LEARNING SUPER-RESOLUTION FOR SENTINEL-2 IMAGES WITH REAL GROUND TRUTH DATA FROM A REFERENCE SATELLITE

M. Galar[1,*], R. Sesma[2], C. Ayala[2], L. Albizua[3], C. Aranda[2]

[1] Institute of Smart Cities (ISC), Public University of Navarre,
Campus de Arrosadía s/n, 31006, Pamplona, Spain - mikel.galar@unavarra.es

[2] Tracasa Instrumental, Calle Cabárceno, 6, 31621 Sarriguren, Navarra - (rsesma, cayala, caranda)@itracasa.es

[3] Tracasa, Calle Cabárceno, 6, 31621 Sarriguren, Navarra - lalbizua@tracasa.es

KEY WORDS: Super-resolution, deep learning, sentinel-2, convolutional neural networks, multi-spectral images

ABSTRACT:

Copernicus program via its Sentinel missions is making earth observation more accessible and affordable for everybody. Sentinel-2 images provide multi-spectral information every 5 days for each location. However, the maximum spatial resolution of its bands is 10m for RGB and near-infrared bands. Increasing the spatial resolution of Sentinel-2 images without additional costs, would make any posterior analysis more accurate. Most approaches on super-resolution for Sentinel-2 have focused on obtaining 10m resolution images for those at lower resolutions (20m and 60m), taking advantage of the information provided by bands of finer resolutions (10m). Otherwise, our focus is on increasing the resolution of the 10m bands, that is, super-resolving 10m bands to 2.5m resolution, where no additional information is available. This problem is known as single-image super-resolution and deep learning-based approaches have become the state-of-the-art for this problem on standard images. Obviously, models learned for standard images do not translate well to satellite images. Hence, the problem is how to train a deep learning model for super-resolving Sentinel-2 images when no ground truth exist (Sentinel-2 images at 2.5m). We propose a methodology for learning Convolutional Neural Networks for Sentinel-2 image super-resolution making use of images from other sensors having a high similarity with Sentinel-2 in terms of spectral bands, but greater spatial resolution. Our proposal is tested with a state-of-the-art neural network showing that it can be useful for learning to increase the spatial resolution of RGB and near-infrared bands of Sentinel-2.

1. INTRODUCTION

The European Space Agency under Sentinel missions are promoting and easing research on earth observation. Thanks to their open data initiative, data gathered from Sentinel satellites can be freely accessed, allowing research and multiple services to take advantage of this situation. Among the variety of Sentinel satellites, Sentinel-2 (S2) is focused on high-resolution optical imagery, having vegetation, soil and coastal areas as its main objectives (Drusch et al., 2012). Thirteen spectral bands are captured by the sensor of S2. These bands are in the visible/near infrared (VNIR) and short-wave infrared spectral range (SWIR) at different resolutions. The greatest spatial resolution provided is 10m for RGB and NIR bands, whereas the rest are given either at 20m or 60m.

Recent developments in single image super-resolution (SISR) (Yang et al., 2018) suggest that these spatial resolutions could be improved without using additional information. Having higher resolutions, the posterior analyses could be carried out with greater details. However, except for a few approaches (Liebel, Körner, 2016, Wagner et al., 2019), previous works have mainly developed methods for obtaining all thirteen bands at 10m resolution (Lanaras et al., 2018, Gargiulo et al., 2018). Although this is also a challenging setting, it has the advantage of having additional information at 10m resolution (RGB and NIR bands) that can help during the super-resolution of the other bands. Anyway, these strategies cannot be applied to improve the resolution of 10m bands to 5m or 2.5m, since no reference data at these resolutions exist.

Currently, deep learning-based methods have become the standard for image processing and computer vision (Goodfellow et al., 2016). To deal with images, Convolutional Neural Networks (CNNs) (Lecun et al., 1998) are usually considered. SISR (Yang et al., 2018) is a scenario where CNNs have excelled. A number of methods have been proposed in the literature for this purpose (Kim et al., 2016, Ledig et al., 2017). Their advantage with respect to more classical super-resolution methods such as bicubic interpolation or reconstruction methods (Yan et al., 2015) has been already proven. Although networks trained for standard images do not translate well to satellite images for different reasons (Liebel, Körner, 2016), they seem to be the way to go for increasing the resolution of S2 10m resolution bands.

Therefore, training a specific network for improving the resolution of S2 10m bands is required, which must take into account the characteristics of these kinds of images. In this work, we propose a full end-to-end CNN to super-resolve RGB and NIR bands to 2.5m resolution. However, notice that CNNs for super-resolution fall in the category of supervised machine learning. Consequently, the network is trained with labeled data, which means that for each S2 image used to train the network, the same scene at 4 time more resolution is required. Indeed, this is the main problem addressed in this work.

The most straightforward way to create pairs of images for super-resolution would be to consider S2 10m bands as the target resolution and downsample the same image to 40m. This way, a network could be learned for super-resolving 40m to 10m, with the assumption that this network would then translate well to super-resolve 10m resolution to 2.5m resolution. Similar approaches can be found in (Liebel, Körner, 2016, Wagner et al., 2019). However, this is somewhat similar to the effect of apply-

*Corresponding author

ing CNNs trained for standard super-resolution to satellite images, not all characteristics are properly learned and conquently, the CNN does not generalize as it could be expected.

