The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-1/W3-2023
2nd GEOBENCH Workshop on Evaluation and BENCHmarking of Sensors, Systems and GEOspatial Data
in Photogrammetry and Remote Sensing, 23–24 October 2023, Krakow, Poland

CHALLENGES IN PREPARING DATASETS FOR SUPER-RESOLUTION ON THE EXAMPLE OF SENTINEL-2 AND PLANET SCOPE IMAGES

A. Malczewska[1,*], J. Malczewski [2], B. Hejmanowska[1]

[1] AGH University of Krakow, Faculty of Geo-Data Science, Geodesy, and Environmental Engineering,
Department of Photogrammetry Remote Sensing of Environment and Spatial Engineering - (zadlo,galia)@agh.edu.pl,
[2] The Henryk Niewodniczański Institute of Nuclear Physics Polish Academy of Sciences,
The LHCb Experiment Department - jakub.malczewski@ifj.edu.pl
[*] zadlo@agh.edu.pl

KEY WORDS: super-resolution, satellite images, dataset preparation, earth observation, Sentinel-2, PlanetScope

ABSTRACT:

Benchmark datasets is an significant aspect in in many areas such as computer vision, deep learning, geospatial data as they serve as standardized test sets for evaluating the performance of models. Among many techniques of image processing, there is super-resolution (SR) which is aimed at reconstructing a low-resolution (LR) image into a high-resolution (HR) image. For training and validation SR models as a dataset the pairs of HR and LR images are needed, which should be the same apart from resolution. There is a lot of benchmark datasets for super-resolution methods, but they usually include conventional photographs of an common objects, while remote sensing data have different characteristic in general. This paper focuses on the process of preparing datasets for super-resolution in satellite images, where high-resolution and low-resolution image data come from different sources. The case of the single-image super-resolution method was considered. The experiment was performed on Sentinel-2 and PlanetScope data, but the assumptions can also be transferred to data obtained from other satellites. The procedure on how to make the pairs of HR and LR images consistent in terms of time, location and spectral values was proposed. The impact of the processes carried out was measured using image similarity measurement methods such as PSNR, SSIM and SCC.

1. INTRODUCTION

The recently developed super-resolution (SR) techniques based on machine learning are aimed at reconstructing a low-resolution (LR) image into a high-resolution (HR) image. These methods are also used in satellite images, especially on public, free of charge data to enhance their quality. In the acquisition of satellite images, there is always a trade-off between the spatial, spectral and temporal resolution (Yue et al., 2016). The super-resolution techniques improve the quality of spatial components. There are two main methods of super-resolution: single image (SISR) and multi-image (MISR). SISR methods assume that for training the model pairs of HR and LR are needed, while in MISR the portion of data contains one HR and many LR images. This paper focuses on the process of preparing datasets for super-resolution of satellite images, where HR and LR data were obtained from different satellites. The case of SISR is considered.

To prepare the dataset for SISR, pairs of images are needed, and one of them should be higher resolution. The best case is when exactly the same LR and HR data are available (e.g. made with the same sensor from a different distance or with a different resolution). If this kind of data is not available, training is conducted on approximate data that does not reflect exactly what needs to be achieved. The resulting images should be similar in all aspects to the input LR image except for the resolution. In the case of training models on approximate data, the result is not free from errors caused by differences between the input data and the reference data. However, for satellite images, and real-world cases, very often there is no other option than training on approximated data. As a reference, data from another sensor are used with frequently exhibit varying temporal, spatial, and spectral characteristics. Therefore, it is important to prepare such images taking into account the reduction of differences in the LR and HR pairs.

Benchmark datasets fulfil an important function in deep learning. They are carefully curated to cover a wide range of challenges and scenarios, allowing to compare new approaches against existing state-of-the-art methods. There are a lot of datasets for super-resolution methods which include conventional photographs of common objects, landscapes or people. An example of such a dataset is DIV2K (Agustsson and Timofte, 2017) which consists of 1000 high-resolution with corresponding low-resolution images. Low-resolution images were degraded from high-resolution in two ways: bicubic interpolation and more realistically by using low-pass filters, Poisson noise, pixel shifting, and motion blur. Other examples of frequently used data for image denoising and super-resolution testing are: Set5 (Bevilacqua et al., 2012), Set14 (Zeyde et al., 2012), and Urban100 (Huang et al., 2015).

