remote sensing

MDPI

Article

Sentinel-2 Image Fusion Using a Deep Residual Network

Frosti Palsson, Johannes R. Sveinsson * and Magnus O. Ulfarsson

Department of Electrical Engineering, University of Iceland, Hjardarhagi 2-6, Reykjavik 107, Iceland; frostip@gmail.com (F.P.); mou@hi.is (M.O.U.)
* Correspondence: sveinsso@hi.is

Received: 4 July 2018; Accepted: 7 August 2018; Published: 15 August 2018

Abstract: Single sensor fusion is the fusion of two or more spectrally disjoint reflectance bands that have different spatial resolution and have been acquired by the same sensor. An example is Sentinel-2, a constellation of two satellites, which can acquire multispectral bands of 10 m, 20 m and 60 m resolution for visible, near infrared (NIR) and shortwave infrared (SWIR). In this paper, we present a method to fuse the fine and coarse spatial resolution bands to obtain finer spatial resolution versions of the coarse bands. It is based on a deep convolutional neural network which has a residual design that models the fusion problem. The residual architecture helps the network to converge faster and allows for deeper networks by relieving the network of having to learn the coarse spatial resolution part of the inputs, enabling it to focus on constructing the missing fine spatial details. Using several real Sentinel-2 datasets, we study the effects of the most important hyperparameters on the quantitative quality of the fused image, compare the method to several state-of-the-art methods and demonstrate that it outperforms the comparison methods in experiments.

Keywords: residual neural network; image fusion; convolutional neural network; Sentinel-2

1. Introduction

Image fusion can be defined as the fusion of two or more images of different properties or modalities such that the fused image has the same properties as the source images, e.g., spatial and spectral resolution, and is thus more informative. The fusion process must, as best as possible, preserve the salient information found in each source image, and it must avoid introducing spectral and/or spatial distortion into the fused image.

One of the earliest and most established types of image fusion in remote sensing is so-called pansharpening [1,2]. There, a multispectral (MS) image of high spectral resolution but low spatial resolution is fused with a single band panchromatic (PAN) image of high spatial resolution to yield a high spatial resolution MS image, which has the same spatial resolution as the PAN image and the same spectral resolution of the original MS image. By performing this kind of image fusion, more use is made of the available data, and this can be useful for many applications such as classification [3], target detection, snow cover analysis [4], etc. In recent years, more fusion scenarios are becoming possible, such as the fusion of hyperspectral (HS) images and PAN images, referred to as hypersharpening [5–10] to yield HS images of high spatial resolution and the fusion of MS and HS images [11–22] to yield high spatial resolution HS images. Both MS/HS fusion and hypersharpening can be seen as extensions of the pansharpening problem, where the source images have more bands.

What all these fusion problems have in common is that there is significant spectral overlap between the low and high spatial resolution source images. For example, in pansharpening, the PAN image is a wide-band image of a single channel which means that the PAN sensor is sensitive to a wide band of the electromagnetic spectrum. The spectral response of the MS sensor has a significant overlap with the spectral response of the PAN sensor. Recently, more advanced MS sensors have been developed, which acquire images from more spectral bands covering a wider band of the electromagnetic spectrum. The Sentinel-2 constellation of satellites operated by the European Space Agency (ESA), and the Worldview-3 and Worldview-4 satellites operated by DigitalGlobe (Westminster, CA, USA), are examples of such sensors. They typically acquire images in the visible, near-infrared (NIR) and shortwave infrared (SWIR) regions of the electromagnetic spectrum, and at different spatial resolution. For example, the Sentinel-2 sensor acquires four bands at 10 m resolution (ground sampling distance), six bands at 20 m resolution and three bands at 60 m resolution.

Since all the acquired images show the same scene, this presents an opportunity for single sensor image fusion or super-resolution, i.e., to enhance the resolution of the coarse resolution bands using information from the finer bands. However, due to the lack of spectral overlap between bands, this problem is more challenging than the pansharpening problem. A widely used constraint in pansharpening is that a linear combination of the fused bands gives an approximation to the PAN image. This constraint makes the problem easier to solve. However, in the single sensor case, this assumption is false since the bands are spectrally disjoint.

Several methods have been developed to fuse spectrally disjoint images. In [23], the 90 m resolution thermal bands of the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) were sharpened using data from the 15 m resolution Visible Near Infrared (VNIR) bands using a method based on Generalized Laplacian Pyramid (GLP) [24]. The 20 m bands of Sentinel-2 were sharpened in [25] using geostatical stochastic simulation and genetic programming. The 500 m resolution bands of the Moderate Resolution Imaging Spectroradiometer (MODIS) were upscaled to 250 m resolution in [26] using area-to-point regression kriging (ATPRK). Wavelet multiresolution analysis was used to enhance the 500 m bands using the 250 m bands in [4]. Sentinel-2 super-resolution was performed in [27] by solving a convex deconvolution problem in a lower dimensional subspace and using a roughness penalty as a regularizer, and its extension was given in [28], by using cyclic decent on a manifold. A two-stage method for Sentinel-2 fusion was given in [29] that separates band-dependent geometrical information from band specific reflectance and then applies this model to the lower resolution bands while preserving their reflectance using spectral unmixing techniques.

Recently, deep learning based methods have been demonstrated to outperform traditional signal processing approaches in areas such as speech and pattern recognition [30]. These are methods based on deep neural networks, and specifically in pattern recognition and related fields [31,32], the so-called convolutional neural network (CNN) has been shown to be effective. Deep learning methods have been used to solve the pansharpening problem [33–35], multispectral/hyperspectral image fusion [11], and super-resolution [36]. In [37], the authors trained a deep residual neural network to super-resolve Sentinel-2 images using extensive training data with global coverage. In this study, we focus more on the single image case and provide a study of how the performance of the network is affected by several important hyperparameters of the method. It is not meant to find an optimal set of the hyperparameters, but rather to study the effects of each hyperparameter on the quantitative quality metric values. However, hyperparameter tuning for deep learning algorithms is an important issue, and several methods have been proposed to automate this task. In [38,39], particle swarm optimization was used, Ref. [40] used greedy sequential algorithms using the expected improvement criterion, Ref. [41] used the Covariance Matrix Adaptation Evolution Strategy (CMA-ES), Ref. [42] used Bayesian optimization based on Gaussian Processes, and, finally, Ref. [43] extrapolated learning curves to speed up the task of parameter selection.

Residual neural networks (ResNets) [44] have recently been shown to give good performance in image recognition tasks. The residual design allows deeper networks to be trained more easily and there is indeed evidence that deeper networks perform better than shallower nets [30].

In this paper, we propose a method for single sensor image fusion based on a deep residual neural network architecture. The network consists of a number of residual blocks, where each residual block contains two convolutional (conv) layers. The last layer in each block is an element-wise sum layer where the output of the layer preceding the block is added to its output. Apart from the residual blocks, there is also an important residual aspect to the network inspired by the fusion model. The upscaled coarse part of the input is added to the last layer of the network, relieving it from having to learn the low-pass structure of the image. The lack of high-resolution reference during training is circumvented by reducing the resolution of the observed data before training, by the resolution ratio between the coarse and fine bands. Therefore, the observed coarse bands can be used as the reference or target during training. This strategy is inspired by Wald's protocol [45], and is commonly used in image fusion in remote sensing to evaluate the performance of fusion methods. The assumption being made here is that the relationship learned between the reduced resolution level and the observed level also applies to the higher level [46].

We perform several experiments using real Sentinel-2 datasets and compare the proposed method to three state-of-the-art methods for single sensor fusion. These are the model-based Super-Resolution for Multispectral Multiresolution Estimation (SupReME) method from [27], the Area-To-Point Regression Kriging (ATPRK) method from [47] and the Superres method from [29].

The outline of this paper is as follows. In Section 2, we give a brief overview of residual networks and describe the proposed method in detail. In Section 3, we discuss implementation issues, choice of hyperparameters, experimental results, and finally, in Section 4, the conclusions are drawn.

2. The ResNet

Recently, deep and very deep CNNs have been demonstrated to perform significantly better than shallower networks in many image recognition tasks [30,48,49]. In general, deep networks are difficult to train due to the problem of vanishing/exploding gradients [50,51], however this has currently been largely mitigated by techniques such as batch-normalization [52], self normalizing networks [53] and better initialization techniques [51,54]. Still, a problem remains with training deep networks. With an increasing number of layers, the accuracy of the network saturates and then starts decreasing. This phenomenon is known as degradation [44] and is not caused by overfitting of the network. Once the network accuracy becomes saturated, adding more layers will only result in higher training error.

A solution to this problem is the deep residual learning framework or ResNets [44]. The main idea behind ResNets is that, instead of letting a stack of layers learn the desired mapping $\mathcal{H}(\mathbf{x})$, shortcut connections are constructed which skip over the stack and are added to the output of the previous layers. The shortcut connection is the identity mapping, and it forces the "skipped" layers to fit a residual mapping, $\mathcal{F}(\mathbf{x}) = \mathcal{H}(\mathbf{x}) - \mathbf{x}$, which is easier to optimize. In this way, the originally desired mapping has been transformed into $\mathcal{F}(\mathbf{x}) + \mathbf{x}$. This is illustrated in Figure 1.

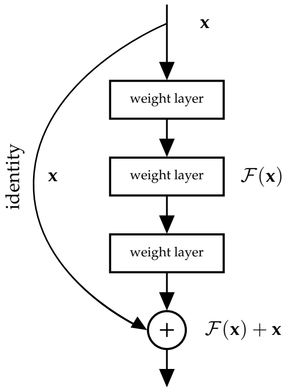


Figure 1. A residual building block. Instead of letting the layers learn the desired mapping $\mathcal{H}(\mathbf{x})$, a skip connection (identity mapping) is constructed that forces the skipped layers to fit a residual mapping $\mathcal{F}(\mathbf{x}) = \mathcal{H}(\mathbf{x}) - \mathbf{x}$, that is easier to optimize since the skipped layers are relieved of learning $\mathbf{x}$.

In this paper, we use the following notation: the observed fine bands are denoted by $\mathbf{Y} \in \mathbb{R}^{d_1 \times d_2 \times L_1}$ and observed coarse bands are denoted by $\mathbf{X} \in \mathbb{R}^{d_1/r \times d_2/r \times L_2}$, where $d_1 \times d_2$ is the dimension of the fine resolution bands, L_1 is the number of fine resolution bands, L_2 is the number of coarse resolution bands and r is the resolution ratio between the fine and coarse resolution bands. Filtering and subsequent downsampling by the factor r, i.e., decimation, is denoted by the operator $\mathbf{D}$ and upsampling by a factor r and subsequent filtering is denoted by $\mathbf{U}$. Finally, square brackets denote the concatenation of matrices along the spectral dimension, i.e., $[\mathbf{X},\ \mathbf{Y}]$. Note that the first two dimensions of the matrices need to be the same. The operators $\mathbf{U}$ and $\mathbf{D}$ operate band-wise.