For this reason, our main motivation is to consider real ground truth images at 2.5m resolution. This is in line with the approach in (Galar et al., 2019), where the authors proposed to use another satellite (RapidEye) to improve S2 RGB bands resolution to 5m. However, we consider several important aspects that were not required when dealing with RapidEye and 2x super-resolution. Moreover, we also introduce the usage of NIR band, move from 8bit radiometric resolution to the native 16bit of S2 images, avoid manual validation and propose a way to match the reflectance of both satellites. The work in (Beaulieu et al., 2018) also follows the same idea for 2x super-resolution, but no clear justification for the satellite is given and very limited experiments are carried out. Hence, in this work we propose to use satellites with similar spectral bands to those of S2 as a source for target images for training a neural network. With this aim, we have opted for PlanetScope (PS) constellation of Dove satellites [1]. Although they provide images at 3.125m resolution (in the case of the Ortho Tile product we have used), we resample them to 2.5m for easing the architecture and learning of the CNN. As we detail in Section 3, proper preprocessing is the key for having good pairs of S2 and PS images.

The experimental study to validate the proposed approach consists of a set of images obtained from Open California, which were freely available from[2]. We consider a state-of-the-art model called EDSR (Enhanced Deep Residual Networks) (Lim et al., 2017) with some modifications to avoid checkerboard patterns (Aitken et al., 2017, Sugawara et al., 2018). We evaluate different strategies for learning the network using commonly considered metrics for super-resolution: the peak signal to noise ratio (PSNR) and the structural similarity (SSIM) (Zhou Wang et al., 2004). We will show that the proposed methodology leads to promising results and spectral validation shows that super-resolution preserves the coherence with the original S2 image. We must keep in mind that preserving the radiometric information when super-resolving images is of vital importance in satellite imagery.

The rest of this work is organized as follows. Section 2, briefly recalls deep learning and CNNs, mainly focusing on SISR. Afterwards, we present our proposal for super-resolving S2 images in Section 3. The experiments are carried out in Section 4. Finally, conclusions and future work are presented in Section 5.

2. PRELIMINARIES

CNNs (Lecun et al., 1998) are currently the standard to address computer vision tasks due to their performance. Although image classification was the first problem addressed by CNNs, their usage have been extended to a series of problems in computer vision, including SISR. Hereafter, we briefly recall several approaches for this purpose in Section 2.1 and focus on EDSR model (the model considered for the experiments) and the modifications we carried out in Section 2.2.

2.1 CNNs for Single Image Super-Resolution

SISR is the problem of increasing the spatial resolution of an image using the information in the image itself together with some knowledge acquired in the form of an algorithm or a model (Yang et al., 2018). There are three types of methods for doing SISR: interpolation-based (bicubic interpolation (Keys, 1981)), reconstruction-based (applying prior knowledge to generate sharp details (Yan et al., 2015)) and learning-based methods (learning a model from source and target data (Yang et al., 2018)). We focus on the latter and more specifically, on deep learning (CNN) approaches due to the excellent results they have shown working with standard images (Yang et al., 2018).

To learn a CNN for SISR, one needs to train the network with pairs of images at low (source) and high (target) resolution. Then, the network is expected to find high-level abstractions from the low resolution image to bridge the gap with respect to the high resolution space. The common way for obtaining these pairs is to consider very high quality images as target and downsample them to obtain the source images. However, the problem in this work is that there are no S2 images available at high resolution (2.5m).

With the training set available, different architectures and optimization objectives have been developed (Yang et al., 2018). SRCNN (Dong et al., 2014), the first CNN for SISR, used a bicubic interpolation of the source image as input to the network, resulting in high computational costs. VDSR (Kim et al., 2016) followed the same idea but increased the depth of the network. One way to reduce the computational effort by methods using bicubic interpolation as input was presented in (Shi et al., 2016), as a part of ESPCN. The authors proposed a Pixel Shuffle layer with sub-pixel convolution for upsampling at the end of the network. Later, ICNR initialization (Aitken et al., 2017) of these layers allowed to remove the checkerboard pattern recurrently appearing in CNN-based approaches, which was further improved by blurring in (Sugawara et al., 2018). EDSR (Lim et al., 2017) followed the same idea of upsampling at the end of the network. It was based on SRResNet (Ledig et al., 2017), which stacked several ResBlocks commonly used for image classification, but unnecessary modules were removed and the loss function was changed from L2 to L1 norm. We focus on this network in the next section as it is the base for our proposal.

EDSR was not the only proposal dealing with the loss function used for training the CNN. The L2 norm, i.e., Mean Square Error (MSE), has been the most widely used loss function. In EDSR, the authors justified using the L1 norm, i.e, Mean Absolute Error (MAE), for its better convergence. Using Generative Adversarial Networks (GANs) can also be understood as a different form of training. In SRGAN (Ledig et al., 2017), the authors trained a SRResNet using GAN learning, that is, having a discriminator network to distinguish between super-resolved and real high resolution images. Although this led to good visual results, this is not usually reflected on the numerical evaluation due to their ability to picture missing pixels.

