DEEP RESIDUAL LEARNING FOR SINGLE-IMAGE SUPER-RESOLUTION OF MULTI-SPECTRAL SATELLITE IMAGERY

Lena Wagner*, Lukas Liebel*, Marco Körner

Computer Vision Research Group, Remote Sensing Technology, Technical University of Munich (TUM)
Arcisstr. 21, 80333 Munich, Germany -(lena.wager, lukas.liebel, marco.koerner)@tum.de

KEY WORDS: Single-Image Super-Resolution, Convolutional Neural Networks, Deep Learning, Residual Learning, Remote Sensing, Sentinel-2

ABSTRACT:

Analyzing optical remote sensing imagery depends heavily on their spatial resolution. At the same time, this data is adversely affected by fixed sensor parameters and environmental influences. Methods for increasing the quality of such data and concomitantly optimizing its information content are, thus, in high demand. In particular, single-image super-resolution (SISR) approaches aim to achieve this goal solely by observing the individual images.
We propose to adapt a generic deep residual neural network architecture for SISR to deal with the special properties of remote sensing satellite imagery, especially taking into account the different spatial resolutions of individual Sentinel-2 bands, *i.e.*, ground sampling distances of 20 m and 10 m. As a result, this method is able to increase the perceived resolution of the 20 m channels and mesh all spectral bands. Experimental evaluation and ablation studies on large datasets have shown superior performance compared to the state-of-the-art and that the model is not bound by its capacity.

1. INTRODUCTION

Single-image super-resolution (SISR) is a promising technique for various fields and applications. It allows enhancing the spatial resolution of a single image without having access to additional information, such as its acquisition properties or other images from the same sequence. By taking into account learned features from a training phase, the information content of the source image is increased. Many applications which suffer from a limited image size can benefit greatly from these methods. Possible causes for low resolution (LR) imagery can be restrictions in space if the sensor is too far away or the object to be observed is too small. Furthermore, high-quality sensors may be too costly for certain purposes or may not have been available at all when dealing with historic images. Hence, SISR has benefits for fields from medical or security imaging to remote sensing, which is the application we selected for our experiments. In remote sensing, images are usually acquired from large distances which is why achieving a high spatial resolution is often not feasible. However, high image quality is crucial for many specific applications where fine details are required, such as land cover classification, target detection, or the determination of object dimensions.

Very basic approaches to SISR are interpolation methods. An LR image is transformed to a high resolution (HR) grid and the intermediate pixels are estimated using a specific function, *e.g.*, bilinear interpolation. Aliasing effects are reduced but no high-frequency components are predicted. Super-resolution based on machine learning provides a better solution, as the relationship between LR and HR images is explicitly learned. Figure 1 exemplarily shows the improvement of image quality using our approach compared to bilinear interpolation. A promising deep learning approach for SISR which uses a convolutional neural network (CNN) is the so-called very

Figure 1. RGB composites with a spatial resolution of 10 m, up-sampled by a factor of two using bilinear interpolation (left) and our deep residual learning SISR approach (right).

deep super-resolution (VDSR) network (Kim et al., 2016). It has shown remarkable performance compared to other state-of-the-art methods. Its depth of 20 layers proofed to be very effective as it considers the neighboring contextual information utilizing a large receptive field. Furthermore, by making use of residual learning and a high learning rate, they achieve fast convergence.

Originally, the VDSR network was trained and tested on conventional 8 bit/px RGB images. In this paper, we present our approach to adapt the VDSR network to multi-spectral satellite imagery. We have to consider the different radiometric and spectral resolution, the varying topography of the Earth's surface that leads to texture changes, the acquisition circumstances like the distance of the sensor to the ground, and weather conditions like cloud coverage or snow. For our experiments, we used satellite imagery from the Sentinel-2 mission with a spectral resolution of 13 channels (Figure 2). Four of them have a spatial resolution of 10 m, which we use as ground truth data. As LR input data for training,

*Corresponding author

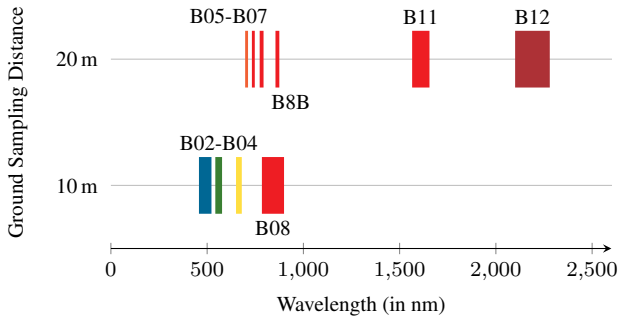


Figure 2. Spectral and spatial resolution of the Sentinel-2 bands with a GSD of 10 m and 20 m which we used in our experiments.

we down-sample them to 20 m. After training our network with these input ground-truth pairs, it is able to enhance images from 20 m to 10 m spatial resolution. We show that it also achieves good results for images from regions that were not used for training, as well as for images from different spectral bands. A quantitative comparison to a state-of-the-art method, the multi-spectral satellite image super-resolution (msiSRCNN) (Liebel and Körner, 2016), demonstrates the superior performance of our network.

