	26.11	26.21	26.22	26.39	26.81	**28.10**
7	25.76	26.75	27.41	28.02	27.89	28.25	28.79	**29.43**
8	27.53	28.09	28.24	28.35	28.35	28.44	28.16	**28.57**
9	27.36	28.28	28.69	29.27	29.16	29.52	30.45	**32.19**
10	35.21	35.92	36.15	36.43	36.39	**36.51**	34.43	34.13
11	21.25	22.11	22.82	23.29	23.32	23.63	26.55	**27.66**
12	26.48	27.20	27.67	28.06	27.99	28.29	29.28	**29.83**
13	25.68	26.54	27.06	27.58	27.48	27.76	27.21	**27.80**
14	22.25	23.25	23.89	24.34	24.30	24.59	26.05	**27.73**
15	24.59	25.30	25.65	26.53	26.19	26.58	27.77	**29.01**
16	21.75	22.59	23.11	23.34	23.37	23.69	24.95	**25.84**
17	28.12	28.71	28.89	29.07	29.03	29.12	28.89	**28.79**
18	29.30	30.25	30.61	31.01	30.93	31.15	32.53	**32.85**
19	28.34	29.33	29.40	30.23	29.94	**30.53**	29.81	29.22
20	29.97	30.86	31.33	31.92	31.87	**32.17**	29.02	31.08
21	29.75	30.62	30.98	31.34	31.32	**31.58**	30.76	31.51
AVG	27.46	28.27	28.66	29.09	29.02	29.28	29.52	**30.26**



HR BC (20.46dB) SC (21.51dB)

SRCNN (23.30dB) FSRCNN (24.23dB) CNN-7 (23.25dB)

LGCNet (25.67dB) DCM (27.60dB) **Proposed (31.81dB)**

Fig. 6. Qualitative assessment of the UC Merced road test image considering a 4× scaling factor.

D. Discussion

According to the quantitative results reported in Tables II and III, there are some relevant points that need to be emphasized. The first important aspect concerns the impact of the scaling factor on the final performance of the tested SR methods. In this sense, Table II reveals that both PSNR and SSIM metrics decrease when considering higher scaling factors. This is because the amount of available visual information logically diminishes with the LR input image size. However, it is also possible to observe that the proposed approach provides quantitative improvements with respect to the other considered methods. When considering the PSNR metric, the proposed architecture achieves the best average result, which is substantially higher than the one obtained by any other SR method. More specifically, the gains obtained by the proposed method with regards to the other ones in terms of average PSNR are as follows: +0.71dB (DCM), +0.91dB (LGCNet), +1.10dB (FSRCNN), +1.16dB (CNN-7), +1.41dB (SRCNN), +1.66dB (SC), and +2.88dB when compared to the bi-

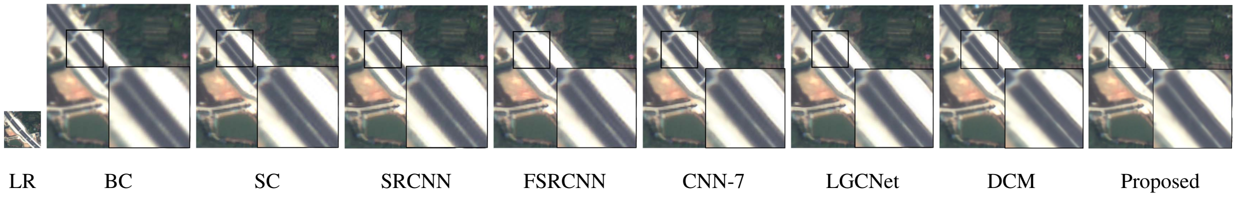


Fig. 7. Qualitative assessment of the GaoFen-2 airport test image considering a 3× scaling factor.

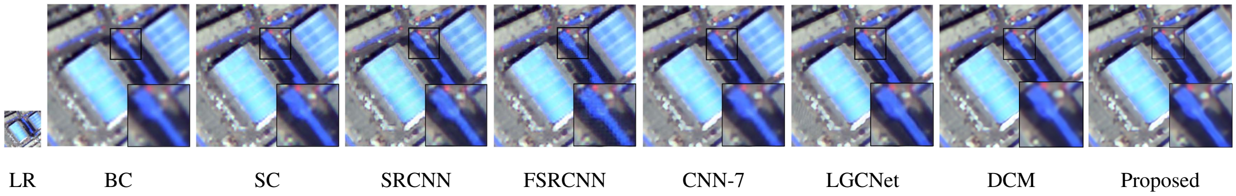


Fig. 8. Qualitative assessment of the GaoFen-2 factory test image considering a 4× scaling factor.

cubic interpolation baseline. In the case of the SSIM metric, the proposed approach also outperforms, on average, DCM (+0.010 units), LGCNet (+0.023), FSRCNN (+0.029), CNN-7 (+0.030), SRCNN (+0.037), SC (+0.041), and the bi-cubic interpolation (+0.079).