2.2 EDSR: Enhanced Deep Residual Networks

EDSR modified SRResNet according to the properties of SISR. Batch normalization was removed from ResBlocks, in such a way that information suffer less changes, which is desired in SISR. Additionally, a residual scaling factor was added to stabilize learning (default value of 0.1).

Two main parameters define the architecture of EDSR: the number of ResBlocks and the number of filters. We consider the simplest version of EDSR with 8 ResBlocks and 64 filters, which

[1] https://directory.eoportal.org/web/eoportal/satellite-missions/d/dove
[2] https://www.planet.com/trial/

has a good trade-off between accuracy and complexity. After the 8 ResBlocks, Pixel Shuffle upsampling is used to increase the resolution of the image at the end of the network, making EDSR faster than previous alternatives using bicubic interpolation as input to the network. In the case of 4x super-resolution, the upsampling is implemented by doing Pixel Shuffle twice (each time duplicating the number of pixels). This allows using 2x super-resolution as a pretrained model for 4x super-resolution, making convergence faster. The network simply needs to be enlarged adding another Pixel Shuffle layer duplicating the number of pixels. We will test the usefulness of this strategy in the experiments. A scheme of the EDSR used in this work is presented in Figure 1.

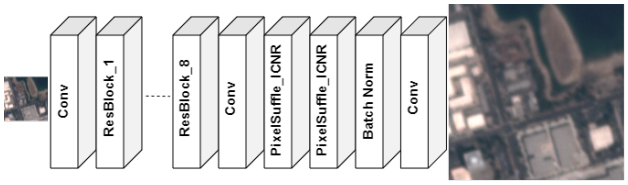

Figure 1. Architecture of EDSR8.

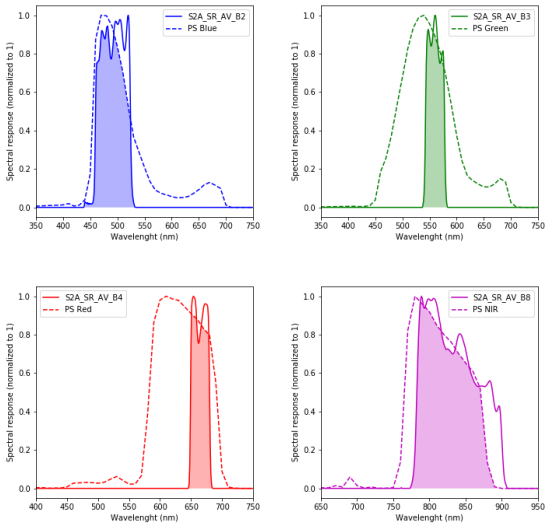

Figure 2. Comparison between PS (group 0exx) and S2 (A) spectral response functions for RGB and NIR.

3. S2PS: SENTINEL-2 TO PLANETSCOPE

In this section, we describe our proposal for super-resolving S2 images, using images coming from PS as a reference (target) to learn a modified EDSR network. The motivation for the usage of PS is presented in Section 3.1. Details about how we have generated the training set are given in Section 3.2, whereas the specific properties of the EDSR network that we have implemented and its learning are presented in Section 3.3.

3.1 Proposal

The problem we want to address in this work is that we do not have S2 images at 2.5m resolution to be able to learn a CNN for SISR. As a consequence, we tried to find a sensor that matches the spectral bands of S2, also providing higher resolution images. We found that PS constellation of Dove satellites [3] operating since 2016 could be a good candidate for our purpose. The Ortho Tile Analytic products of PS (the ones used in this work) are provided at 3.125m[4] and hence, we resample them to 2.5m using cubic interpolation for easing the architecture and learning of the network. We acknowledge that in this case, we are not truly super-resolving S2 4x, but we are close to it. The fact that spectral bands were similar and the possibility of accessing the images of PS freely with Open California program (detailed thereafter) were key points to opt for PS. Figure 2 depicts how the spectral bands of S2 and PS match. More specifically, for PS those corresponding to 0exx group of Doves is depicted (there are different groups of PS satellites with slightly different spectral bands).

Evidently, ortho products from PS and S2 are provided in different processing levels and magnitudes. S2 is provided in reflectance with scale factor of 10000 and the processing level can be selected (e.g., L1C is top of atmosphere reflectance and L2A is bottom of atmosphere reflectance). In the case of the product we were able to download for PS, the rasters contained digital numbers, as most of the satellite images are supplied. We used the metadata of the product to convert these numbers to top of atmosphere reflectance. We will detail every aspect regarding the processing of the images for learning the network in the next section. We should notice that we restrict our experiments to top of atmosphere reflectance because this is the product we were able to download from PS. Notice that the same procedure could be followed with bottom of atmosphere reflectance if those products were available. As we explain in the next section, each S2 and PS pair of images were selected under the same acquisition conditions (same day, close time and close zenital observation angles) to ensure that TOA reflectance are comparable.

3.2 Datasets

We downloaded PS images using 14 days free trial from Planet[5] giving access to Open California. In the future, we plan to extend this study with strategically selected images from different regions of interest. Therefore, all the images in our dataset are from California state (United States of America, USA).