Since there is no high-resolution reference image available, the data need to be degraded in resolution before training to be able to use the observed coarse bands as the reference. This approach, depicted in Figure 2, is the most widely used method for quantitative evaluation of image fusion methods such as pansharpening. We denote the spatially degraded fine and coarse bands by $\mathbf{DY}$ and $\mathbf{DX}$, respectively, where the operator $\mathbf{D}$ degrades their spatial dimensions by the factor r, which is the resolution ratio between the fine and coarse bands. Since $\mathbf{DX}$ is smaller than $\mathbf{DY}$ by a factor of r along each spatial dimension, it needs to be interpolated to the same size as $\mathbf{DY}$. We denote the interpolated degraded coarse bands by $\mathbf{X}^D \in \mathbb{R}^{d_1/r \times d_2/r \times L_2} = \mathbf{UDX}$ and the degraded fine bands by $\mathbf{Y}^D \in \mathbb{R}^{d_1/r \times d_2/r \times L_1} = \mathbf{DY}$. Now, $\mathbf{X}^D$, $\mathbf{Y}^D$ and $\mathbf{X}$, i.e., the observed coarse bands, have the same spatial size.

To make the training of the network computationally feasible, the input images are divided into many small overlapping patches of suitable size. Thus, the input to the network are patches of the stacked degraded bands, i.e., patches of $[\mathbf{X}^D,\ \mathbf{Y}^D] \in \mathbb{R}^{d_1/r \times d_2/r \times (L_1 + L_2)}$, denoted by $[\mathbf{X}_i^D,\ \mathbf{Y}_i^D] \in \mathbb{R}^{p \times p \times (L_1 + L_2)}$, $i = 1, \ldots, M$, where M is the number of patches and p is the patch-size. The target patches during training come from $\mathbf{X}$ and are denoted by $\mathbf{X}_i \in \mathbb{R}^{p \times p \times (L_1 + L_2)}$, $i = 1, \ldots, M$. As all the bands have the same size, the ith patch covers the same part of the scene for all the images, i.e., $\mathbf{X}^D$, $\mathbf{Y}^D$ and $\mathbf{X}$.

We can formulate the fusion problem as

$$\hat{\mathbf{X}}^H = \mathbf{UX} + R(\mathbf{Y}, \mathbf{UX}), \tag{1}$$

where the estimated fused image is denoted by $\hat{\mathbf{X}}^H$ and $R(\mathbf{Y}, \mathbf{UX})$ is the mapping from the fine bands $\mathbf{Y}$ and the upscaled coarse bands $\mathbf{UX}$, i.e., the input bands. Thus, the residuals $R(\mathbf{Y}, \mathbf{UX})$ are fine details that are added to $\mathbf{UX}$, which can be viewed as the low-pass component of the fused image $\hat{\mathbf{X}}^H$. In this framework, the residual mapping $R(\mathbf{Y}, \mathbf{UX})$ is learned by the network during training. The design of the network is shown in Figure 3. The network is residual on two-levels. First, we have the residual blocks that are designed as shown in Figure 1. Each residual block consists of a conv layer with leaky rectified linear unit (ReLU) activation. activation [55], followed by another conv layer with linear activation and finally an element-wise sum layer where the output of the layer preceding the residual block is added to the output of its last conv layer. There are a total of K residual blocks in the network, where the parameter K is a tuning parameter of the method. Aside from the residual blocks themselves, there are two other skip connections, one that skips the entire stack of residual blocks and one that skips from the input layer to the output layer of the network. Only the coarse part of the input, i.e., the part $\mathbf{UX}$ is added via the skip connection. This design reflects the model given in Equation (1). The last skip connection effectively enforces the residual nature of the fusion process according to the model.

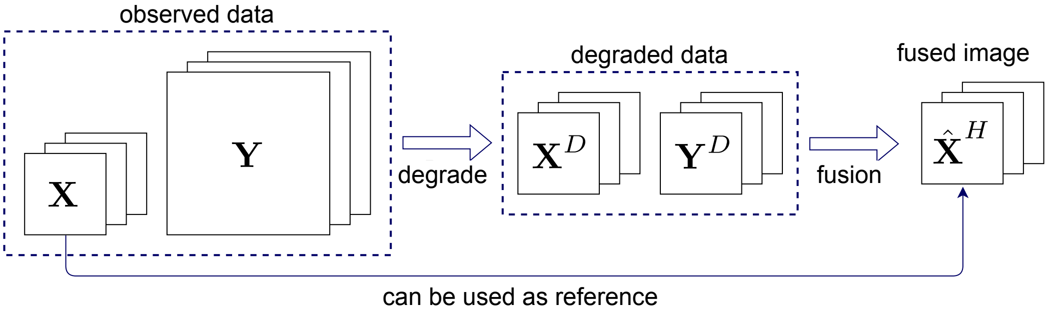


Figure 2. To obtain a reference image, the observed data are reduced in resolution by their respective resolution ratio. Then, the observed lower resolution image can be used as the reference image.

2.1. Training

The key idea of the method is to degrade the resolution of the input images **X** and **Y** before training to use the observed coarse bands **X** as the training targets. The input and target images are split into M overlapping patches of size $p \times p$ pixels with a shift of one pixel between patches such that the patches completely cover the source images. The training and target patches are then obtained by randomly sampling M patches from the input and target images without replacement. The input to the network are then M stacked patches of $\mathbf{X}^D$ and $\mathbf{Y}^D$, i.e., $[\mathbf{X}_i^D,\ \mathbf{Y}_i^D] \in \mathbb{R}^{p\times p\times(L_1+L_2)}$, $i = 1,\ldots,M$. The target patches are $\mathbf{X}_i \in \mathbb{R}^{p\times p\times(L_1+L_2)}$, $i = 1,\ldots,M$. The loss function of the network is given by

$$J(\mathbf{\Theta}) = \frac{1}{M}\sum_{i=1}^{M} \|\mathbf{F}([\mathbf{X}_i^D,\ \mathbf{Y}_i^D];\mathbf{\Theta}) - \mathbf{X}_i\|_2^2, \tag{2}$$

where $F([\mathbf{X}_i^D,\ \mathbf{Y}_i^D];\mathbf{\Theta})$ is the prediction of the network for the ith patch, $\mathbf{\Theta}$ are the network parameters and M is the number of patches.

An important factor in the method is how the decimation and interpolation of the bands is performed, i.e., the operators **D** and **U**. For downsampling, we use the GLP filters described in [56], which are implemented in the Scikit-image Python library [57] as the function *pyramid_reduce*. The parameter σ of the Gaussian filter is chosen as $\frac{r}{3}$, which is the default value of this parameter. For interpolation of the bands, we use spline interpolation of order 5. This is implemented in the Scikit-image function *rescale*. The CNN was implemented using the high level Keras [58] deep learning library, which is now a part of Tensorflow 1.4 [59].

2.2. Testing

The main hypothesis behind the method is that the relationship learned by the network between the data at the reduced scale, i.e., between $[\mathbf{X}^D,\ \mathbf{YD}]$ and **X** also holds for $[\mathbf{UX},\ \mathbf{Y}]$ and the hypothetical fine resolution image $\mathbf{X}^H$, which are the coarse bands **X** at the next higher resolution scale, i.e., the scale of **Y**. When the network has been trained, the entire stack of the input images at the observed resolution scale is processed at once to yield the high-resolution bands $\hat{\mathbf{X}}^H$.

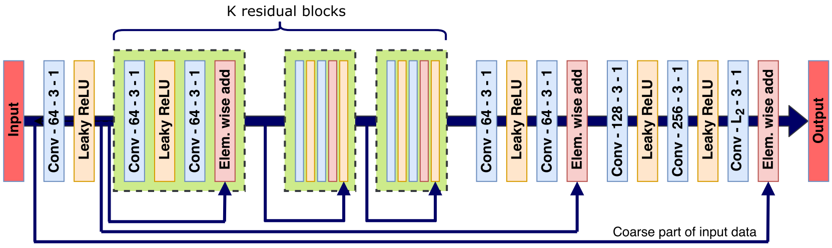


Figure 3. The structure of the residual network. Conv-64-3-1 denotes a conv layer with 64 filters of size 3 and stride 1.

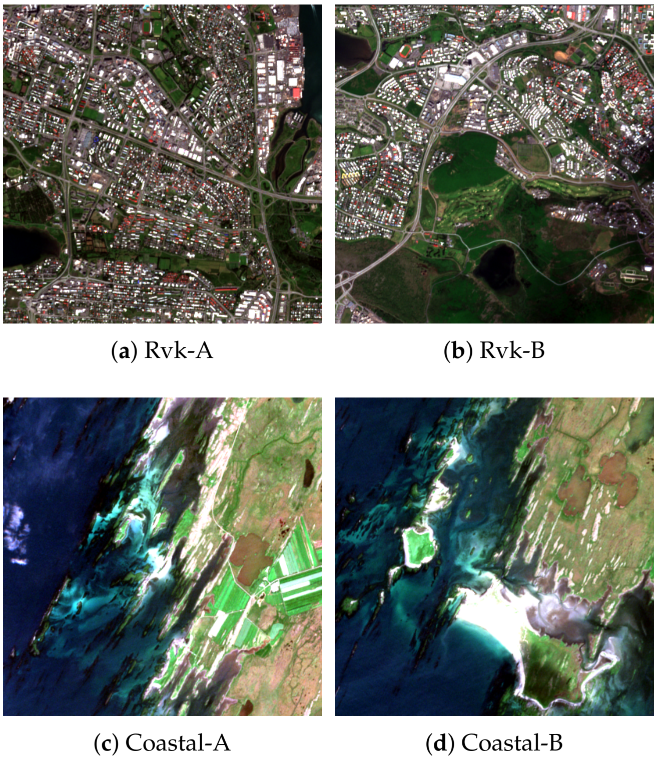
3. Experiment Results

3.1. Data

The Sentinel-2 constellation is a part of ESA's Copernicus Programme and consists of the twin Sentinel-2A and Sentinel-2B polar-orbiting satellites that are located in the same orbit, phased at 180° to each other. Built by Airbus DS (Ottobrunn, Germany) and operated by the ESA, these satellites were designed to deliver global coverage of high-resolution multi-spectral imagery with high revisit frequency and provide observation data for the next generation of operational products such as land-cover change-detection maps, managing of natural disasters and forest monitoring. Each satellite is provided with a multi-spectral instrument (MSI) that has a total of 13 spectral channels in the visible/NIR and SWIR range. The MSI splits the incoming reflected light at a filter and focuses it onto two distinct focal plane assemblies/detectors. There is a separate detector for visible and NIR bands (VNIR) and another one for SWIR bands. Stripe filters mounted on top of the detectors separate each spectral band into individual wavelengths. The data are acquired on 13 spectral bands in the VNIR and SWIR range. There are four bands in the visible and NIR range at a 10 m resolution (ground sampling distance), six bands in the NIR and SWIR range at 20 m resolution and three bands at a 60 m resolution in the visible NIR and SWIR ranges. The radiometric resolution of all bands is 12 bits per pixel, i.e., there are 4096 possible light intensity values. Finally, the temporal resolution, i.e., the revisit frequency of the combined constellation is five days and ten days for a single satellite.