The remainder of this paper is structured as follows: In Section 2, we explain how the problem of SISR is solved in other approaches. Section 3 explains the VDSR network, its architecture and features, and to what extent we build on it. Furthermore, problems are specified that occur due to the difference between satellite and conventional images. We describe our conducted experiments in Section 4, including data processing, training, and quantification metrics. Our results are presented and discussed in Section 5, as well as compared to another approach. In Section 6, we conduct an ablation study, where the impact of the depth of the network as well as the amount of input data is analyzed. The conclusion (Section 7) summarizes our contribution and specifies possible future work.

2. RELATED WORK

With the emergence of machine learning, more sophisticated solutions for SISR became possible. The review of Hayat (2018) describes the recent developments for super-resolution via deep learning and differentiates between two types of algorithms, *i.e.*, reconstruction and learning methods.

CNNs are designed to deal with image data. A SISR approach using a CNN has been proposed (Dong et al., 2014, 2016) which was a cornerstone for a lot of the following research. It was later outperformed by a network called VDSR (Kim et al., 2016) that was able to obtain considerable results with a very deep architecture.

In remote sensing, SISR has a high significance as it tackles problems that occur due to the exceptional sensor requirements for satellite imagery. These are caused, *e.g.*, by the large acquisition distance which prohibits a small ground sampling distance (GSD), *i.e.*, distance between centers of the ground area covered by neighboring image pixels. Fernandez-Beltran et al. (2017) provide an overview of SISR methods for the challenges of remote sensing. Their taxonomy differentiates reconstruction-based, learning-based and hybrid methods. An example of learning-based methods is utilizing CNNs, as in state-of-the-art SISR approaches for conventional photographs. The msiSRCNN (Liebel and Körner, 2016) adapts such a CNN (Dong et al., 2014, 2016) for Sentinel-2 images. The authors show the necessity of developing special networks for remote sensing applications that have to be trained with satellite imagery. Another approach in the field of remote sensing was proposed by Mei et al. (2017). They developed the so-called 3D-FCNN that exploits not only the spatial context of the neighboring pixels but also the spectral correlation of neighboring bands. Some work has been dedicated to the development of unsupervised methods as well. Haut et al. (2018) propose such an unsupervised deep generative network for remote sensing.

Closely related to our approach is the work of Huang et al. (2017), who build on the VDSR architecture and adapt it to multi-spectral satellite imagery from Sentinel-2. Based on first experiments, they conclude that this architecture is not capable of scaling remote sensing imagery, as it was not able to outperform the bicubic interpolation baseline. While they proceeded with using a custom network architecture, we analyze potential sources of errors and strategies to overcome them. The outcome is presented in the following sections.

3. A SUPER-RESOLUTION CNN FOR MULTI-SPECTRAL SATELLITE IMAGERY

As introduced before, SISR is a class of methods to improve the spatial resolution of an image without using any additional information. We propose to apply a deep learning approach, in particular a deep CNN for which we chose the VDSR architecture, presented by Kim et al. (2016).

This section describes the architecture, its characteristics (Section 3.2), and how to adapt it to multi-spectral satellite imagery (Section 3.1). As there are significant differences to conventional images regarding the acquisition geometry and sensor requirements, this poses a major challenge.

3.1 Problem

The VDSR network is designed to enhance the spatial resolution of conventional images of arbitrary scenes taken by consumer-grade handheld cameras. They usually only cover the visible spectrum from around 400–700 nm with three spectral bands (RGB). Most standard camera sensors and raster graphics formats feature a radiometric resolution of 8 bit/px and, thus, cover intensity values in the range of 0–255.

In comparison to handheld digital cameras, sensors of Earth observation satellites have to cope with a challenging acquisition scenario. Since images are expected to cover a huge area of the Earth's surface, they have to be taken from a large distance. The Sentinel-2 satellites orbit the Earth at an altitude of 786 km while acquiring images with a swath width of 290 km. Thus, the GSD is high, even though HR scanning sensors are utilized. For Sentinel-2, the GSDs are 10 m, 20 m, and 60 m, depending on the spectral band (Figure 2). Hence, the resulting satellite images do not cover structures with as much detail as conventional images.

Besides this difference in spatial resolution, multi-spectral satellite imagery differs from conventional images in its spectral and radiometric resolution. As the name implies, multi-spectral