Although we are limited with respect to the images from PS, we have access to all the metadata from them. One key issue to build a proper dataset for SISR is that the difference between the input and target images should be minimal, except for the resolution. Commonly, this is achieved by learning to super-resolve by taking a very high resolution image as target and downsampling it to obtain the input to the network. In our case, we have images coming from different satellites and hence, we should make an effort to make them be almost the same, except for the resolution. To do so, in the following we explain the process we have followed.

First, we need to download product pairs from S2 and PS. We maximize the matching between the product of S2 and PS in terms of acquisition date. This means that we only download product pairs located in the same place and obtained in the same date. Moreover, we establish some requirements for the query: acquisition data between 01/01/2017 and 01/01/2020; cloud cover less than 10% in PS and equal to 0% in S2; minimum usable data in PS of 90%; the PS product must be completely inside the S2 product. Notice that these restrictions are

[3] https://directory.eoportal.org/web/eoportal/satellite-missions/d/dove

[4] https://assets.planet.com/docs/Planet_Combined_Imagery_Product_Specs_letter_screen.pdf

[5] https://www.planet.com/

not required when the network is deployed, since it can be used for super-resolving any S2 image in L1C level. With these constraints, we found 257 different areas. For the experiments in this work, we have considered the pairs in Table 1. We selected these pairs mainly focused on having a variety of scenarios. The selected locations are also presented in Figure 3. Notice that we divided these images into the common training/validation/test partitions, where a complete image goes always to a single partition, making the evaluation fairer.

Area	Date	S2 time	PS time	Set	#Tiles
La Habra	2018/08/29	18:29:09	18:00:59	test	2475
Shafter	2019/04/17	18:39:21	18:18:33	test	1969
Visalia	2018/09/10	18:39:21	18:12:50	val	2464
Ontario	2019/03/30	18:29:39	18:10:25	val	2435
Willits	2017/08/26	18:59:09	18:17:18	train	2610
Vallejo	2017/09/27	18:51:31	18:14:00	train	1394
Anderson	2019/06/20	18:49:21	18:37:39	train	2463
Folsom	2018/06/29	18:49:19	18:20:12	train	2428
Santa Rosa	2017/07/19	19:03:51	18:11:00	train	995
Patterson	2018/10/07	18:52:49	18:22:33	train	2144
Pasadena	2017/09/26	18:44:09	17:54:54	train	2600
Stockton	2018/10/17	18:53:59	18:19:31	train	2364

Table 1. Summary of the images used from S2 and PS to form our dataset.

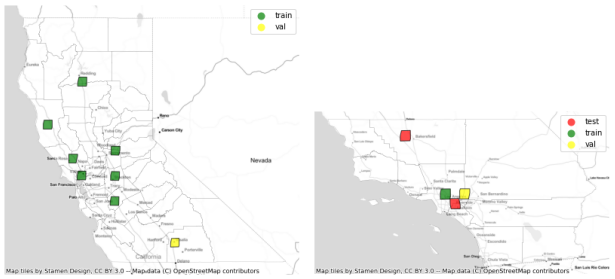

Figure 3. Location of the images considered for the study.

Once we have downloaded the product pairs, we obtain the top of atmosphere reflectance for PS so that both resulting images are comparable. Co-registration accuracy of each satellite may vary. We tried to overcome co-registration issues automatically performing small shifts between image pairs if necessary. In the case of S2, we only take bands corresponding to RGB and NIR, the same that are present in PS. The next step consist in dividing each image into tiles (patches) of 48×48 and 196×196 pixels in S2 and PS, respectively. These will be the inputs and targets to learn the network. Nevertheless, although one could expect to have very similar reflectance values in each pair of tiles, we observed that the color between them did not match exactly. This issue was caused by the difference in the spectral bands. Although we evaluated the possibility of matching the Spectral Bands Functions via SBAF (Pinto et al., 2018), we finally decided to follow a much simpler method that provided us with the desired output. Having both patches in the same place and almost in the same time set out the ideal situation for applying histogram matching between patches (Gonzalez, Woods, 2008). Histogram matching is the transformation of the image so that its histogram matches that of another image. Our idea is to transform the target image (from PS), so that it better resembles a S2 image at 2.5m. To do so, a mapping between the values of each band of PS and those of S2 is obtained for each tile. This is done by computing the cumulative histogram of each image and then, linearly interpolating unique pixel values in the PS tile that closely match the quantiles of the unique pixel values in S2 tile. This way, we end up with a PS tile that better matches the reflectance values of the corresponding S2 tile. This process also minimizes the effect of PS satellites having slightly different spectral bands. Figure 4 shows the effect and clear benefit of applying this process.

S2 PS PS + Histogram matching

Figure 4. Example of histogram matching applied to a PS tile with S2 tile as template.

After histogram matching, we still need to filter some undesirable patches that do not represent the ideal situation where both are the same except for the resolution (clouds may be present, flat surfaces too similar to be useful, i.e., sea, lakes, etc.). To do so, we designed an automatic validation process, where we computed the similarity between the corresponding patches of S2 and PS using the PSNR and SSIM metrics (explained in Section 3.4). To make both patches comparable, we applied bicubic interpolation to S2 image and upsample it to 196×196 pixels. By visual inspection, we established the following thresholds: we took patches with PSNR between 25 and 40 and from those, the ones with SSIM greater than 0.7. This way we only took patches that are good enough, avoiding those not providing interesting information. The result of this automatic validation process for the image over Vallejo area can be observed in Figure 5.