The MSI products are compilations of granules of a fixed size. For ortho-rectified products such as Level-1C products, the granules are 100 km^2 ortho-images in UTM/WGS84 projection. All data in this paper come from a single Level-1C product which means that the data have been converted from radiances into top-of-atmosphere (TOA) reflectances and projected into cartographic coordinates using a digital elevation model (DEM) and then resampled at constant ground sampling distance of 10 m, 20 m or 60 m depending on the native resolution of the respective spectral bands.

The dataset used is a 100 km by 100 km tile showing part of western Iceland, including the capital Reykjavik and the fjords of Hvalfjörður and Borgarfjörður. The acquisition date is 27 July 2017, and it is almost cloud free. From this image, we have made four smaller ones for 20 m super-resolution of size 408 by 408 pixels at 10 m resolution that translates to 4.08 km by 4.08 km and one large data set of size 3452 by 3452 pixels, i.e., 34.52 km by 34.52 km, which is used for experiments involving the super-resolution of 60 m bands. The datasets come in pairs of similar images denoted by A and B. One pair is of a rural coastal area north of Borgarfjörður (64.501°N, 22.010°W), and the other pair shows part of the capital of Iceland, Reykjavik (64.80°N, 21.56°W). The datasets are referred to as Coastal-A, Coastal-B, and Rvk-A, Rvk-B, and they are shown as RGB color images in Figure 4.



(**a**) Rvk-A

(**b**) Rvk-B

(**c**) Coastal-A

(**d**) Coastal-B

Figure 4. Datasets used for the super-resolution of 20 m bands. The datasets are displayed as RGB color images using bands B2, B3 and B4 as the red, green and blue channels, respectively.

3.2. Experiment Methodology

Two end-user scenarios can be envisioned for the proposed method. The first one is a single image approach, where the network is trained and tested on the same image. This means that, if the user is interested in super-resolving a specific image, he/she would need to train the network specifically for that image. The second scenario is to pre-train the network on a large number of different scenes such that it can generalize well-enough to make training for specific images unnecessary. This is the approach in [37], where a deep residual network was trained on a large number of different images. In this paper, the focus will be on the single image approach. However, we do experiments where we train on one image and test on another image. The experiments are divided into experiments involving 20 m bands, and experiments involving the 60 m bands. Experiments, where we try to estimate the effect of different network parameters such as the number of residual blocks or the patch-size, are based on 40 m to 20 m super-resolution. To test the generalization capacity of the method, we train on one image, referred to as image A and test on another similar image, referred to as image B.

3.3. Quality Metrics and Reduced Resolution Evaluation

To be able to evaluate the fused image quantitatively, a reference image is needed. Obviously, such a reference image is not available. However, in image fusion such as pansharpening, it is common practice to reduce the observed data in resolution by a factor that is equal to the resolution factor between the higher resolution bands and the lower resolution bands. In this way, the observed lower resolution bands can be used as the reference image for the fusion. This is the method that is used for the training of the network. The drawback is however that, obviously, the reduced resolution image fusion problem is not the same problem as the fusion at the observed resolution scale. By degrading the observed data, information is lost. However, this method enables the comparison of image fusion methods since it can give an idea of their relative performance as measured by quantitative quality evaluation metrics. The experiments are divided into 40 m to 20 m super-resolution, and 360 m to 60 m super-resolution and the main focus will be on super-resolution of the 20 m bands.

By using the reduced resolution method, we can use standard performance metrics or indices to evaluate the quality of the estimated high-resolution images. For this purpose, we use the following quality indices: signal-to-reconstruction error (SRE), Erreur Relative Globale Adimensionnelle de Synthese (ERGAS) and Spectral Angle Mapper (SAM).

The SRE is the ratio of the power of the signal to the error, and it is given in decibels by

$$\mathrm{SRE} = 10 \log_{10} \frac{\mu_x^2}{\|\mathbf{X} - \hat{\mathbf{X}}\|^2/N}, \tag{3}$$

where $\mathbf{X}$ denotes the reference image, $\hat{\mathbf{X}}$ denotes the estimated image, μ_x^2 is the mean of $\mathbf{X}$ and N is the number of pixels in each band.

ERGAS [60] calculates the amount of spectral/spatial distortion in the enhanced image based on the mean square error (MSE) and is given by

$$\mathrm{ERGAS} = \frac{100}{r^2} \sqrt{\frac{1}{L_2} \sum_{l=1}^{L_2} \left(\frac{\sqrt{\sum_{n=1}^{N} (\hat{\mathbf{X}}_{n,l} - \mathbf{X}_{n,l})^2 / N}}{\sum_{n=1}^{N} \hat{\mathbf{X}}_{n,l} / N} \right)^2}, \tag{4}$$

where r^2 is the ratio of high-resolution pixels to low-resolution pixels, L_2 denotes the number of bands and $\mathbf{X}_{n,l}$ denotes pixel n of band l.

Finally, SAM calculates the spectral similarity between two vectors as an angle. The value of SAM for the entire image is the average of all the angles for each pixel. It is given in degrees by

$$\mathrm{SAM}(\hat{\mathbf{X}}, \mathbf{X}) = \frac{1}{N} \sum_{n=1}^{N} \arccos \left(\frac{\sum\limits_{l=1}^{L_2} \hat{\mathbf{X}}_{n,l} \mathbf{X}_{n,l}}{\sqrt{\sum\limits_{l=1}^{L_2} \hat{\mathbf{X}}_{n,l}^2 \sum\limits_{l=1}^{L_2} \mathbf{X}_{n,l}^2}} \right). \tag{5}$$

For ERGAS and SAM, the optimal value is 0, while higher values are better for the SRE.

The data are degraded in resolution by first filtering each band with a Gaussian filter, with the standard deviation parameter σ chosen as $\frac{1}{r}$, then filtering again with a moving average filter with kernel size 3 × 3 and finally downsampling by factor r.

3.4. *20 m Bands—Study of the Effect of Network Parameters*

In this part of the experiments, we investigate the effect of various network hyperparameters, such as the number of training patches, the number of training epochs, the number of residual blocks K, and the patch size. We train on the image A and test on both images A and B.

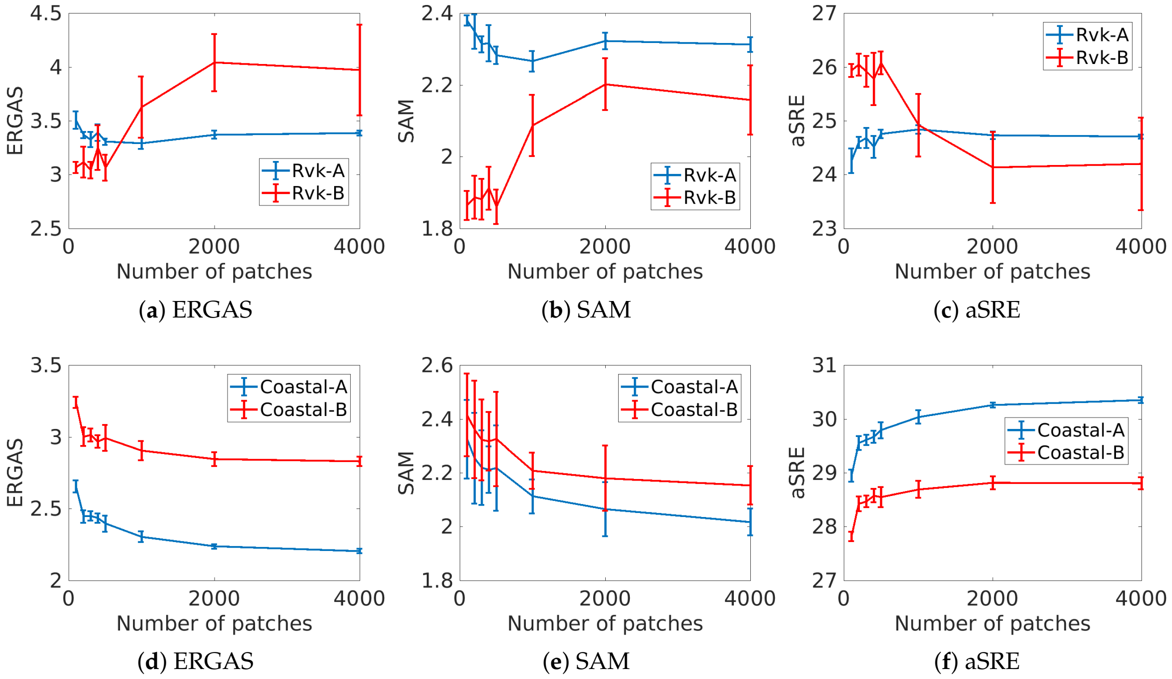
The results of all the experiments are the values of the quantitative quality metrics for both train and test datasets for each value of the parameter under study. The default value of the parameters are as follows: the number of training patches is 500, the number of residual blocks is 24, the size of patches is 8 × 8 pixels, and the number of epochs is 200. For a single experiment where one parameter is varied, the random seed for the random patches selected for training is kept fixed to see the effects of that parameter better.

3.4.1. Effect of the Number of Training Patches

In this experiment, we estimate the effect of the number of training patches on the values of the quantitative quality metrics. The number of patches varies from 100 to 4000 in increments of 100 and 1000 patches. For each number of patches, 30 trials are performed, and the results of the experiment are the mean of the test metrics for both A and B images. The results are shown in Figure 5. For the urban Rvk datasets, all evaluation metrics reach an optimal value at 1000 or fewer training patches for the train image A and test image B. Using more than 1000 patches results in worse performance and especially so for the test image. This is a clear sign of overfitting of the network when the number of training patches increases. For the coastal data set, the behavior is very different. Using more training patches gives better results according to all the metrics. It seems that overfitting is much less of an issue for the coastal images than for the urban images, but it is not clear why that is so. These results indicate that the optimal number of training patches is largely scene dependent.