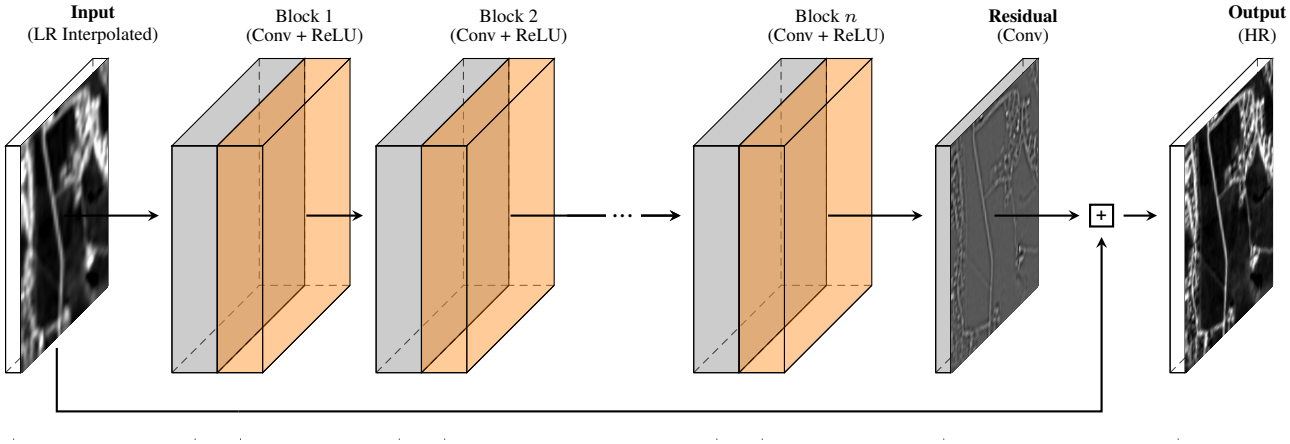


Figure 3. Architecture of the VDSR network, consisting of alternating convolutional layers and ReLU activations in n blocks. The output of the final convolutional layer is a residual image, which is added to the interpolated LR input to obtain the desired HR output.

imagery consists of several spectral channels that are acquired simultaneously. The 13 spectral channels of Sentinel-2 have a sampling depth of 12 bit/px, which is further specified in Section 4.1.

The goal of our approach is to enhance the quality of Sentinel-2 images. However, to train our network we need vast training datasets, consisting of samples with high spatial resolution as ground-truth and lower resolution as input images. Such image pairs are not directly available since Sentinel-2 acquires only one image with a fixed resolution per band. A workaround is to down-sample the original HR images and use these simulated LR images as input data (Dong et al., 2014, 2016; Liebel and Körner, 2016). Hence, we use the images of the four bands with a spatial resolution of 10 m as ground truth and down-sample them to a spatial resolution of 20 m to get the corresponding input data. After optimizing our network with these input/ground-truth pairs, it can be used to up-sample images of spectral bands with 20 m resolution to enhance the image quality.

3.2 Network

In this section, we describe the VDSR network (Kim et al., 2016), used in our SISR approach. The first part of its name originated from the relatively large number of 20 layers, as shown in Figure 3.

Prior to feeding images through the network, they are interpolated to the desired output size. As the network is expected to enhance the quality of this interpolated image, we use basic bilinear interpolation here. To speed up training, several images are concatenated to a batch. This batch of input images is then processed through the network that consists of 20 alternating convolutional layers and non-linear ReLU activation functions. Each convolutional layer is composed of 64 filters with a kernel size of 3×3. A specific feature of the VDSR approach is to learn a residual image that is added to the LR input to get the final HR result. Therefore, the sizes of input and output images are required to match. In order to preserve the image size, zero padding is applied before each convolution. A major advantage of residual learning is that the network only has to predict the high-frequency components of the image, while low-frequency components are directly transferred from input to output. According to Kim et al. (2016), faster convergence and superior performance can be achieved by exploiting the high correlation of input and output. Figure 4(b) shows the residual image along with the bicubic interpolated LR input and HR output. The high-frequency components are clearly visible.

In order to optimize the network, the deviation between the ground-truth and HR output image has to be minimized. The mean squared error

$$\mathrm{MSE}(I_A, I_B) = \frac{1}{N} \sum_{n=1}^{N} (I_{A,n} - I_{B,n})^2 \tag{1}$$

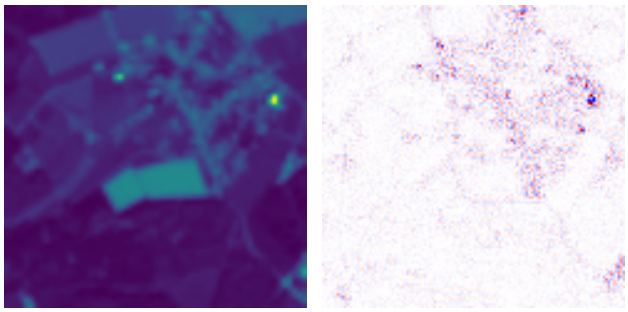
with $N =$ number of pixels per image
$I_{A,n}, I_{B,n} =$ corresponding pixels in images I_A and I_B

between both images, I_A and I_B serves as an easy tool to evaluate similarity metric and loss function for our approach. The loss is minimized by back-propagation using stochastic gradient descent.