Sentinel-2 PlanetScope Validation Mask

Figure 5. Automatic patch validation over Vallejo area.

Finally, when using these pairs of tiles from S2 and PS to learn the network, we normalize them to $[0, 1]$ interval using the 12bit range. The following steps summarize the whole process:

1. To download PS Ortho Tile Analytic and S2 L1C products and consider RGB and NIR bands.
2. To convert PS data to Top of Atmosphere Reflectance and resample it to 2.5m (from 3.125m).
3. Match PS image with S2 image and crop accordingly (since PS products are smaller in size).
4. To create non-overlapping patches of 48×48 from S2 and 196×196 from PS that match.
5. Carry out histogram matching for each PS patch, making it resemble to be a S2 image at 2.5m.
6. To perform automatic validation of patches using the PSNR and SSIM thresholds.
7. To normalize patches to $[0, 1]$ interval using 12bit range.

After carrying out this process, we end up with the number of tiles for each image presented in Table 1. Recall that we divided

the set of images into training, validation and test sets. We will use training images to learn the network, validation set to decide which network we finally take during learning and finally, the test set to evaluate the model in a set of images that the network has never seen. The number of tiles obtained in each set is shown in Table 2. This numbers are close to those usually considered for training and evaluating machine learning models. Most of the data is used for training (approximately 65% of the patches) and less data is considered for validation and testing (approximately 20% for each set).

Set	#Images	#Patches	Ratio %
Train	8	16998	64%
Val	2	4899	19%
Test	2	4444	17%
Total	12	26341	100

Table 2. Number of patch pairs generated for training, validation and test.

3.3 Network training

There are several key issues that we must take into account for super-resolving S2. We are working with four channels, RGB and NIR, whereas most networks are usually designed for the typical RGB channels. This means that we have to properly adapt the network to receive four channels and also output another four channels. This modification is almost trivial, but it has several implications. The main one is with respect to the loss function considered. Previous works (Galar et al., 2019) showed that using a combination of pixel loss (L1), a feature loss based on VGG16 (Simonyan, Zisserman, 2014) and a style loss (Johnson et al., 2016) based on the same network, provided good results, avoiding the blurry effect of only focusing on pixel loss. Recall that the feature loss consists of computing the L1 loss between the activations obtained by the target and the super-resolved images when they are forwarded through VGG16 network. Otherwise, the style transfer tries to force the super-resolved image to have similar correlations to those of the target image among the activations of the different channels in several layers of VGG16.

In our case with RGB and NIR bands, the pixel loss can directly work with the four channels. Nonetheless, the VGG16 network used for the feature loss and the style loss is only pretrained for RGB images. To solve this issue, we divided the loss into two parts: RGB loss and NIR loss, each part having the same three components (pixel, feature and style losses). Then, we compute the losses as usual for RGB and we convert the NIR band into an RGB image by copying the same band into the three RGB channels. This way, we are able to go through VGG16 and obtain the corresponding loss for the NIR. Finally, to combine both parts we scale the losses so that all bands get the same importance in the final loss. Details on the weights to achieve it are given in Table 4.

With respect to our implementation of EDSR, we introduce some novelties apart from the specific loss function so as to completely avoid the appearance of any checkerboard patterns. We applied ICNR (Aitken et al., 2017) initialization combined with a blurring carried out by an average pooling operation (Sugawara et al., 2018). This mainly means that after Pixel Shuffle, values are averaged in 2×2 windows, eliminating any undesirable checkerboard pattern.

For training the network we carried out different tests with the progressive resizing idea presented by the authors of EDSR. The idea resides in first learning a model for 2x super-resolution and then, using this model to train the 4x super-resolution faster by simply adding another Pixel Shuffle layer to the previous one. This way of learning is expected to accelerate convergence and improve generalization. Having in mind that we are using images from a different sensor as target images, we will test different ways of performing this methodology to understand their advantages. To do so, we first resample the patches to different resolutions (either PS or S2 patches, depending on the experiments): 40m, 20m, 10m and 5m. Notice that PS patches are originally at 2.5m and S2 ones at 10m. It is clear that our final objective is to translate S2 images from 10m to 2.5m (S2PS). Nevertheless, we can first pretrain the network to learn 2x super-resolution and then use that network to faster and better learn 4x super-resolution. Furthermore, these pretrainings can be performed in different ways. All our experiments are summarized in Table 3, where we show the different models learned, including from which model they have started learning (pretrained model). They mainly differs on which satellite is used for pretraining and in which order we go from 2x to 4x super-resolution. We also want to remark that Baseline refers to the commonly used method for super-resolving S2, that is, learn from 40m to 10m (4x) and use it to super-resolve from 10m to 2.5 (also 4x). We use this as a reference for comparison together with bicubic interpolation. Moreover, model 4 refers to not carrying out progressive resizing.