This initial quantitative comparison shows that the most recent deep learning SR methods in the literature (DCM, LGCNet and FSRCNN) provide the best results for all the considered scaling factors. Nonetheless, the proposed approach is able to obtain superior average PSNR and SSIM results, especially when considering the highest scaling factors, i.e. 3× and 4×. Note that the remote sensing SR problem becomes particularly challenging as the scaling ratio increases, because less visual information is available for the SR process itself. In this context, the novel attention mechanism integrated within our newly presented architecture allows the proposed approach to exploit better those LR image regions that require more computations to effectively introduce additional HR spatial details, and which cannot be easily recovered from a global deep learning SR perspective (in which all image features are equally relevant).

Another important point is related to the consistency of the proposed approach regarding the obtained quantitative results per class (Table III). As it is possible to observe, the proposed remote sensing SR architecture obtains the best PSNR results in 15 of the 21 UC Merced categories when considering a 3× scaling factor, which certainly indicates that our approach exhibits a great potential to manage a broader range of remotely sensed imagery. Despite the fact that other recent deep learning models, such as DCM and LGCNet, also exhibit a good overall SR performance, the PSNR results per class reported in Table III reveal that our method is particularly effective when dealing with classes that contain spatially detailed structures mixed with relatively invariant land-cover surfaces, which is a typical scenario in remotely sensed images. The PSNR performance improvements obtained in some classes, such as harbor, mobile-home-park or parking-lot, indicate that the implemented attention mechanism is able to effectively focus the network computations towards the image components that require most of the HR details, for example, boats, houses or cars, providing competitive advantages (from a remote sensing standpoint) with regards to other SR methods.

The effectiveness of the proposed approach is also supported by the qualitative results displayed in Figs. 5-8. Specifically, Figs. 5-6 show the super-resolved output of two UC Merced test images, i.e. airplane and road, considering two different scaling factors, i.e. $3\times$ and $4\times$, respectively. As these visual results show, each specific SR model encourages a particular kind of feature on the super-resolved output. Whereas SC and SRCNN methods appear to very sensitive to aliasing and moire effects, because they are unable to distinguish between the relevant high-frequency image components and the up-scaling noise, the most recent deep learning approaches, i.e. FSCNN, CNN-7, LGCNet, DCM and the proposed approach can effectively attenuate these undesirable anomalies. This is because these methods use deeper architectures, which allows them to recover more precise HR image patterns. A clear example of this can be seen in the airplane wing of Fig. 5, where BC, SC and SRCNN introduce a significant aliasing effect, while the other methods are able to generate a super-resolved result with higher quality.

Despite the fact that FSRCNN, CNN-7, LGCNet and DCM are generally able to provide satisfactory SR performance, it is possible to appreciate some important advantages of the proposed architecture in the task of super-resolving remote sensing data by analyzing the visual results in more detail. Specifically, according to Fig. 6, the proposed approach certainly provides the sharpest edges and the most similar output with regards to the corresponding ground-truth HR counterpart. In fact, there is a main factor that differentiates the result obtained by proposed approach result from the other results: the image blur, and also the noise reduction. As it is possible to observe in Fig. 5, the proposed method is the most effective one when reducing the noise present in the airplane wing. Besides, Fig. 6 shows that the proposed approach is able to produce the clearest road lines and the most homogeneous concrete surface. In addition to all these observations, the qualitative SR results presented in Figs. 7 and 8 also suggest the robustness of the proposed approach in the task of transferring the knowledge acquired from the UC Merced collection to the GaoFen-2 dataset, in spite of the existing spatial resolution differences. More specifically, it is possible to see that the proposed method is able to reduce blur and ringing artifacts, eventually leading to a better visual quality in the super-resolved output.