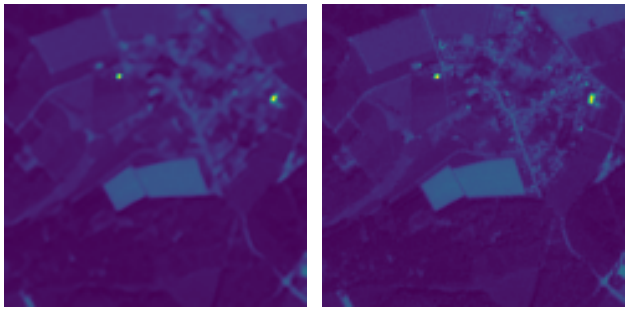
The authors of the VDSR suggested further measures for improving the performance during training and inference that we adopt. A relatively high initial learning rate of $\alpha = 0.1$ is used to efficiently train the deep network quickly. In order to ensure convergence, a multi-step schedule is applied that decreases the α after a certain number of iterations. Prominent problems that occur when training with high α include vanishing or exploding gradients. In order to avoid this, gradient clipping is applied. If the absolute value of gradients gets larger than a certain threshold value θ, they are clipped to the range $[-\theta, \theta]$. This technique helps to achieve convergence faster and more reliably.

4. EXPERIMENTS

In this section, we present the experiments that have been conducted. Section 4.1 describes the data acquisition and its pre-processing. In Section 4.2, we specify the training procedure, while Section 4.3 explains the used validation metrics.



(a) Bilinearly interpolated LR input (b) Residual



(c) HR output: input + residual (d) HR ground-truth

Figure 4. Qualitative comparison of results. The bilinear interpolation (a) is used as input for our network that predicts a residual image (b). By adding (b) to (a), the HR output (c) is generated. In comparison to the interpolation result, our method produces much clearer images that are more similar to the ground-truth (d).

4.1 Dataset

As Sentinel-2 is part of the Copernicus program, its images are freely available and can easily be downloaded from the Copernicus Open Access Hub[1]. We use Level 1C processing products that are already geometrically and radiometrically corrected as well as georeferenced. The multi-spectral instrument (MSI) aboard both Sentinel-2A and Sentinel-2B features 13 spectral bands, whereof four—the blue (B2, 490 nm), green (B3, 560 nm), red (B4, 665 nm) and near infra-red (B8, 842 nm) channels—have a spatial resolution of 10 m. These are used as ground truth data. Every image covers an area of 100 km^2, with a size of $10\,980 \times 10\,980$ px. The radiometric resolution of the MSI is 12 bit/px, but the images are encoded in 16 bit/px JPEG2000 format (Sentinel-2 User Handbook, 2013).

In total, ten images were downloaded. All of them are showing areas from central Europe, mainly Germany, as to be seen in Figure 5. Topography variations from the Alps to the Baltic and Northern Sea are covered. The images were taken from April 2018 to October 2018 to consider temporal variations in vegetation. Furthermore, we only selected images with cloud coverage of less than 1%.

We extracted each of the four 10 m bands to get a total of 40 single-band images and proceeded to cut them into small tiles of 60×60 px as the ground-truth patches. In order to obtain corresponding input patches, the tiles were down-sampled to 30×30 px, simulating a GSD of 20 m. The resulting dataset

[1] https://scihub.copernicus.eu

Figure 5. Footprints of the Sentinel-2 images used in our experiments. The distribution considers the varying topography of Germany from North to South, as well as temporal variations of the vegetation throughout the year. Cloud coverage in all used images is less than 1%. Background map: © OpenStreetMap contributors.

containing approximately $13.5 \cdot 10^6$ samples was partitioned 9:1 into training and test data. The test samples were extracted from the very right border of each of the 40 images in the dataset, thus representing the whole study area from North to South and all utilized spectral bands.

Since the image size is preserved during a forward pass through our network by design, the input patches need to be up-sampled to the desired output size prior to inference. As the network is expected to learn how to enhance the quality of these input patches, basic bilinear interpolation is expected to suffice here. Using such pairs of input and ground-truth samples the network can be optimized. No further data augmentation or pre-processing measures, such as normalization or color space conversion, have been implemented. Note that our dataset contains image patches of four different bands and mini-batches were randomly drawn from the whole dataset during training. We, thus, expect the network to learn weights that focus on spatial rather than spectral features, and hence also be able to process data from bands that have not been seen during training, in particular, the 20 m bands (*cf.* Figure 2: B05-B07, B8B, B11, B12).

4.2 Training Procedure

We re-implemented the VDSR in Pytorch, a scientific deep learning framework. This allowed us to train our model on GPUs fast and efficiently.

Like the original VDSR approach, we used 20 blocks, each consisting of a convolutional layer with 64 filters and a kernel size of 3×3 px, and non-linear ReLU activation. The final layer is a single convolutional layer that predicts the residual image. The learning rate was initially set to $\alpha = 0.1$ and

decreased by a factor of 10 every 2000 iterations. This was done five times until $\alpha = 10^{-7}$ was reached that was maintained for the remaining training time. Input parameters for the stochastic gradient descent optimizer are the momentum and a weight decay settings for which we chose 0.9 and 0.0001. For training, mini-batches with a size of 128 were used, which occupied 8 GB of GPU memory. We used $\theta = 0.4$ for gradient clipping.

Training the network to full convergence on an NVIDIA GeForce RTX 2070 took approximately four days. Only one epoch of training was necessary to obtain the best results, presented in the following.