Model	Pre	Name	Ep	Bs	Lr
Baseline	–	$S2^{40} \rightarrow S2^{10}$	50	256	1e-3
1.1	–	$PS^{10} \rightarrow PS^{5}$	10	96	5e-4
1.2	1.1	$PS^{5} \rightarrow PS^{2.5}$	10	24	5e-6,1e-5
1.3	1.2	$S2^{10} \rightarrow PS^{2.5}$	50	32	5e-4,5e-3
2.2	1.1	$S2^{10} \rightarrow PS^{5}$	10	96	5e-6,5e-5
2.3	2.2	$S2^{10} \rightarrow PS^{2.5}$	50	32	5e-4,5e-3
3.1	–	$S2^{10} \rightarrow PS^{5}$	10	96	1e-3
3.2	3.1	$S2^{10} \rightarrow PS^{2.5}$	50	32	5e-4,1e-3
4	–	$S2^{10} \rightarrow PS^{2.5}$	50	32	1e-3
5.1	–	$S2^{20} \rightarrow S2^{10}$	10	192	1e-3
5.2	5.1	$S2^{10} \rightarrow PS^{5}$	10	96	5e-6,5e-5
5.3	5.2	$S2^{10} \rightarrow PS^{2.5}$	50	32	5e-4,5e-3
5.4	5.1	$S2^{10} \rightarrow PS^{2.5}$	50	32	5e-4,5e-3

Pre: Pretrained model; Ep: Epochs; Lr: Learning rate;

Table 3. Configurations considered in the experiments.

Following current guidelines for training (Smith, 2018), we established the largest batch size fitting into the GPU memory (a NVIDIA RTX 2080Ti with 11GB of RAM), shown in Table 3. Likewise, we used one-cycle learning policy with learning rate finder to establish the learning rate to train each model (shown in Table 3). Other parameters used for training the network are provided in Table 4.

Parameter name	Value
VGG16 layers (feature/style losses)	First 3 Max-pooling inputs
VGG16 layer weights feature / style	0.2, 0.7, 0.1 / 200, 2450, 50
RGB / NIR Loss	0.75 / 0.25
Optimizer	Adam
Learning strategy / Weight decay	Once Cycle Policy (pct_start=0.7) / 1e-7

Table 4. Common parameters for all configurations.

3.4 Evaluation measures

The most widely applied metrics for super-resolution evaluation are considered in this work, both for evaluating the model and

for the automatic path validation process explained in Section 3.2. These metrics are the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) index (Zhou Wang et al., 2004).

The PSNR is tightly related to the mean square error between the super-resolved image (from S2) and the target image (PS with histogram matching):

$$\text{PSNR}(y, \hat{y}) = 10 \cdot \log_{10}(\frac{v_{max}^2}{\text{MSE}}) \quad (1)$$

where v_{max}^2 is the greatest possible difference between two pixel values.

On the contrary, the SSIM is more related to human perception, becoming more important than PSNR in certain scenarios such as ours. Recall that there is no real ground truth image of S2 at 2.5m and hence, we are working with an well-thought approximation obtained from PS.

4. EXPERIMENTAL STUDY

4.1 Results

In Table 5, we present the results in terms of PSNR and SSIM for all the models super-resolving 10m to 2.5m. Recall that their main difference is whether they perform progressive resizing and how they do it. Moreover, bicubic interpolation and baseline (learning with S2 to go from 40m to 10m) are used as a reference, although in the future we would like to extend this comparison with more complex methods.

Configuration	PSNR	SSIM
Bicubic	34.71	0.9230
Baseline $S2^{40} \rightarrow S2^{10}$	33.30	0.8921
1.3 $S2^{10} \rightarrow PS^{2.5}$	**35.47**	0.9387
2.3 $S2^{10} \rightarrow PS^{2.5}$	35.43	**0.9399**
3.2 $S2^{10} \rightarrow PS^{2.5}$	35.43	0.9392
4 $S2^{10} \rightarrow PS^{2.5}$	35.17	0.9342
5.3 $S2^{10} \rightarrow PS^{2.5}$	35.43	0.9392
5.4 $S2^{10} \rightarrow PS^{2.5}$	35.27	0.9390

Table 5. Results obtained by the different configurations in test set for both PSNR and SSIM.

Additionally, Figure 6 provides several examples of super-resolved patches so that the visual differences of the proposed super-resolution with bicubic and baseline model can be appreciated. For this purpose, we have considered model 2.3 as our proposal, the one achieving the best performance metrics (although the rest provide similar visual quality). Five examples of RGB bands are presented together with one example of NIR super-resolution

4.2 Discussion

We start analyzing the results of the different configurations in Table 5. We can observe that the worst performer is the baseline. Although this idea has been previously used for super-resolving S2 images (Liebel, Körner, 2016, Wagner et al., 2019), these experiments where real ground truth is considered show that it is not very accurate. This was partially expected as the details that can be observed when going from 40m to 10m and from 10m to 2.5m rather differ. Comparing bicubic interpolation with the rest of the models show that EDSR-based models get between 0.5 and 0.7 more points in terms of PSNR, whereas greater differences are obtained in terms of SSIM (1.2-1.7 points in all cases). The fact that we are not comparing with real S2 images seems to slightly benefit the PSNR for bicubic interpolation and its blurring effect. However, as it can be observed in Figure 6, the proposed method provides much sharper details with better defined edges, which explain the differences in terms of SSIM. In general, it becomes difficult to differentiate between the PS and the super-resolved image, except for the details that can be hardly recovered from 10m images. Among the different ways for carrying progressive resizing, no significant differences are found. However, carryin it out seems to be slightly beneficial when compared to model 4.