3.4.2. Effect of the Number of Training Epochs

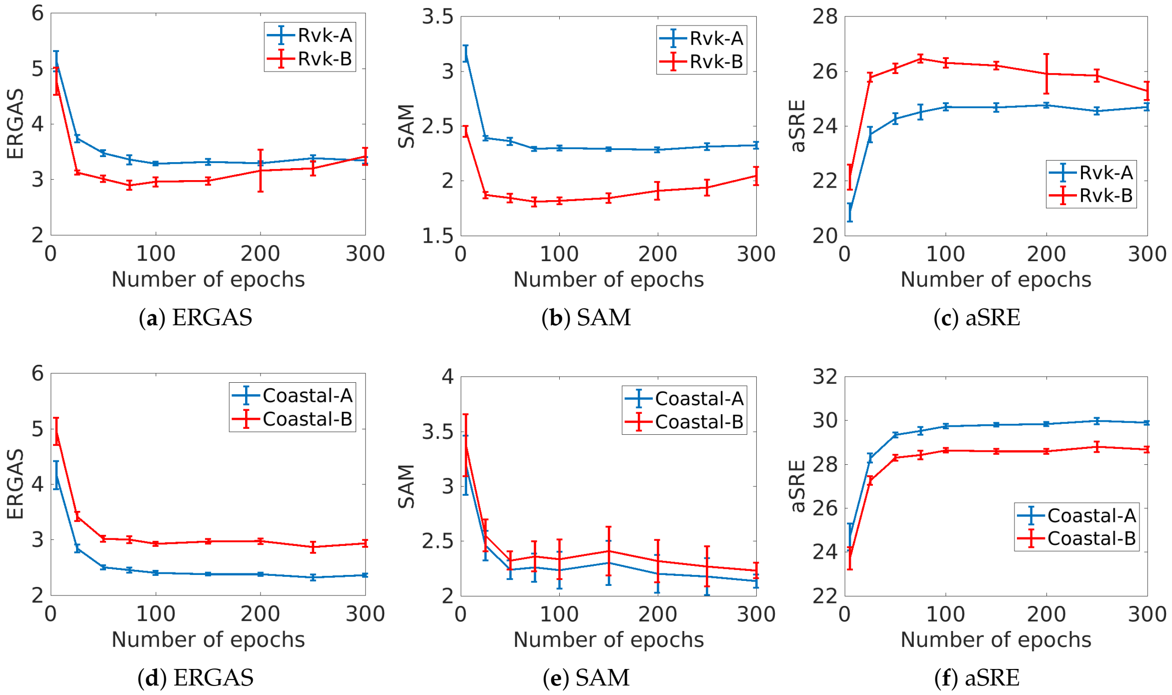
Now, we consider how the evaluation metrics change as a function of the number of epochs used to train the network. The setup of the experiment is the same as before, and now only the number of training epochs varies from 15 to 100 in increments of 25 epochs and from 100 to 300 in increments of 50 epochs. The results are shown in Figure 6. For the urban Reykjavik dataset, the performance metrics for both A and B images indicate optimal performance is reached at around 100 epochs of training. Training for a higher number of epochs does not improve the performance of the network. For the coastal dataset, the best results are obtained at 250 epochs for ERGAS and aSRE and at 300 epochs for SAM. This experiment reveals that the network converges quickly and that, when the training data is limited, training the network for a large number of epochs does not improve the performance, even if the cost function is still decreasing. Note that the network is trained at a lower resolution scale than used when predicting the fused image and the relation between the network loss and the values of the metrics is not strictly linear.



(**a**) ERGAS (**b**) SAM (**c**) aSRE

(**d**) ERGAS (**e**) SAM (**f**) aSRE

Figure 5. Values of the quality evaluation metrics as a function of the number of training patches for the Coastal and urban Rvk datasets. The network was trained on the A image and tested on A and B. aSRE is the average SRE of the bands and is given in dB and SAM is given in degrees.



(**a**) ERGAS (**b**) SAM (**c**) aSRE

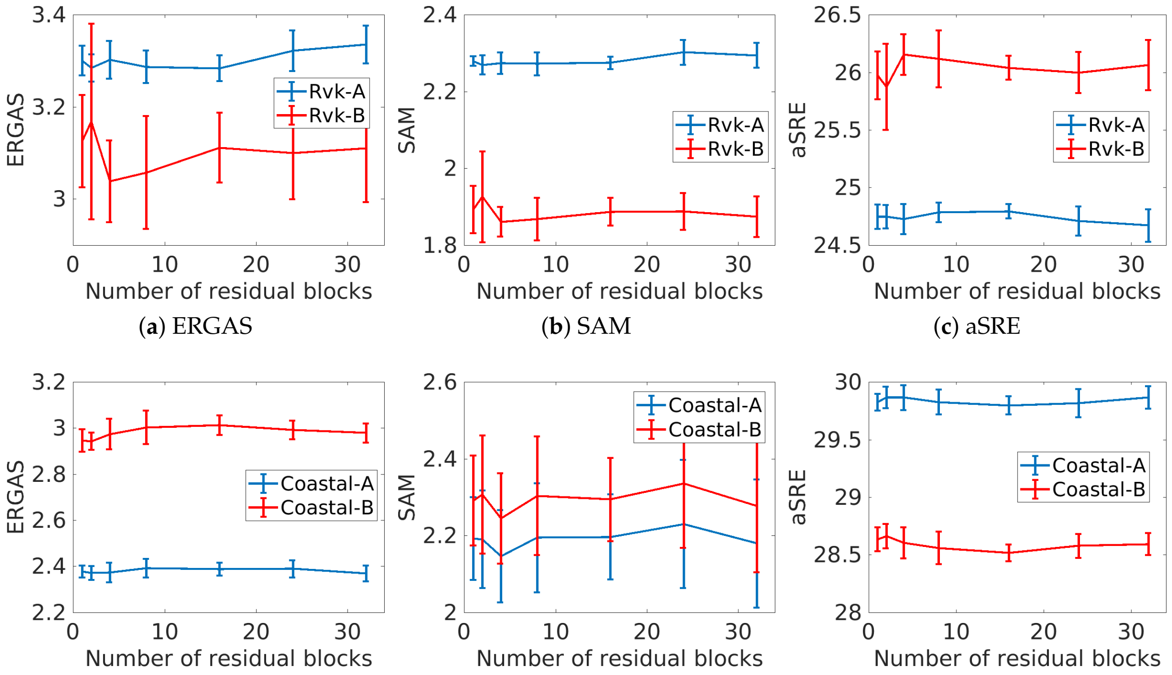
(**d**) ERGAS (**e**) SAM (**f**) aSRE

Figure 6. Values of the quality evaluation metrics as a function of the number of training epochs for the Coastal and urban Rvk datasets. The network was trained on the A image and tested on A and B. aSRE is the average SRE of the bands and is given in dB and SAM is given in degrees.

3.4.3. Effect of the Number of Residual Blocks

The number of residual blocks is an important tuning parameter; in this experiment, we estimate the effect of the network depth in terms of the number of residual blocks on the values of the quantitative quality metrics. The network was trained on the A images for the following values of *K*, i.e., number of residual blocks : 2, 4, 8, 16, 24, 32, and 40. Each residual block contains four layers, i.e., two conv layers, one Leaky ReLU activation layer and one element-wise sum layer. Aside from the residual blocks, the network has six conv layers, and thus the network depth in terms of the number of conv layers varies from eight conv layers to 70 conv layers. For each value of *K*, 30 trials were performed using patch size of 16 by 16 pixels, and the number of patches was set to 500 for the Rvk dataset and 2000 for the Coastal dataset. After training on the A image in each trial, the network was tested on both the A and B images. The test results are shown in Figure 7. The results indicate that, for the urban Reykjavik image, best results for the evaluation metrics are obtained with a low number of residual blocks. For the single image case, i.e., image A, the optimal number of *K* is eight, and it is the same for the test image B. For the coastal images, four to eight residual blocks are optimal. Perhaps the most surprising result of this experiment is that the performance of the network is not very sensitive to the network depth.

Figure 8 shows a plot of the evaluation metrics as a function of number of epochs for the Rvk-A dataset. There are three cases, i.e., the number of residual blocks *K* is 1, 16, and 32. This plot shows that there seems to be no relation between convergence speed and number of residual blocks. Actually, 32 residual blocks give worse performance than one.



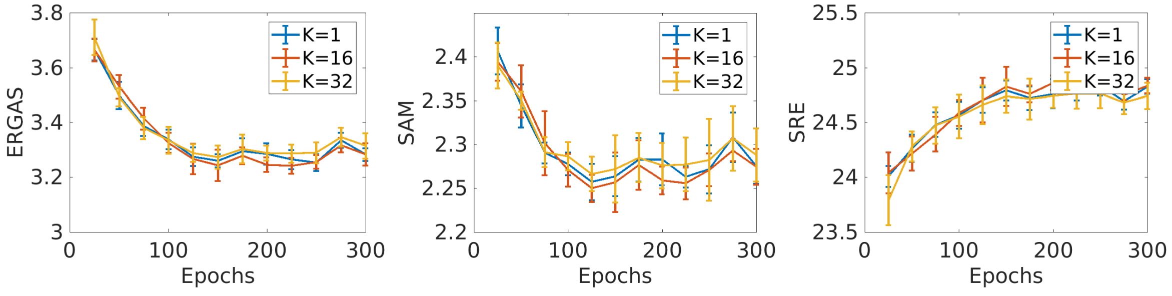
(**a**) ERGAS (**b**) SAM (**c**) aSRE

(**d**) ERGAS (**e**) SAM (**f**) aSRE

Figure 7. Values of the quality evaluation metrics as a function of the number of residual blocks for the Coastal and urban Rvk datasets. The network was trained on the A image and tested on A and B. aSRE is the average SRE of the bands and is given in dB and SAM is given in degrees.