4.3 Quantification Metrics

For validation, we use two error metrics to quantify our results. The peak signal to noise ratio (PSNR)

$$\text{PSNR} = 10 \cdot \log_{10} \frac{R}{\text{MSE}} \tag{2}$$

is usually used to compare the compression quality of images. It depends on the gray value range of the images R, which is the difference between the minimum and maximum of possible values, and the MSE. For Level 1C Sentinel-2 imagery we set $R = 10^4$. As the PSNR depends on the radiometric resolution of the image, it has no specific range, but the higher, the better accordance between both images. To still allow for a fair comparison with other approaches that use the full range of 16 bit/px here, we also calculate the PSNR with $R = 2^{16}$. Since the PSNR is directly derived from the MSE (*cf.* Equations (1) and (2)), which we used as our loss function, the PSNR was expected to converge with the MSE objective during training.

To get an additional quantitative evaluation of our results that is independent of the loss function, we calculated the structural similarity index (SSIM)

$$\text{SSIM}(I_A, I_B) = \frac{(2\mu_{I_A}\mu_{I_B} + C_1)(\sigma_{I_A I_B} + C_2)}{(\mu_{I_A}^2 \mu_{I_B}^2 + C_1)(\sigma_{I_A}^2 + \sigma_{I_B}^2 + C_2)} \tag{3}$$

with
μ_{I_A}, μ_{I_B} = mean gray values
$\sigma_{I_A I_B}$ = covariance of I_A and I_B
$\sigma_{I_A}, \sigma_{I_B}$ = variance
C_1, C_2 = constants

which extracts structural information and represents the similarity of images based on the human visual system (Wang et al., 2004). The range of SSIM ranges from zero to one, where SSIM $= 0$ corresponds to no similarity and SSIM $= 1$ indicates that original and reference images are identical. During training, we calculated the PSNR and SSIM for our test set every 500 iterations.

5. DISCUSSION

Figure 4 shows qualitative results for an image patch that was used for validation. Compared to the bilinearly interpolated LR input Figure 4(a), the HR output Figure 4(c), that was obtained by adding the residual Figure 4(b) to the input, was clearly improved. Its structures are more distinct and it looks very similar to the ground-truth image Figure 4(d).

The quantitative results confirm this improvement (Figures 6 and 7). As the curves illustrate, the MSE-based loss converges

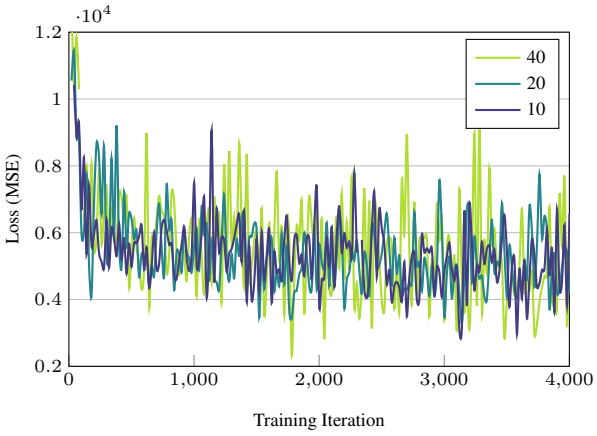


Figure 6. Convergence of the training loss for networks with different depths, *i.e.*, number of convolution + ReLU blocks. All three models converge to a very similar solution.

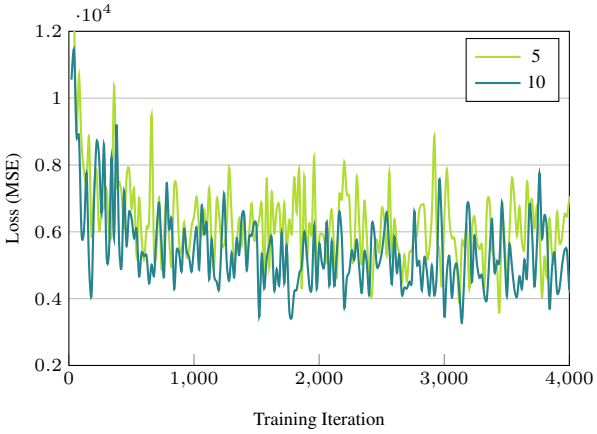


Figure 7. Convergence of the training loss for datasets of different sizes, *i.e.*, number of used Sentinel-2 granules. Training with a bigger dataset yields better results.

after a few thousand iterations, indicating that the differences in pixel intensity values between HR output and ground-truth decrease. In both plots, the turquoise curves show the results of a model utilizing a depth of 20 blocks and 10 Sentinel-2 images for training.

Table 1 shows the obtained values for PSNR and SSIM, compared to bilinear and bicubic interpolation. Our approach outperforms both interpolations with a margin of 2.6212 dB for the bicubic interpolation and even 3.8312 dB for the bilinear one clearly in terms of PSNR. This shows that the prediction result of our trained VDSR model is more similar to the ground-truth than the interpolation results. However, the SSIM is slightly higher for the interpolations than for the deep learning method. Since the SSIM is already very close to one for all methods, this raises the question of how suitable SSIM is for the validation of SISR approaches. Technically, the metric imitates the human visual system, but comparing the interpolation and prediction results visually (Figure 4), we actually find the output of our network to be more satisfying and significantly better than the interpolated version, which was also confirmed by the PSNR.