Finally, we would like to mention that we have carried out a validation of the reflectance values obtained by the super-resolution by comparing the histograms of a complete S2 image with that obtained after its super-resolution with our model. This comparison is presented in Figure 7, showing that histograms are not altered by the super-resolution, which is a very desirable property for further analysis.

5. CONCLUSIONS AND FUTURE WORK

In this work we have proposed a new way for super-resolving S2 RGB and NIR bands to 2.5m resolution. To do so, we have used images coming from another satellite with similar spectral characteristics but providing images at higher spatial resolution. Having image pairs located at the same place and almost at the same time with the same acquisition conditions, we have been able to train a deep learning model based on EDSR network. To make the network learn to super-resolve but not change the S2 images, we have first properly preprocessed image patches radiometry coming from the reference satellite, mainly applying histogram matching. For learning the EDSR, we have made use of the current guidelines for avoiding checkerboard patterns and learning efficiently. The loss function is also specifically designed for the task of super-resolution and has been adapted to work with NIR band. In the experiments, different forms of progressive resizing have been tested, showing their benefits. Both numeric (in terms of PSNR and SSIM) and visual results have shown the advantage of the proposed method over bicubic interpolation or other simpler methods only using S2 images.

Nevertheless, there is still work that remain to be done. The dataset used could be improved including more images for training, validation and testing. Moreover, the location of these images should cover different parts of the world to make the network more robust. With respect to the CNN, we would like to compare EDSR model with other state-of-the-art approaches such as GANs. Likewise, the experimental study should be completed with other methods not based on neural networks. Regarding our specific proposals, more research should be done to improve the quality of NIR bands and better balance the cost function to properly work with every band.

ACKNOWLEDGEMENTS

This work was partially supported by the Public University of Navarre under project PJUPNA13 and Tracasa Instrumental S.L. (OTRI 2018 901 073 and OTRI 2019 901 091).

REFERENCES

Aitken, A., Ledig, C., Theis, L., Caballero, J., Wang, Z., Shi, W., 2017. Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize. *arXiv*.