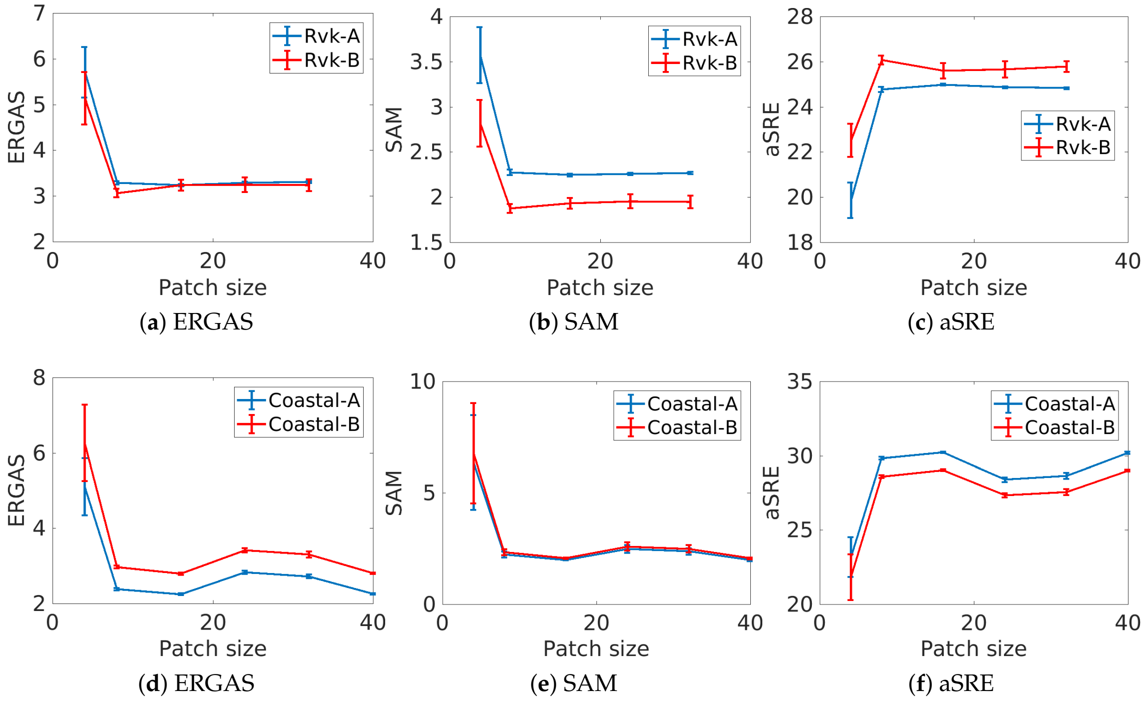
3.4.4. Effect of the Patch Size

In this experiment, the effect of the size of the training patches is investigated. As before, the other hyperparameters are kept fixed and 30 trials are run for each patch size of 4 × 4, 8 × 8, 16 × 16, 24 × 24, 32 × 32, and 40 × 40 pixels, respectively. The network is trained on the image A and tested on both images A and B. The results are summarized in Figure 9. For the Reykjavik image A, a patch size of 16 gives the optimal results, while for the B image, eight pixels gives the best results. For the coastal images, 16 × 16 pixels and 40 × 40 pixels give the best results. A large patch size increases the computational cost of the algorithm significantly, and therefore it is beneficial to keep it as small as possible. According to Figure 9, the optimal patch size for performance and computational complexity is 16 by 16 pixels. Minimal improvement is gained by increasing the patch size beyond that.



(**a**) ERGAS (**b**) SAM (**c**) aSRE

Figure 8. Values of the quality evaluation metrics as a function of the number of Epochs for the Rvk-A dataset and for $K = 1$, $K = 16$, and $K = 32$.



(**a**) ERGAS (**b**) SAM (**c**) aSRE

(**d**) ERGAS (**e**) SAM (**f**) aSRE

Figure 9. Values of the quality evaluation metrics as a function of the patch size for the Coastal and urban Rvk datasets. The network was trained on the A image and tested on A and B. aSRE is the average SRE of the bands and is given in dB and SAM is given in degrees.

3.5. 20 m Bands—Comparison to State-of-the Art

We compare the proposed method to the CNN based method in [11], the SupReME method [27], the ATPRK method [47], and the Superres method [29]. The CNN based method, which we refer to as ConvNet, is based on a three convolutional layer network, which has no skip connections. The size of the filters is 3 by 3 pixels, and there are 32, 64, and 128 filters in each layer, respectively. This method serves as a baseline neural network method for the comparison. The SupReME method depends on solving a deconvolution problem in a lower dimensional subspace. The ATPRK method is based on area-to-point regression kriging of coarse residuals between fine and coarse bands, which are obtained by regression modeling. Finally, the Superres method tries to propagate band-independent details from the fine bands to the coarse bands using spectral unmixing.

For the experiments, we use the same four datasets as in previous experiments, i.e., the Coastal and Rvk pairs of datasets. The hyperparameters of the proposed methods were chosen according to the results of the previous experiments, i.e., the patch was selected as 16 by 16 pixels, the number of training epochs is 200, and the number of residual blocks was set to 24. This also applies to the ConvNet method. Results of the quantitative evaluation of all methods and datasets are summarized in Table 1. The best results are highlighted using a bold typeface.

Table 1. Quantitative evaluation results for all datasets and all methods. The data have been degraded two-fold in resolution in order to use the observed 20 m bands as the reference image. Columns B5 to B12 are the band-wise SRE values, while aSRE is the average SRE of the bands and is given in dB. SAM is given in degrees. Bold typeface indicates the best results.

Coastal-A									
Method	**B5**	**B6**	**B7**	**B8a**	**B11**	**B12**	**aSRE**	**ERGAS**	**SAM**
ResNet	**31.91**	**30.53**	**30.20**	**31.63**	**30.88**	**28.16**	**30.55**	**2.16**	**1.94**
ConvNet	30.94	30.22	29.90	31.41	30.18	27.43	30.01	2.30	2.11
ATPRK	25.33	21.76	21.19	21.19	21.62	20.52	21.94	5.95	5.60
SupReME	25.91	28.17	28.51	29.74	25.11	23.64	26.85	3.43	2.88
Superres	29.41	26.56	26.15	26.76	27.38	26.21	27.08	3.18	2.03
Coastal-B									
Method	**B5**	**B6**	**B7**	**B8a**	**B11**	**B12**	**aSRE**	**ERGAS**	**SAM**
ResNet	**32.1**	**29.32**	**28.62**	**29.92**	**30.23**	**27.50**	**29.63**	**2.53**	**1.93**
ConvNet	30.82	28.82	28.08	29.39	29.37	26.72	28.87	2.77	2.15
ATPRK	25.59	20.82	19.99	20.14	19.31	18.80	20.78	7.51	5.55
SupReME	26.52	27.24	27.15	27.96	21.38	20.91	25.19	4.90	3.19
Superres	29.39	24.98	24.41	24.82	26.09	24.90	25.77	3.98	2.07
Rvk-A									
Method	**B5**	**B6**	**B7**	**B8a**	**B11**	**B12**	**aSRE**	**ERGAS**	**SAM**
ResNet	23.14	**26.32**	26.42	**27.19**	24.55	21.01	24.77	**3.28**	**2.25**
ConvNet	**23.30**	26.27	**26.50**	27.15	**24.59**	**21.15**	**24.83**	3.30	2.30
ATPRK	16.43	18.77	18.73	18.69	17.78	15.68	17.68	7.36	3.64
SupReME	22.33	24.87	25.42	26.08	21.74	18.78	23.20	4.07	2.54
Superres	20.19	23.03	23.08	23.38	21.31	19.66	21.77	4.49	2.37
Rvk-B									
Method	**B5**	**B6**	**B7**	**B8a**	**B11**	**B12**	**aSRE**	**ERGAS**	**SAM**
ResNet	**24.10**	**29.06**	**29.29**	**30.34**	**25.06**	**21.8**	**26.51**	**2.94**	**1.75**
ConvNet	23.05	28.28	28.70	29.69	24.74	20.30	25.79	3.26	2.05
ATPRK	18.80	21.45	21.36	21.48	18.07	16.11	19.55	6.13	2.92
SupReME	22.94	27.48	28.05	28.83	22.76	19.32	24.90	3.60	1.99
Superres	21.32	26.09	26.22	26.44	23.19	20.77	24.00	3.64	1.80

It is evident that the neural network based methods considerably outperform the other methods, with ResNet showing the best performance in every dataset but Rvk-A, where the ConvNet method gives slightly better results for some of the bands. Of the remaining methods, the ATPRK method performs clearly worst in every dataset. The Superres method gives the third best results and SupReME comes fourth. Interestingly, the performance gap between the proposed method and the SupReME method is considerably larger for the coastal datasets than the urban datasets.

Figure 10 shows a visual comparison of the results for the Coastal-A dataset for all bands and methods. A quick glance at the results in Figure 10 reveals that the ATPRK method gives results that are clearly worse than the other methods. It is more difficult to discern the differences between the other methods, but scrutiny of the images shows that the proposed method gives images that are indeed the closest to the reference.

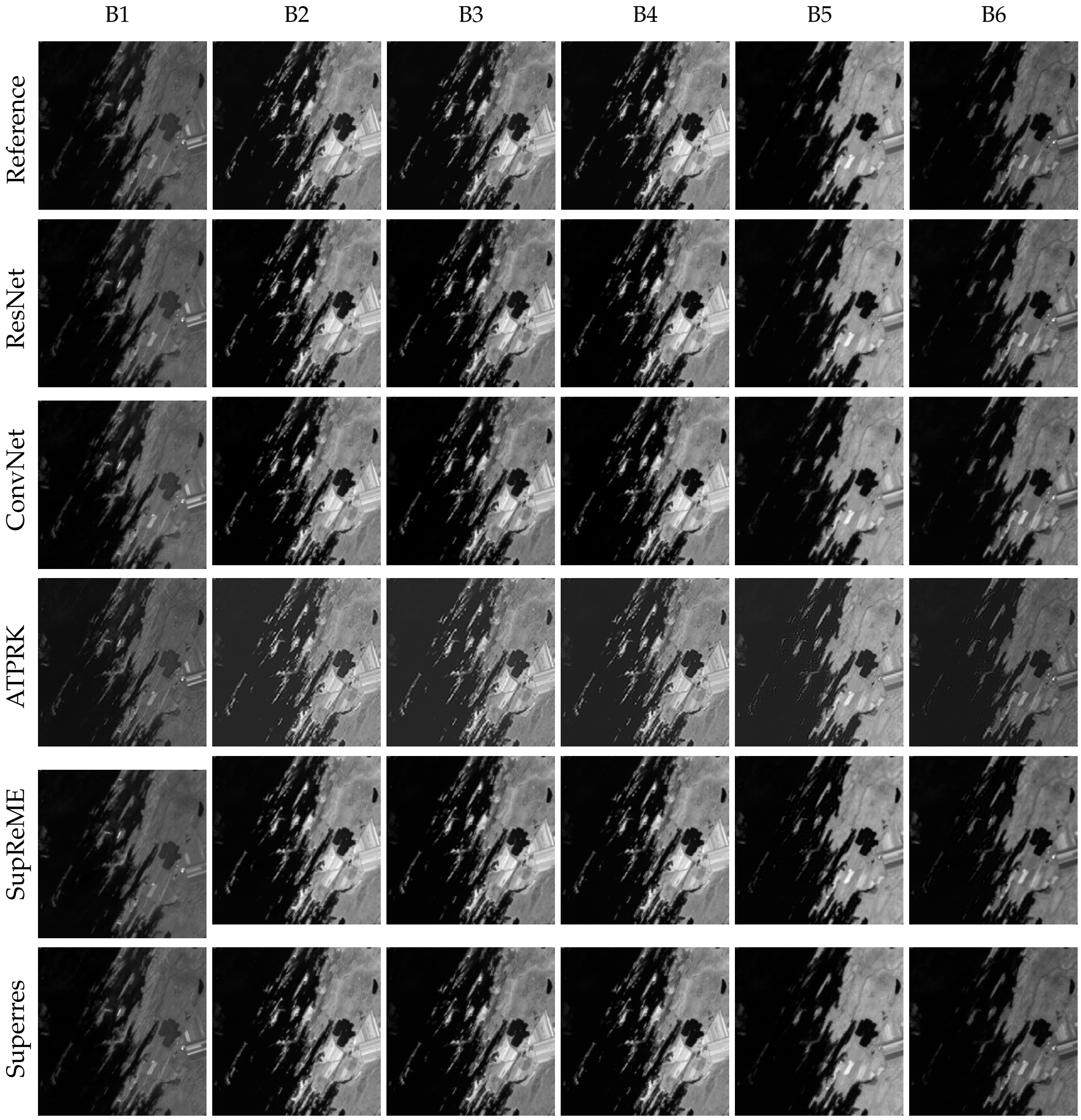


Figure 10. Visual comparison of all methods and all bands for the Coastal-A dataset.