At this point, it is important to emphasize that the most recent deep learning-based SR methods, i.e. [33]–[35], aim to enhance remotely sensed optical data by using deeper architectures, which allows them to uncover more representative convolutional features and introduce additional HR components in the super-resolved result. Nonetheless, these architectures often become very difficult to train because of their large number of hidden layers and parameters, which eventually leads to a poor propagation of activations and gradients in the back-propagation process, i.e. the so-called vanishing gradient problem [37]. Precisely, these undesirable effects, together with the special complexity of remotely sensed imagery, generate a degradation of the convolutional features that may introduce blur and noise artifacts in the final result. The proposed remote sensing SR approach mitigates these problems by implementing a novel attention mechanism over a residual block architecture, which allows the network to focus on those image components that require more computations to be super-resolved. Note that this aspect takes on special importance when dealing with remote sensing data, because Earth surface acquisitions are rather complex aerial captures in which different image regions typically demand different processing levels. A clear example of this fact can be observed in Fig. 6, where the concrete surface does not require substantial changes (unlike the road lines, which need to be completely recovered at a $4\times$ scale). In this sense, the proposed approach intelligently discards the low-frequency image components through the network

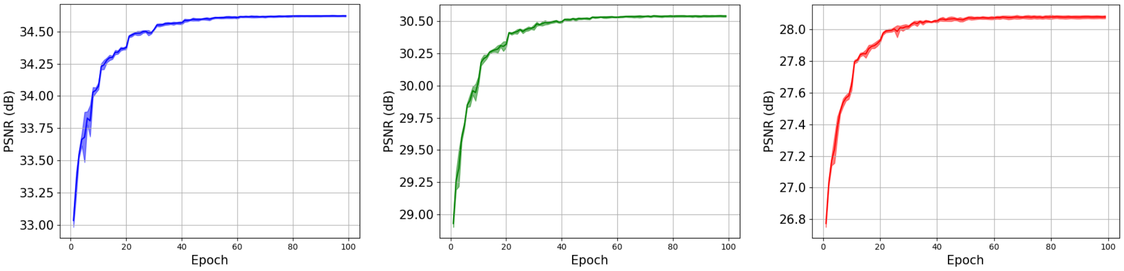


Fig. 9. Average PSNR evolution (per epoch) for the UC Merced validation set when considering 2× (left), 3× (center) and 4× (right) scaling factors.

in order to focus on the most spatially relevant Earth surface areas and, consequently, improve the network convergence. Specifically, Fig. 9 presents the average PSNR evolution per epoch for the UC Merced validation set, in order to illustrate the fast and consistent convergence of our newly proposed architecture. As it is possible to observe, the network is able to converge after 30 epochs for the three considered scaling factors, i.e. 2×, 3× and 4×. Accordingly, the improvements introduced by the proposed approach lead to higher remote sensing SR performance when compared to other state-of-the-art methods.

V. Conclusions and Future Lines

This work presents a new single-image SR approach which has been specially designed for dealing with the particular complexity of remotely sensed imagery. Specifically, the proposed deep learning architecture incorporates a new attention mechanism into a residual-based network design. Such mechanism allows our method to focus the SR process on those Earth surface components that require more computations to be super-resolved. In this way, the less informative low-frequency features (that is, visual characteristics extracted from spatially irrelevant Earth surface areas) are intelligently discarded by means of four different levels of skip connections. Consequently, the performance of the proposed SR approach improves significantly, since the network convergence is driven by the most relevant high-frequency information. Our experiments, conducted using the UC Merced and GaoFen-2 remote sensing image collections, three scaling factors, and eight different single-image SR methods, reveal that the proposed approach offers state-of-the-art performance when super-resolving remotely sensed optical data.

One of the most important conclusions that arises from this work is the importance of adopting an effective attention mechanism within the network design when super-resolving airborne and space-borne optical data with deep learning models. Whereas the current trends in CNN-based SR of remotely sensed data, e.g. [34], [35], do not identify the most important convolutional features from the input acquisition instrument, the qualitative and quantitative SR results obtained by our newly proposed approach reveal that guiding the network training process towards the most informative high-frequency features leads to very competitive performance with respect to other state-of-the-art remote sensing SR models. Despite the fact that the proposed approach exhibits remarkable potential, our future work will be directed towards the following improvements: 1) adapting the proposed architecture to the unsupervised self-learning SR paradigm, 2) extending the model cost function to simultaneously assess multiple image quality metrics, and 3) expanding the network formulation to inter-sensor tandem platforms.

References

[1] C. Toth and G. Jóźków, "Remote sensing platforms and sensors: A survey," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 115, pp. 22–36, 2016.