The PSNR obtained using the msiSRCNN by Liebel and Körner

Table 1. Comparison of quantitative results for up-scaling using different methods on an unseen test dataset (dataset mean). Note that the msiSRCNN was scored on a different test set.

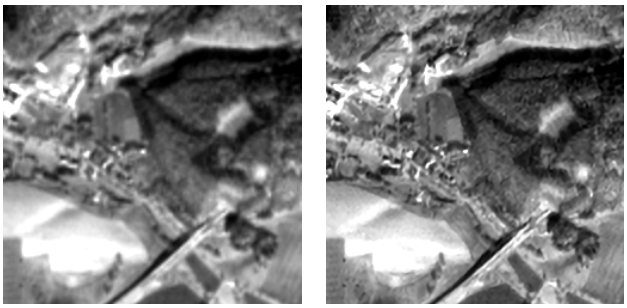
Method	PSNR (in dB)	SSIM
msiSRCNN (Liebel and Körner, 2016)	*60.6527*	*0.9979*
Bilinear Interpolation	61.3960	0.9993
Bicubic Interpolation	62.6060	0.9995
VDSR (ours)	65.2272	0.9989

(2016) is 60.6527 dB which is lower than our value. However, it is important to stress that these values were obtained using different test data which makes a direct comparison less meaningful. Nevertheless, they report their method to outperform bicubic interpolation by only 0.3680 dB on their test set. This indicates the superior performance of our network.

Huang et al. (2017) also tested the VDSR network on Sentinel-2 images. They concluded that the network fails at this task because in their experiments the bicubic interpolation result and network output did not differ significantly in terms of PSNR. A possible reason could be the very limited size of their training dataset, which contains only half as many training images than ours. Furthermore, they trained three models for the spectral bands B02, B03 and B04 separately from each other and thus had even less training data available per network. They also used $\theta = 0.01$ for gradient clipping, which is much smaller than $\theta = 0.4$ that achieved the best results in our experiments. Considering the PSNR of our results, we claim that the VDSR network does not fail in scaling multi-spectral satellite imagery. Yet, apparently, the method depends on a sufficiently large dataset for training. We did an in-depth analysis of this factor and report results in Section 6.

For final testing, an unseen Sentinel-2 image was used. The image was acquired over central Germany and, hence, does not overlap with the training set (*cf.* Figure 5). The goal of this final evaluation was to assess whether our trained model could also enhance the spatial resolution of the 20 m bands that were not used during training. We extracted the corresponding channels B5, B6, B7, B8a, B11 and B12 from our test image, used them as the LR input to our network, and calculated the HR output with an up-sampled GSD of 10 m. Since there is no ground-truth available for quantitative validation, results are shown in Figure 8 for qualitative comparison. Compared to the blurry bicubic interpolation, our super-resolution result looks much sharper and shows more distinct structures. Even though our network was not trained on samples from these channels, the learned weights apparently generalize well to unseen spectral bands. However, up-sampling bands that lie within the range of the electromagnetic spectrum that was used for training, *i.e.*, bands B05 to B07 and B8B (*cf.* Figure 2), yields the best results. For single-band images of B11 and B12 that feature longer wavelengths, the enhancement is less distinct as for the other bands. A visually appealing improvement was still achieved for all bands.

Additionally, we analyzed whether our network is able to further improve images that already have a GSD of 10 m to an up-sampled spatial resolution of 5 m, even though it was not trained for this task. The results, shown in Figure 9, look surprisingly good. Again, no quantitative evaluation is possible due to a lack of ground-truth with higher-resolution. This still shows that the network is able to generalize well by learning to distinguish regions with sharp edges and smooth areas with a constant texture and improve them accordingly.



(a) Bilinear interpolation (b) SISR Result

Figure 9. Enhancement of band 4 (red) with a native GSD of 10 m (left) to 5 m (right).

A comparison of the quantitative results to other state-of-the-art methods that did not make use Sentinel-2 imagery is not possible, first of all, due to the dependence of the PSNR on the radiometric resolution. Furthermore, the utilization of images captured by other sensors, on different scales, and showing different scenes will greatly influence the results. However, judging from qualitative comparison to the results presented by Mei et al. (2017); Haut et al. (2018), our method produces images of similarly high quality.

6. ABLATION STUDIES

In additional ablation studies, we analyzed two aspects of our method with a particular impact on the final performance. We modified the depth of the network by adding or taking away blocks in the network architecture (*cf.* Figure 3), as well as the size of our training dataset.