The residuals for each band are shown in Figure 11. The ATPRK and Superres methods have the largest residuals while the proposed method and the ConvNet method produced images whose residuals have the least structure.

Finally, Figure 12 shows the SRE values for each band, for all methods and all the datasets. For the Coastal datasets in sub-figures (a) and (b), respectively, the CNN based methods give results that are substantially better than the other methods used in the comparison. Interestingly, the SupReME method shows a trend in the band-wise SRE values that is different from what the other methods show. For the urban datasets in (c) and (d), all methods show a similar trend, and now the SupReME method performs relatively better compared to ResNet.

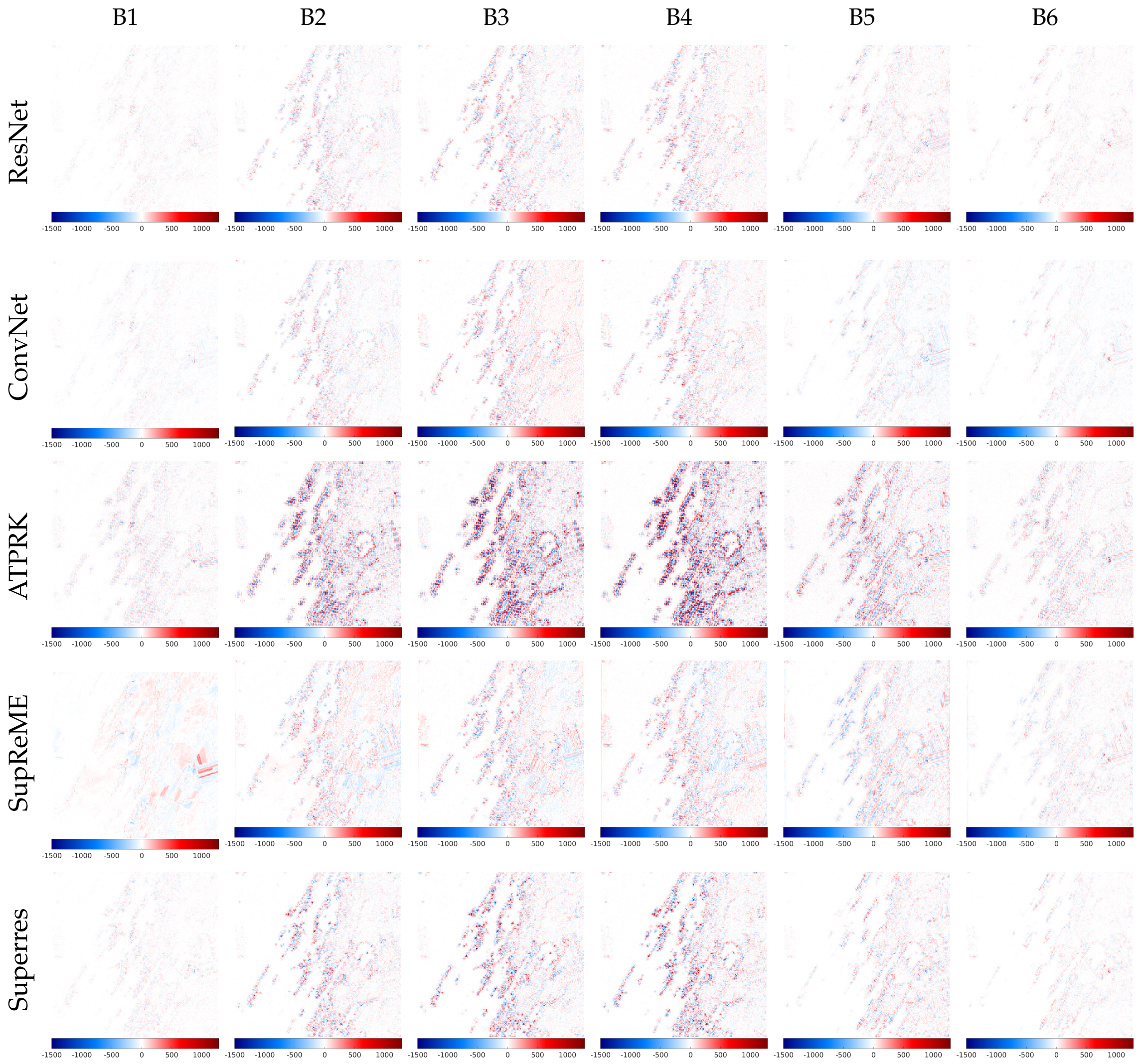
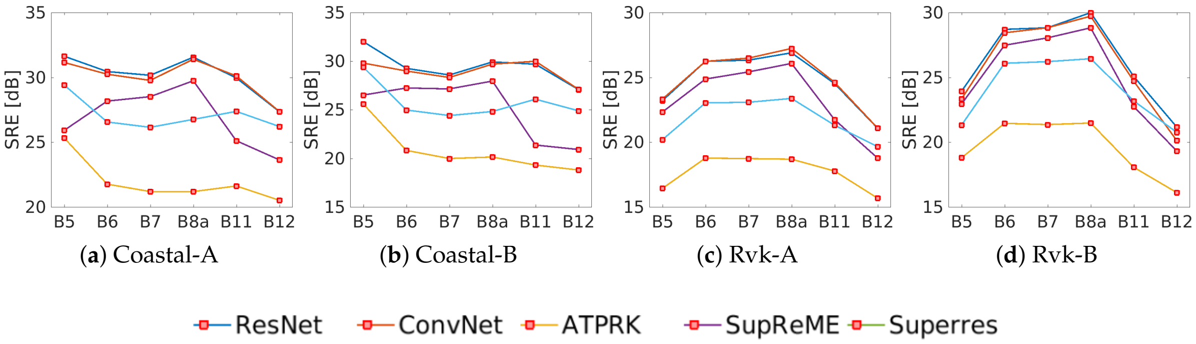


Figure 11. Residual images for all methods and bands of the Coastal-A data set. Blue indicates negative residuals and red indicates positive residuals.



(**a**) Coastal-A (**b**) Coastal-B (**c**) Rvk-A (**d**) Rvk-B

Figure 12. Plot of the band-wise SRE for all datasets and methods.

3.6. 60 m Bands

For the sharpening of the 60 m bands, we use the same network architecture as the 20 m bands. In order to be able to use the observed 60 m bands as targets during training, the input image has to be downgraded in resolution by a factor of 6 and now the input also contains the 60 m bands, thus the network uses information from both the 10 m bands and the 20 m bands to produce the sharpened 60 m bands. After downgrading by factor 6, the 60 m, 20 m , and 10 m bands are at resolution 360 m, 120 m, and 60 m, respectively. For all input bands to be of the same size, the downgraded 60 m and 20 m bands need to be interpolated by a factor of 6 and 3, respectively. To be able to quantitatively evaluate the fusion performance, a reference image is needed. As with the 20 m sharpening, this can be achieved by reducing the observed image by a factor of 6 and using the observed 60 m bands as the reference. This means that the input to the network during training has been downgraded in resolution by a factor of 36. Thus, the only difference between the proposed method for 20 m and 60 m super-resolution is that for the latter, the 60 m bands are added to the input, during training, the input data are downgraded by a factor of 6 instead of 2, and now the observed 60 m bands serve as targets.

The dataset is a 3452 by 3452 pixel subset of the same Level-1C product as for the previous datasets. It shows a part of the western coast of Iceland, including the fjord of Borgarfjörður and its surroundings. An RGB rendering of the data set is shown in Figure 13.

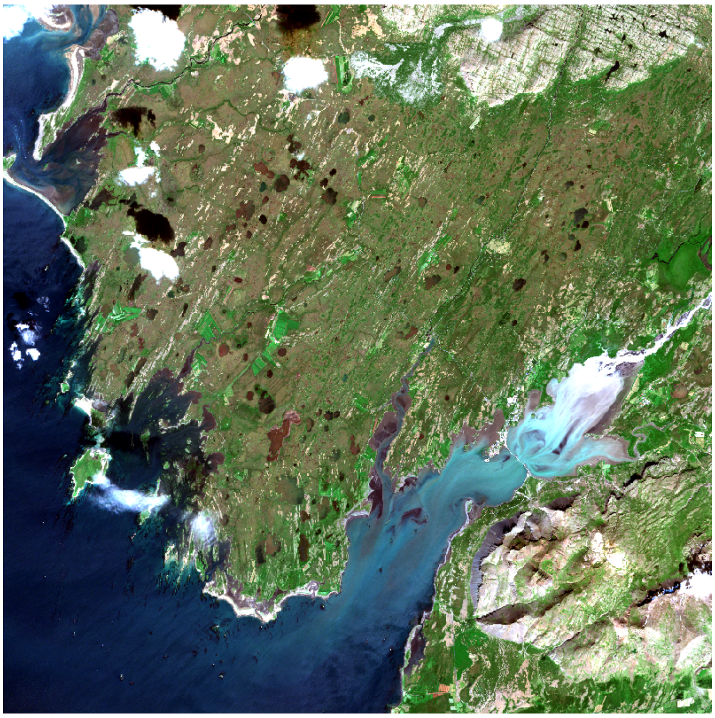


Figure 13. RGB image of the dataset used for the quantitative evaluation of all methods for 360 m to 60 m super-resolution.

For this experiment, we use the same network parameters as for the 20 m bands. The number of residual blocks is 24, the batch size is 48, and the patch size is 16 by 16 pixels. For ResNet, the number of training epochs was set to 200, while for ConvNet the number of epochs was set to 500, and the batch size was set to 16. For the SupReME method, the dimension of the subspace was chosen as 7, and the value of the regularization parameter λ was determined by performing a simple linear search of 30 values. The ATPRK method cannot do sharpening of the 60 m bands and the Superres method is limited to the sharpening of the 20 m bands in reduced resolution mode; therefore, these methods are not included in this experiment.

The results are summarized in Table 2. The results for the neural network based methods is the mean of 30 trials. The proposed method gives the best results with the ConvNet method giving the second best results. The performance difference between the two methods is larger than for the super-resolution of the 20 m bands. This also applies to the SupReME method, which now performs relatively worse. The SAM value for SupReME is more than twice higher than for the proposed method, and the average SRE value is lower by almost 6 dB.

Table 2. Quantitative evaluation results for 360 m to 60 m super-resolution. The data have been degraded six-fold in resolution in order to use the observed 60 m bands as the reference image. Columns B1 and B9 are the band-wise SRE values, while aSRE is the average SRE of the bands and is given in dB. SAM is given in degrees. Bold typeface indicates the best results.