[2] J. A. Benediktsson, J. Chanussot, and W. M. Moon, "Very high-resolution remote sensing: Challenges and opportunities [point of view]," *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1907–1910, 2012.

[3] N. Pettorelli, H. Schulte to Bühne, A. Tulloch, G. Dubois, C. Macinnis-Ng, A. M. Queirós, D. A. Keith, M. Wegmann, F. Schrodt, M. Stellmes *et al.*, "Satellite remote sensing of ecosystem functions: opportunities, challenges and way forward," *Remote Sensing in Ecology and Conservation*, vol. 4, no. 2, pp. 71–93, 2018.

[4] H. Song, B. Huang, Q. Liu, and K. Zhang, "Improving the spatial resolution of landsat TM/ETM+ through fusion with SPOT5 images via learning-based super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 3, pp. 1195–1204, 2015.

[5] F. Li, L. Xin, Y. Guo, D. Gao, X. Kong, and X. Jia, "Super-resolution for gaofen-4 remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 1, pp. 28–32, 2018.

[6] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geoscience and remote sensing magazine*, vol. 1, no. 2, pp. 6–36, 2013.

[7] R. Fernandez-Beltran, A. Plaza, J. Plaza, and F. Pla, "Hyperspectral Unmixing Based on Dual-Depth Sparse Probabilistic Latent Semantic Analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 11, pp. 6344–6360, 2018.

[8] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep Pyramidal Residual Networks for Spectral-Spatial Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, no. 99, pp. 1–15, 2018.

[9] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, J. Li, and F. Pla, "Capsule Networks for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, no. 99, pp. 1–15, 2018.

[10] Y. Xian, Z. I. Petrou, Y. Tian, and W. N. Meier, "Super-Resolved Fine-Scale Sea Ice Motion Tracking," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5427–5439, 2017.

[11] Y. Chen, Y. Ge, G. B. Heuvelink, R. An, and Y. Chen, "Object-Based Superresolution Land-Cover Mapping From Remotely Sensed Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 1, pp. 328–340, 2018.

[12] Y. Chen, Y. Ge, and Y. Jia, "Integrating object boundary in super-resolution land-cover mapping," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 1, pp. 219–230, 2017.

[13] Y. Zhang, Y. Zhang, W. Li, Y. Huang, and J. Yang, "Super-resolution surface mapping for scanning radar: Inverse filtering based on the fast iterative adaptive approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 1, pp. 127–144, 2018.

[14] R. Fernandez-Beltran, J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla, "Multimodal probabilistic latent semantic analysis for sentinel-1 and sentinel-2 image fusion," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 9, pp. 1347–1351, 2018.

[15] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE signal processing magazine*, vol. 20, no. 3, pp. 21–36, 2003.

[16] P. Milanfar, *Super-resolution imaging*. CRC press, 2010.

[17] L. Yue, H. Shen, J. Li, Q. Yuan, H. Zhang, and L. Zhang, "Image super-resolution: The techniques, applications, and future," *Signal Processing*, vol. 128, pp. 389–408, 2016.

[18] H. Zhang, Z. Yang, L. Zhang, and H. Shen, "Super-resolution reconstruction for multi-angle remote sensing images considering resolution differences," *Remote Sensing*, vol. 6, no. 1, pp. 637–657, 2014.

[19] R. Fernandez-Beltran, P. Latorre-Carmona, and F. Pla, "Single-frame super-resolution in remote sensing: a practical overview," *International Journal of Remote Sensing*, vol. 38, no. 1, pp. 314–354, 2017.

[20] A. S. Belward and J. O. Skøien, "Who launched what, when and why; trends in global land-cover observation capacity from civilian earth observation satellites," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 103, pp. 115–128, 2015.

[21] R. Sandau, H.-P. Roeser, A. Valenzuela *et al.*, *Small satellite missions for earth observation*. Springer, 2014.

[22] C.-Y. Yang, C. Ma, and M.-H. Yang, "Single-image super-resolution: A benchmark," in *European Conference on Computer Vision*. Springer, 2014, pp. 372–386.

[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] Y. Zhang, F. Ling, X. Li, and Y. Du, "Super-resolution land cover mapping using multiscale self-similarity redundancy," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 11, pp. 5130–5145, 2015.

[25] J. M. Haut, R. Fernandez-Beltran, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla, "A new deep generative network for unsupervised remote sensing single-image super-resolution," *IEEE*

Transactions on Geoscience and Remote Sensing, vol. 56, no. 11, pp. 6792–6810, 2018.