The original VDSR network has a sizeable depth of 20 blocks, as the authors claim that deeper models show better performance (Kim et al., 2016). By adding layers to a CNN, the size of the receptive field, which is the region of pixels that are taken into account for the calculation of a single pixel in the output, increases. Hence, with more layers, more contextual information is considered for the prediction, which is expected to yield better results. To check if this assumption holds for multi-spectral satellite images with a significantly different ratio of GSD to object size, we re-trained our network with 10 and 40 blocks. Contrary to our expectations, deeper models did not perform better here. The results, thus, did not correspond to the observations of Kim et al. (2016). All three models converged to similar values after a few iterations, as seen in Figure 6. Similarly, the evaluation of the final models yielded comparable values for both PSNR and SSIM, listed in Table 2. While the 20-block model achieved slightly better results in terms of PSNR than the other models, the SSIM was slightly lower, leading to the conclusion that the impact of network depth is negligible here. Since optimizing very deep networks is more expensive in terms of computation time and memory demands, using deeper models is not preferable in this case. We conclude that given the high GSD of multi-spectral satellite imagery and, thus, the small size of objects in the image considering contextual information at a large scale is

Table 2. Quantitative results of our application studies with varying network depth and dataset size.

Hyperparameter		Performance	
Network Depth	Dataset Size	PSNR (in dB)	SSIM
10	10	65.1493	0.9997
20	5	65.0119	0.9965
20	10	65.2272	0.9989
40	10	65.1096	0.9997

neither necessary nor beneficial for our method. We continued our experiments using the best performing network architecture featuring 20 blocks of network layers.

Apart from the network architecture, we expected the size of the utilized training dataset to have the biggest impact on the results. In order to show this, we reduced the size of our training set by removing all but the five westernmost images from the original dataset (*cf.* Figure 5). This employed procedure of selecting a subset of images maintains the North/South distribution of samples over the study area, which exhibits the biggest difference in topography. Re-training with the compiled subset of images yielded significantly different training results. Figure 7 shows a comparison of the training process when utilizing the full set of images *vs.* a 50% smaller subset. The latter quickly converged to an inferior solution, meaning that the model will always perform worse than the baseline—even given a large number of training iterations. The evaluation results (contained in Table 2) confirm this observation, although the difference is surprisingly small. A smaller set of images naturally contains less variation and, hence, does not fully exploit the capacity of a deep CNN. Our experiment showed that the amount of training data limits the performance of a network, which is an intuitive and well-known observation. Considering that no manual annotations are required for training using the suggested procedure, sometimes referred to as self-supervised training, this outcome is still of particular importance, as it impressively shows the importance of carefully assessing the required amount of training data for this application and approach.

7. CONCLUSION

We presented a learning-based method for improving the spatial resolution of multi-spectral Sentinel-2 satellite imagery. The deep CNN architecture VDSR (Kim et al., 2016) that we utilized was shown to be well-suited for this task. Its design is comparably simple as it consists of alternating convolutional layers and ReLU activations, organized in blocks. In our experiments, we were able to show that the depth of the network is not a limiting factor for the performance of the network. The size of the dataset, on the contrary, proved to be crucial.

Additional features of the network are residual learning, a very high learning rate that was decreased in steps, and gradient clipping, which led to fast and stable convergence. The final model is able to produce impressive super-resolution results. A comparison to the existing method msiSRCNN (Liebel and Körner, 2016) showed the superiority of our approach in terms of both employed metrics, *i.e.*, PSNR and SSIM. Furthermore, our model significantly outperformed the more basic bilinear and bicubic interpolation methods in qualitative and quantitative comparison. In this sense, our results did not go in line with the observations of Huang et al. (2017) who stated that the used network architecture is not suitable for scaling Sentinel-2 imagery, as it was not able to outperform the interpolation baseline in their experiments.

As our study area in Central Europe only covers a comparably small area of the globe, it should be noted that, even though we considered a varying topography ranging from mountains to the sea, results may differ when applying pre-trained models to regions with a different appearance. Thus, for application on a global scale, training should be conducted on likewise datasets. This would enable enhancing arbitrary Sentinel-2 images from which many applications could benefit. One example where our SISR method could be used as a plug-in pre-processing step is land use and land cover classification. Increasing the spatial resolution reveals more detailed structures and surface textures that can, thus, be distinguished more easily and with higher spatial accuracy.

Our objective was up-sample the 20 m bands of Sentinel-2 to a GSD of 10 m which can be accomplished even for channels that feature a considerably larger wavelength than the channels used for training. In further evaluation experiments, we showed that our network is able to provide satisfying super-resolution results even when being applied to higher scales, such as scaling the 10 m bands to 5 m resolution. This demonstrates that our model is scale-independent, up to a certain degree.

We successfully adapted and re-trained the VDSR CNN architecture for enhancing the resolution of multi-spectral Sentinel-2 imagery. The proposed method outperformed existing methods in the conducted experiments, both quantitatively and qualitatively. In particular, the visual quality of up-sampled Sentinel-2 images can be significantly improved as compared to interpolation methods.

References

Dong, C., Loy, C. C., He, K., Tang, X., 2014. Learning a deep convolutional network for image super-resolution. *European Conference Computer Vision*,184–199.

Dong, C., Loy, C.C., He, K., Tang, X., 2016. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38, 295-307.

Fernandez-Beltran, R., Latorre-Carmona, P., Pla, F., 2017. Single-frame super-resolution in remotesensing: A practical overview. *International Journal of Remote Sensing*, 38, 314-354.

Haut, J.M., Fernandez-Beltran, R., Paoletti, M.E., Plaza, J., Plaza, A., Pla, F., 2018. A new deep generative network for unsupervised remote sensing single image super resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 56, 6792–6810.