Method	B1	B9	aSRE	ERGAS	SAM
ResNet	**33.37**	**26.09**	**29.73**	**0.69**	**0.72**
ConvNet	32.00	24.26	28.13	0.92	0.95
SupReME	26.82	18.25	22.53	1.80	2.02

Figure 14 shows band B1 and B9 obtained using all methods as well as the reference. The images produced by the two CNN based methods look noticeably sharper and more detailed than the images obtained using the SupReME method. This applies especially to band B9. The corresponding residual plots are shown in Figure 15 and there one can see that the residuals for the SupReME method contain more information and structure than for the proposed method, especially for band B9. Finally, visual results for 60 m to 10 m super-resolution are shown in Figure 16, where a subset of the dataset showing the town of Borgarnes in the western part of Iceland. The results are shown in false color using band B1 as the red and blue channels and B9 as the green channel. The colors in the image obtained using the proposed method look closer to the colors in the observed bands B1 and B9 than for the SupReME method, where the colors look darker. In addition, the details obtained using the proposed method seem more natural, and there are fewer artifacts, such as ringing around sharp edges, which are visible in the SupReME results.

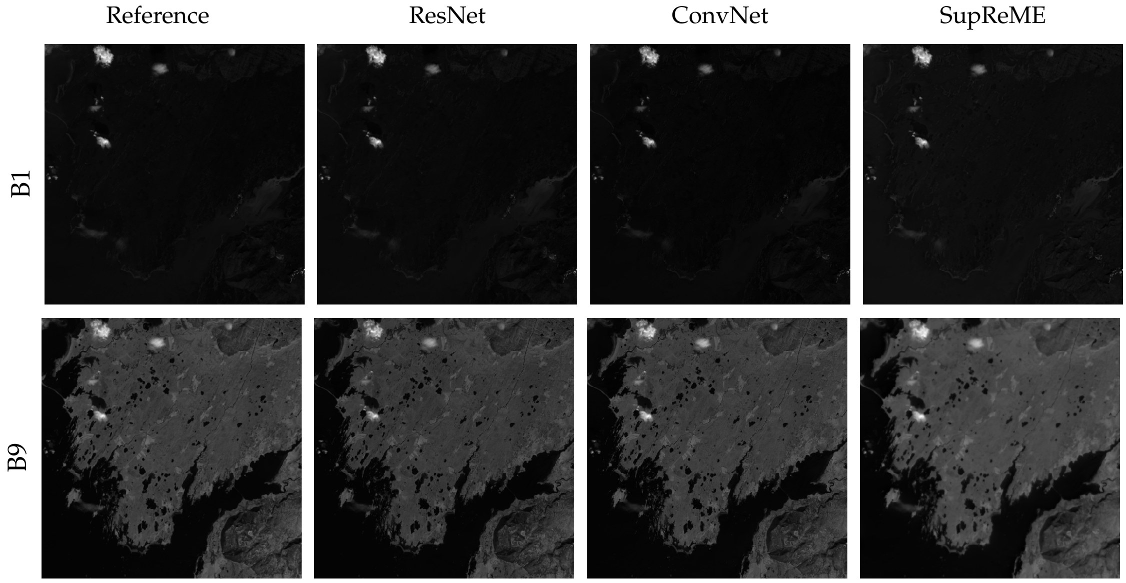


Figure 14. Visual comparison of all methods for bands B1 and B9, which have been fused at the reduced resolution of 360 m to obtain 60 m resolution image.

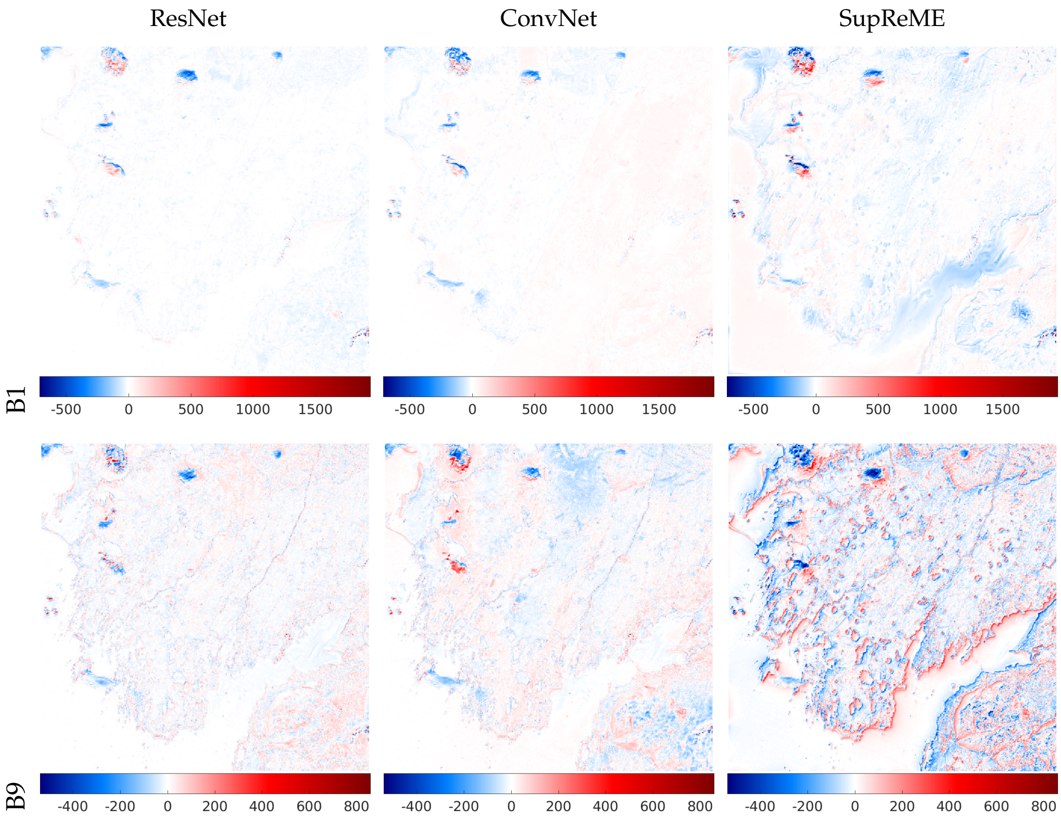


Figure 15. Residual plots for bands B1 and B9 obtained for all methods at 60 m resolution. Residuals for B1 are shown in the top row while residuals for B9 are shown in the bottom row.

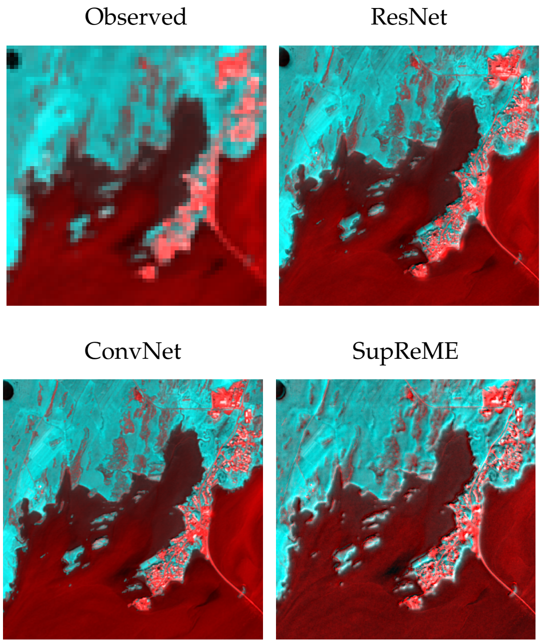


Figure 16. Full-scale (10 m) fusion results of 60 m bands rendered in false color for all methods. The subset shows the town of Borgarnes in Borgarfjörður in western Iceland.

4. Conclusions

In this paper, we have proposed a deep ResNet for the super-resolution of 20 m and 60 m Sentinel-2 bands, and focused on the single image scenario, where the amount of training data is limited. For experiments involving 20 m data, we used two pairs of datasets, one urban and one rural, which all originate from the same Sentinel-2 level-1C product showing a large portion of the western part of Iceland. For experiments using 60 m data, we used a single large portion of the original tile. The first set of experiments, involving twice reduced 20 m data, focused on the effects of the various hyperparameters on the quantitative quality of the fused bands, as measured by three quantitative metrics. We trained on one image and tested on the other, for each pair of datasets. This study revealed that the performance of the ResNet is relatively insensitive to important hyperparameters such as the number of residual blocks and patch-size. Comparison of the proposed method to several state-of-the-art methods, involving both 20 m and 60 m bands at reduced resolution, showed that the proposed method gives the best results, and often outperforming the comparison methods by a significant margin, especially when performing sharpening of the 60 m bands.

Author Contributions: F.P. did all the experiments and F.P., J.R.S., and M.O.U. wrote the paper.

Funding: This research was funded in part by The Icelandic Research Fund grant number 174075-05.

Acknowledgments: We would like to thank Charis Lanaras for providing the source code for the SupReME method.

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

Abbreviations

The following abbreviations are used in this manuscript:

NIR	Near Infrared
SWIR	Shortwave Infrared
MS	Multi Spectral
PAN	Panchromatic
HS	Hyper Spectral
VNIR	Visible Near Infrared
ATPRK	Area-to-Point Regression Kriging
SupReME	Super-Resolution for Multispectral Multiresolution Estimation
CNN	Convolutional Neural Network
DEM	Digital Elevation Model
SRE	Signal-to-Reconstruction Error
SAM	Spectral Angle Mapper

References

1. Vivone, G.; Alparone, L.; Chanussot, J.; Dalla Mura, M.; Garzelli, A.; Licciardi, G.; Restaino, R.; Wald, L. A Critical Comparison Among Pansharpening Algorithms. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 2565–2586. [CrossRef]
2. Palsson, F.; Sveinsson, J.R.; Ulfarsson, M.O. A New Pansharpening Algorithm Based on Total Variation. *IEEE Geosci. Remote Sens Lett.* **2014**, *11*, 318–322. [CrossRef]
3. Palsson, F.; Sveinsson, J.R.; Benediktsson, J.A.; Aanaes, H. Classification of Pansharpened Urban Satellite Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 281–297. [CrossRef]
4. Sirguey, P.; Mathieu, R.; Arnaud, Y.; Khan, M.M.; Chanussot, J. Improving MODIS Spatial Resolution for Snow Mapping Using Wavelet Fusion and ARSIS Concept. *IEEE Geosci. Remote Sens. Lett.* **2008**, *5*, 78–82. [CrossRef]
5. Licciardi, G.; Khan, M.; Chanussot, J. Fusion of hyperspectral and panchromatic images: A hybrid use of indusion and nonlinear PCA. In Proceedings of the 2012 19th IEEE International Conference on Image Processing, Orlando, FL, USA, 30 September–3 October 2012; pp. 2133–2136.