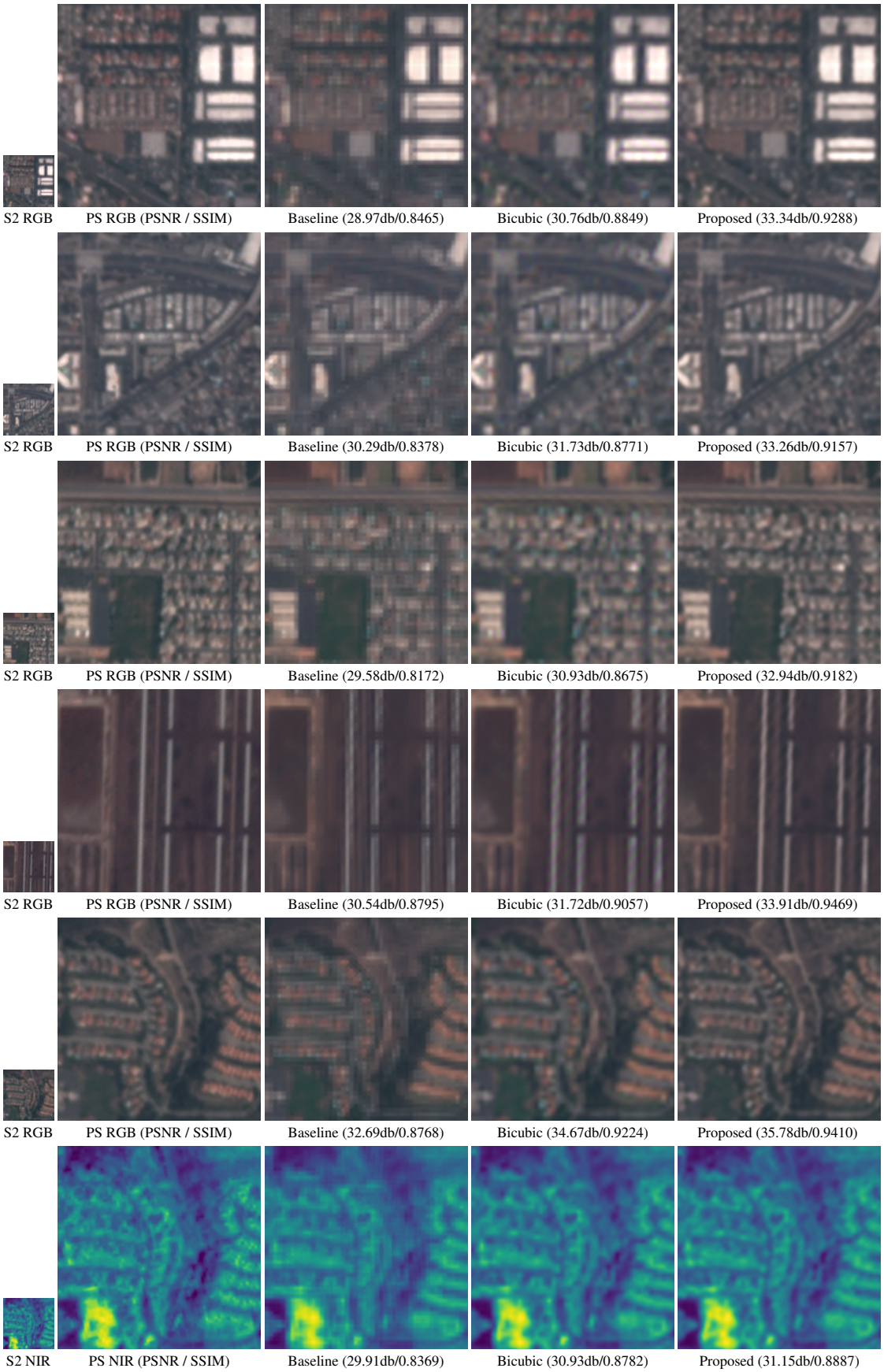

Figure 6. Visual comparison between the bicubic interpolation Baseline and our proposal (images correspond to model 2.3).

Beaulieu, M., Foucher, S., Haberman, D., Stewart, C., 2018. Deep image-to-image transfer applied to resolution enhancement of sentinel-2 images. *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2018-July, 2611–2614.

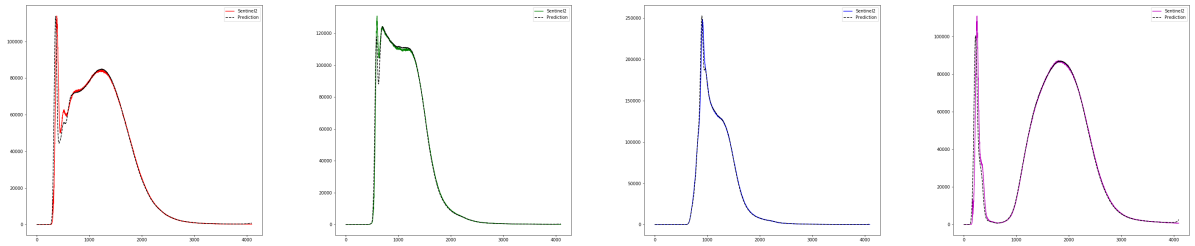

Figure 7. Comparison between input and predicted (scaled by 16) histograms with the complete S2 product associated with La Habra area (test set).

Dong, C., Loy, C. C., He, K., Tang, X., 2014. Learning a deep convolutional network for image super-resolution. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., Bargellini, P., 2012. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment*.

Galar, M., Sesma, R., Ayala, C., Aranda, C., 2019. Super-resolution for Sentinel-2 images. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W16, 95–102.

Gargiulo, M., Mazza, A., Gaetano, R., Ruello, G., Scarpa, G., 2018. A CNN-Based Fusion Method for Super-Resolution of Sentinel-2 Data. *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 4713–4716.

Gonzalez, R. C., Woods, R. E., 2008. *Digital Image Processing (3rd ed.)*. Prentice Hall.

Goodfellow, I., Bengio, Y., Courville, A., 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Keys, R. G., 1981. Cubic Convolution Interpolation for Digital Image Processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*.

Kim, J., Lee, J. K., Lee, K. M., 2016. Accurate image super-resolution using very deep convolutional networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Lanaras, C., Bioucas-Dias, J., Galliani, S., Baltsavias, E., Schindler, K., 2018. Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*.

Lecun, Y., Bottou, L., Bengio, Y., Ha, P., 1998. LeNet. *Proceedings of the IEEE*.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W., 2017. Photo-realistic single image super-resolution using a generative adversarial network. *Procs. 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*.

Liebel, L., Körner, M., 2016. Single-image super resolution for multispectral remote sensing data using convolutional neural networks. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 41(July), 883–890.

Lim, B., Son, S., Kim, H., Nah, S., Lee, K. M., 2017. Enhanced Deep Residual Networks for Single Image Super-Resolution. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2017-July, 1132–1140.

Pinto, C. T., Shrestha, M., Hasan, N., Leigh, L., Helder, D., 2018. SBAF for cross-calibration of Landsat-8 OLI and Sentinel-2 MSI over North African PICS. *Earth Observing Systems XXIII*, 10764, SPIE, 294 – 304.

Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., Wang, Z., 2016. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. *Procs. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv 1409.1556*.

Smith, L., 2018. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay. *arXiv*.

Sugawara, Y., Shiota, S., Kiya, H., 2018. Super-resolution using convolutional neural networks without any checkerboard artifacts. *2018 25th IEEE International Conference on Image Processing (ICIP)*, 66–70.

Wagner, L., Liebel, L., Körner, M., 2019. Deep residual learning for single-image super-resolution of multi-spectral satellite imagery. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W7, 189–196.

Yan, Q., Xu, Y., Yang, X., Nguyen, T. Q., 2015. Single image superresolution based on gradient profile sharpness. *IEEE Transactions on Image Processing*.

Yang, W., Zhang, X., Tian, Y., Wang, W., Xue, J.-H., 2018. Deep Learning for Single Image Super-Resolution: A Brief Review. *arXiv*, 1–17.

Zhou Wang, Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600-612.