Hayat, K., 2018. Super resolution via deep learning. *Digital Signal Processing*, 81, 198–217.

Huang, N., Yang, Y., Liu, J., Gu, X., Cai, H., 2017. Single image super resolution for remote sensing data using deep residual learning neural network. *Neural Information Processing*, Lecture Notes in Computer Science, 10635, 622 - 630.

Kim, J., Lee, J.K., Lee, K.M., 2016. Accurate image super resolution using very deep convolutional networks. *The IEEE Conference on Computer Vision and Pattern Recognition*,1646–1654.

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*, XLI-B3, 883–890.

Mei, S., Yuan, X., Ji, J., Zhang, Y., Wan, S., Du, Q., 2017. Hyperspectral image spatial super-resolutionvia 3D full convolutional neural network. *Remote Sensing*, 9,1139.

Sentinel-2 User Handbook, 2013. https://earth.esa.int/documents/247904/685211/Sentinel-2_User_Handbook

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*,13,600–612.

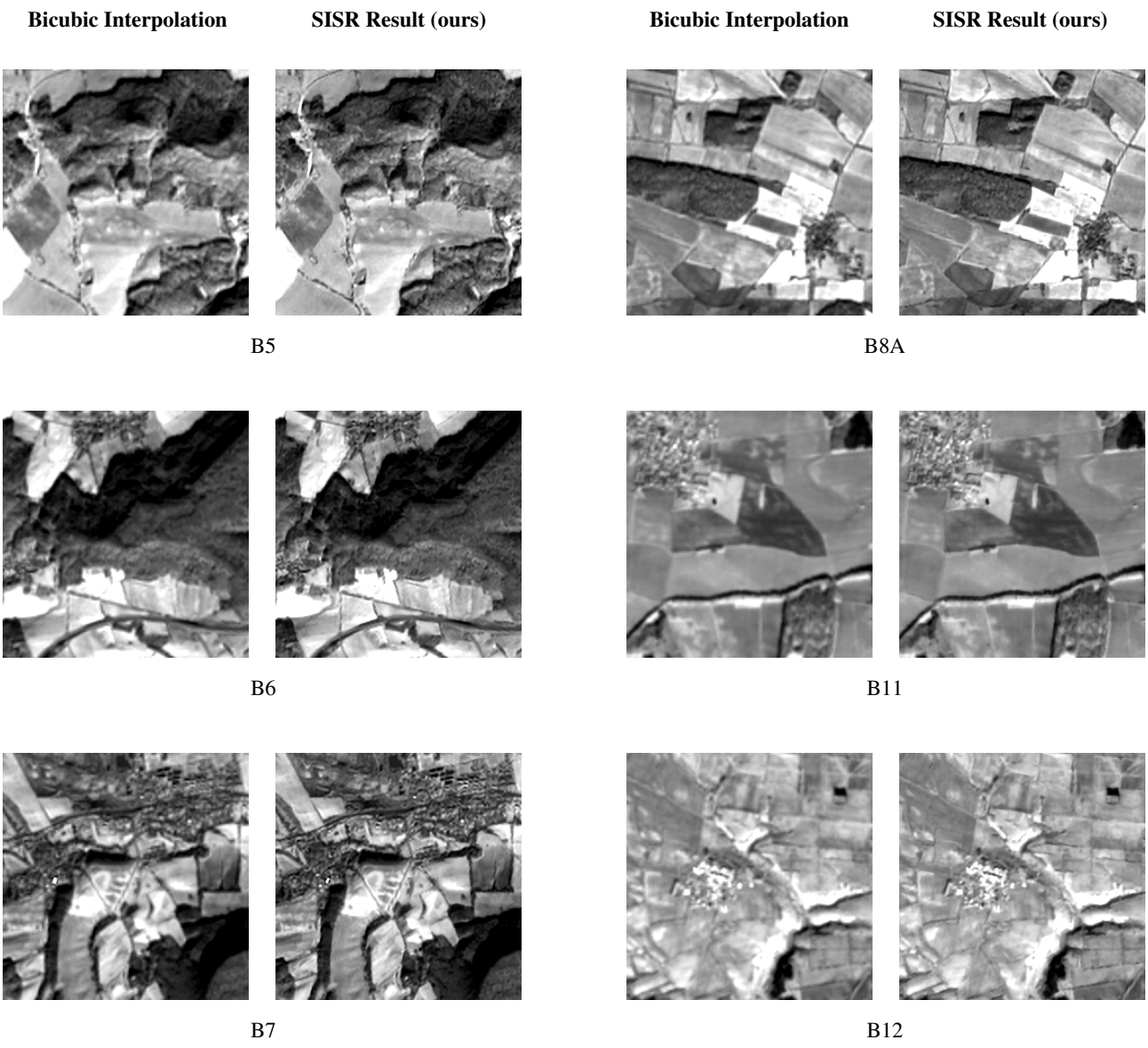


Figure 8. Super-resolution results with a GSD of 10 m for the six Sentinel-2 bands with a native GSD of 20 m.