[26] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[27] S. Gou, S. Liu, S. Yang, and L. Jiao, "Remote sensing image super-resolution reconstruction based on nonlocal pairwise dictionaries and double regularization," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 12, pp. 4784–4792, 2014.

[28] Y. Zhang, Y. Du, F. Ling, S. Fang, and X. Li, "Example-based super-resolution land cover mapping using support vector regression," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 4, pp. 1271–1283, 2014.

[29] R. Fernandez-Beltran, P. Latorre-Carmona, and F. Pla, "Latent topic-based super-resolution for remote sensing," *Remote Sensing Letters*, vol. 8, no. 6, pp. 498–507, 2017.

[30] R. Fernandez-Beltran and F. Pla, "Sparse multi-modal probabilistic latent semantic analysis for single-image super-resolution," *Signal Processing*, vol. 152, pp. 227–237, 2018.

[31] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[32] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *European Conference on Computer Vision*. Springer, 2016, pp. 391–407.

[33] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.

[34] S. Lei, Z. Shi, and Z. Zou, "Super-resolution for remote sensing images via local–global combined network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 8, pp. 1243–1247, 2017.

[35] J. Haut, M. Paoletti, R. Fernandez-Beltran, J. Plaza, A. Plaza, and J. Li, "Remote sensing single-image superresolution based on a deep compendium model," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2019.

[36] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[37] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Advances in neural information processing systems*, 2015, pp. 2377–2385.

[38] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[39] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 286–301.

[40] Y. Lu, Y. Zhou, Z. Jiang, X. Guo, and Z. Yang, "Channel attention and multi-level features fusion for single image super-resolution," *arXiv preprint arXiv:1810.06935*, 2018.

[41] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 12, pp. 7405–7415, 2016.

[42] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks," *arXiv preprint arXiv:1508.00092*, 2015.

[43] K. Nogueira, O. A. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognition*, vol. 61, pp. 539–556, 2017.

[44] M. Längkvist, A. Kiselev, M. Alirezaie, and A. Loutfi, "Classification and segmentation of satellite orthoimagery using convolutional neural networks," *Remote Sensing*, vol. 8, no. 4, p. 329, 2016.

[45] C. Tuna, G. Unal, and E. Sertel, "Single-frame super resolution of remote-sensing images by convolutional neural networks," *International Journal of Remote Sensing*, vol. 39, no. 8, pp. 2463–2479, 2018.

[46] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 22–40, June 2016.

[47] X. X. Zhu, D. Tuia, L. Mou, G. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, Dec 2017.

[48] L. S. Glezer, X. Jiang, and M. Riesenhuber, "Evidence for highly selective neuronal tuning to whole words in the visual word form area," *Neuron*, vol. 62, no. 2, pp. 199–204, 2009.

[49] I. Kuzovkin, R. Vicente, M. Petton, J.-P. Lachaux, M. Baciu, P. Kahane, S. Rheims, J. R. Vidal, and J. Aru, "Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex," *Communications biology*, vol. 1, 2018.

[50] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[51] W. Yang, X. Zhang, Y. Tian, W. Wang, and J.-H. Xue, "Deep learning for single image super-resolution: A brief review," *arXiv preprint arXiv:1808.03344*, 2018.

[52] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) workshops*, vol. 1, no. 2, 2017, p. 4.

[53] 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)*, July 2017, pp. 2790–2798.

[54] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[56] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[57] D. Yang, Z. Li, Y. Xia, and Z. Chen, "Remote sensing image super-resolution: Challenges and approaches," in *2015 IEEE International Conference on Digital Signal Processing (DSP)*, July 2015, pp. 196–200.

[58] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2368–2378, May 2018.

[59] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 1155–1167, Feb 2019.

[60] A. Borji, H. R. Tavakoli, and Z. Bylinskii, "Models of bottom-up attention," *arXiv preprint arXiv:1810.05680*, 2018.

[61] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[62] Y. Hu, J. Li, Y. Huang, and X. Gao, "Channel-wise and spatial feature modulation network for single image super-resolution," *arXiv preprint arXiv:1809.11130*, 2018.

[63] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[64] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 185–207, 2012.

[65] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 4809–4817.

[66] K. Zhang, M. Sun, T. X. Han, X. Yuan, L. Guo, and T. Liu, "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.

[67] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[68] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*. ACM, 2010, pp. 270–279.

[69] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics letters*, vol. 44, no. 13, pp. 800–801, 2008.

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