6. Licciardi, G.; Khan, M.M.; Chanussot, J.; Montanvert, A.; Condat, L.; Jutten, C. Fusion of Hyperspectral and panchromatic images using multiresolution analysis and nonlinear PCA band reduction. *EURASIP J. Adv. Signal Process.* **2011**, *207*, 1783–1786.
7. Capobianco, L.; Garzelli, A.; Nencini, F.; Alparone, L.; Baronti, S. Spatial enhancement of hyperion hyperspectral data through ALI panchromatic image. In Proceedings of the 2007 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Barcelona, Spain, 23–28 July 2007; pp. 5158–5161.
8. Zhao, Y.; Yang, J.; Chan, J.C.W. Hyperspectral Imagery Super-Resolution by Spatial-Spectral Joint Nonlocal Similarity. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 2671–2679. [CrossRef]
9. Licciardi, G.; Veganzones, M.A.; Vivone, G.; Loncan, L.; Chanussot, J. Impact of hybrid pansharpening approaches applied to hyperspectral images. In Proceedings of the 2015 7th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Tokyo, Japan, 2–5 June 2015.
10. Picone, D.; Restaino, R.; Vivone, G.; Addesso, P.; Chanussot, J. Pansharpening of hyperspectral images: Exploiting data acquired by multiple platforms. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 7220–7223. [CrossRef]
11. Palsson, F.; Sveinsson, J.R.; Ulfarsson, M.O. Multispectral and Hyperspectral Image Fusion Using a 3-D-Convolutional Neural Network. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 639–643. [CrossRef]
12. Palsson, F.; Sveinsson, J.R.; Ulfarsson, M.O.; Benediktsson, J.A. Model based PCA/wavelet fusion of multispectral and hyperspectral images. In Proceedings of the 2014 IEEE Geoscience and Remote Sensing Symposium, Quebec City, QC, Canda, 13–18 July 2014; pp. 1532–1535. [CrossRef]
13. Nezhad, Z.H.; Karami, A.; Heylen, R.; Scheunders, P. Fusion of Hyperspectral and Multispectral Images Using Spectral Unmixing and Sparse Coding. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 2377–2389. [CrossRef]
14. Zhang, Y.; Wang, Y.; Liu, Y.; Zhang, C.; He, M.; Mei, S. Hyperspectral and multispectral image fusion using CNMF with minimum endmember simplex volume and abundance sparsity constraints. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 1929–1932.
15. Wei, Q.; Bioucas-Dias, J.; Dobigeon, N.; Tourneret, J.Y. Hyperspectral and Multispectral Image Fusion Based on a Sparse Representation. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3658–3668. [CrossRef]
16. Zhukov, B.; Oertel, D.; Lanzl, F.; Reinhackel, G. Unmixing-based multisensor multiresolution image fusion. *IEEE Trans. Geosci. Remote Sens.* **1999**, *37*, 1212–1226. [CrossRef]
17. Gomez, R.B.; Jazaeri, A.; Kafatos, M. Wavelet-based hyperspectral and multispectral image fusion. *Geo-Spatial Image Data Exploit. II* **2001**, *4383*, 36–42.
18. Kim, Y.; Choi, J.; Han, D.; Kim, Y. Block-Based Fusion Algorithm With Simulated Band Generation for Hyperspectral and Multispectral Images of Partially Different Wavelength Ranges. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 2997–3007. [CrossRef]
19. Wei, Q.; Dobigeon, N.; Tourneret, J.Y. Bayesian fusion of multispectral and hyperspectral images with unknown sensor spectral response. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 698–702. [CrossRef]
20. Yokoya, N.; Chanussot, J.; Iwasaki, A. Hyperspectral and multispectral data fusion based on nonlinear unmixing. In Proceedings of the 2012 4th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Shanghai, China, 4–7 June 2012; pp. 1–4. [CrossRef]
21. Chen, Z.; Pu, H.; Wang, B.; Jiang, G.M. Fusion of Hyperspectral and Multispectral Images: A Novel Framework Based on Generalization of Pan-Sharpening Methods. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 1418–1422. [CrossRef]
22. Grohnfeldt, C.; Zhu, X.X.; Bamler, R. Jointly sparse fusion of hyperspectral and multispectral imagery. In Proceedings of the 2013 IEEE International Geoscience and Remote Sensing Symposium—IGARSS, Melbourne, Australia, 21–26 July 2013; pp. 4090–4093. [CrossRef]
23. Aiazzi, B.; Alparone, L.; Baronti, S.; Santurri, L.; Selva, M. Spatial resolution enhancement of ASTER thermal bands. *Image Signal Process. Remote Sens. XI* **2005**, *5982*, 59821G.
24. Adelson, E.H.; Anderson, C.H.; Bergen, J.R.; Burt, P.J.; Ogden, J.M. Pyramid methods in image processing. *RCA Eng.* **1984**, *29*, 33–41.

25. Pereira, M.J.; Ramos, A.; Nunes, R.; Azevedo, L.; Soares, A. Geostatistical Data Fusion: Application to Red Edge Bands of Sentinel 2. In Proceedings of the 2016 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 15–17 December 2016; pp. 758–761. [CrossRef]
26. Wang, Q.; Shi, W.; Atkinson, P.M.; Zhao, Y. Downscaling MODIS images with area-to-point regression kriging. *Remote Sens. Environ.* **2015**, *166*, 191–204. [CrossRef]
27. Lanaras, C.; Bioucas-Dias, J.; Baltsavias, E.; Schindler, K. Super-Resolution of Multispectral Multiresolution Images from a Single Sensor. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1505–1513.
28. Ulfarsson, M.; Mura, M.D. A low-rank method for Sentinel-2 sharpening using cyclic descent. In Proceedings of the 2018 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Valencia, Spain, 22–27 July 2018.
29. Brodu, N. Super-Resolving Multiresolution Images With Band-Independent Geometry of Multispectral Pixels. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4610–4617. [CrossRef]
30. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
31. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Handwritten digit recognition with a back-propagation network. In *Neural Networks, Current Applications*; Chappman and Hall: London, UK, 1992.
32. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
33. Huang, W.; Xiao, L.; Wei, Z.; Liu, H.; Tang, S. A New Pan-Sharpening Method With Deep Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1037–1041. [CrossRef]
34. Masi, G.; Cozzolino, D.; Verdoliva, L.; Scarpa, G. Pansharpening by Convolutional Neural Networks. *Remote Sens.* **2016**, *8*, 594. [CrossRef]
35. Yuan, Q.; Wei, Y.; Meng, X.; Shen, H.; Zhang, L. A Multi-Scale and Multi-Depth Convolutional Neural Network for Remote Sensing Imagery Pan-Sharpening. *arXiv* **2017**, arXiv:1712.09809.
36. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [PubMed]
37. Lanaras, C.; Bioucas-Dias, J.M.; Galliani, S.; Baltsavias, E.; Schindler, K. Super-Resolution of Sentinel-2 Images: Learning a Globally Applicable Deep Neural Network. *arXiv* **2018**, arXiv:1803.04271.
38. Lorenzo, P.R.; Nalepa, J.; Kawulok, M.; Ramos, L.S.; Pastor, J.R. Particle Swarm Optimization for Hyper-parameter Selection in Deep Neural Networks. In *Proceedings of the Genetic and Evolutionary Computation Conference*; ACM: New York, NY, USA, 2017; pp. 481–488. [CrossRef]
39. Lorenzo, P.R.; Nalepa, J.; Ramos, L.S.; Pastor, J.R. Hyper-parameter Selection in Deep Neural Networks Using Parallel Particle Swarm Optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*; ACM: New York, NY, USA, 2017; pp. 1864–1871. [CrossRef]
40. Bergstra, J.S.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for Hyper-Parameter Optimization. In *Advances in Neural Information Processing Systems 24*; Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F., Weinberger, K.Q., Eds.; Curran Associates, Inc.: New York, NY, USA, 2011; pp. 2546–2554.
41. Loshchilov, I.; Hutter, F. CMA-ES for Hyperparameter Optimization of Deep Neural Networks. *arXiv* **2016**, arXiv:1604.07269.
42. Snoek, J.; Larochelle, H.; Adams, R.P. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*; NIPS Foundation: San Diego, CA, USA, 2012; pp. 2951–2959.
43. Domhan, T.; Springenberg, J.T.; Hutter, F. Speeding Up Automatic Hyperparameter Optimization of Deep Neural Networks by Extrapolation of Learning Curves. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015; pp. 3460–3468.
44. 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, NV, USA, 27–30 June 2016; pp. 770–778.
45. Wald, L.; Ranchin, T.; Mangolini, M. Fusion of satellite images of different spatial resolutions: assessing the quality of resulting images. *Photogramm. Eng. Remote Sens.* **1997**, *63*, 691–699.

46. Palsson, F.; Sveinsson, J.R.; Ulfarsson, M.O.; Benediktsson, J.A. Quantitative Quality Evaluation of Pansharpened Imagery: Consistency Versus Synthesis. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 1247–1259. [CrossRef]
47. Wang, Q.; Shi, W.; Li, Z.; Atkinson, P.M. Fusion of Sentinel-2 images. *Remote Sens. Environ.* **2016**, *187*, 241–252. [CrossRef]
48. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; NIPS Foundation: San Diego, CA, USA, 2012; pp. 1097–1105.
49. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
50. Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **1994**, *5*, 157–166. [CrossRef] [PubMed]
51. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010; pp. 249–256.
52. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.
53. Klambauer, G.; Unterthiner, T.; Mayr, A.; Hochreiter, S. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems*; NIPS Foundation: San Diego, CA, USA, 2017; pp. 972–981.
54. Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA, 11–13 April 2011; pp. 315–323.
55. Maas, A.L.; Hannun, A.Y.; Ng, A.Y. Rectifier nonlinearities improve neural network acoustic models. In Proceedings of the 30th International Conference on Ma-chine Learning, Atlanta, GA, USA, 17–19 June 2013; p. 3.
56. Burt, P.; Adelson, E. The Laplacian pyramid as a compact image code. *IEEE Trans. Commun.* **1983**, *31*, 532–540. [CrossRef]
57. Van der Walt, S.; Schönberger, J.L.; Nunez-Iglesias, J.; Boulogne, F.; Warner, J.D.; Yager, N.; Gouillart, E.; Yu, T.; The Scikit-Image Contributors. Scikit-image: Image processing in Python. *PeerJ* **2014**, *2*, e453. [CrossRef] [PubMed]
58. Chollet, F. Keras. Available online: https://github.com/fchollet/keras (accessed on 15 April 2018).
59. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available online: https://www.tensorflow.org/ (accessed on 15 April 2018).
60. Wald, L. Quality of high resolution synthesized images: Is there a simple criterion? In Proceedings of the Third Conference "Fusion of Earth Data: Merging Point Measurements, Raster Maps and Remotely Sensed Images", Sophia Antipolis, France, 4 June 2000.