Using datasets prepared on standard images in the case of satellite imagery does not result in a positive and stable effect. This is primarily due to the characteristics of the satellite data. Very often they have more spectral channels, different radiometric resolutions, and show scenes that differ from standard images. A lot of datasets dedicated to satellite images have been created and the summary was presented by (Bakuła et al., 2019). The main purposes for which they were created are instance segmentation, semantic segmentation, scene classification, or

* Corresponding author

object detection. The datasets are based mostly on open access images such as Sentinel-1, Sentinel-2, Landsat, MODIS, and Google Earth, but also on commercial satellite data like PlanetScope, WorldView, DigitalGlobe and moreover on aerial and drone images. Some of the datasets contain georeferences and some do not. Access to the benchmark datasets with descriptions and reference to articles can be found in official repositories such as ISPRS Benchmarks (ISPRS Team, 2021), IEEE Geoscience and Remote Sensing Society (IADF TC and IEEE GRSS, n.d.), or TorchGeo databased (TorchGeo Team, Microsoft AI for Good program and the PyTorch Team, 2021). Many datasets are also collected in unofficial repositories, e.g. on GitHub: (Cole, 2022) or (Ali Ahmadi, 2021). In the case of the super-resolution, several datasets were also created. They are concerned mostly with resolutions lower than 10m/pix and the multi-image super-resolution approach. Datasets for super-resolution and their use in various models are described in more detail in the section 2.

Preparing benchmark datasets for a super-resolution of images with higher resolution than 10m/pix is always limited because higher-resolution data is not available free of charge and cannot be published. Although it is not always possible to share such a set of data, it is practicable to implement a stable procedure for such cases. The goal of the data preparation procedure presented in this article is to obtain the most similar pairs of HR and LR images from the images collected by different satellites. In the presented example Sentinel-2 and PlanetScope multispectral images are used respectively for low-resolution and high-resolution datasets. Presented algorithms and image similarity assessment metrics could be applied to images collected from other satellites, with different resolutions, and different characteristics.

The rest of the work is organised as follows: Section 2 is about benchmark dataset and algorithm for super-resolution of satellite images, Section 3 describes experiment design, data and methods, Section 3.4 presents results of data preparation. A short discussion is in Section 4, and conclusions in Section 5.

2. SUPER-RESOLUTION FOR SATELLITE IMAGES

2.1 Benchmark datasets for SR for satellite images

To the best of our knowledge, from the remote sensing multispectral benchmark dataset, there are only a few for super-resolution. One of the newest and very promising datasets consists of 10m/pix and 20m/pix surface reflectance Sentinel-2, with their reference spatially-registered 5m/pix images acquired on the same day by the VENμS satellite (SEN2VENμS) (Michel et al., 2022). Data can be used for training 8 bands of Sentinel-2 images down to 5m/pix. This dataset covers 29 locations and consists of 132,955 patches of 256x256 pixels images at 5m/pix resolution. The authors of the SEN2VENμS dataset took spatial registration and radiometric adjustment into consideration when preparing the data and additionally filtered invalid patches. Other examples have been created for multi-image super-resolution. The first example is an official dataset of ESA's Kelvins competition for "PROBA-V Super-Resolution" (Märtens et al., 2019). The data shows 74 regions around the globe at different points in time and includes channels Red and Nir in resolution 300m/pix and 100m/pix. Another example which focuses on Sentinel-2 is The WorldStrat Dataset. This dataset contains 10,000 km2 of high-resolution images SPOT6/7 with the resolution of 1.5m/pix for panchromatic and 6m/pix for multispectral channels and 10m/pix Sentinel-2 images (Cornebise et al., 2022).

2.2 Models used for SISR

Although there is not much prepared data available for developing super-resolution methods on satellite images, a lot of articles describing models have been published. SRCNN (super-resolution convolutional neural network) was tested firstly on Sentinel-2 (Liebel and Körner, 2016) and presented also on Landsat images with reference training data from Sentinel-2 (Pouliot et al., 2018). Enhanced Deep Residual Network is widely used for satellite images, also for example in enhancing the resolution of Sentinel-2 based on PlanetScope (Galar et al., 2020). While preparing the dataset for that experiment, consideration was also given to the co-registration of images and spectral matching. SARNet (Spectral Attention Residual Network) in addition to the convolution layer contains Residual Channel Attention Blocks (RCAB) to improve the results and visual aspect of super-resolved images (**?**). This example was also prepared for Sentinel-2 based on PlanetScope. The approach for data preparation contained in that article is the closest to the one presented in Section 3. A lot of modifications to existing and new network models are published, not only for Sentinel-2 images. They have been collected in review articles (Wang et al., 2022a), or (Wang et al., 2022b).

3. EXPERIMENT

3.1 Experiment design

The process of data preparation presented in this work was performed on the example of Sentinel-2 data (LR) and PlanetScope data (HR). The Sentinel-2 dataset is one of the largest sources of public and free of charge data provided by the European Space Agency (ESA). Sentinel-2 images contain 13 spectral channels (R, G, B, near and far infrared) with a spatial resolution of 10 m, 20 m and 60 m depending on the channel. Thanks to easy and free access, they are widely used. PlanetScope images are also multispectral, with a resolution of 3m/pix and the spectral characteristics of channels are very similar to Sentinel-2. The data are commercial but are available under the "Education and Research Standard Plan" with limitations. In this paper channels Red, Green, Blue and Nir Infrared are used, Sentinel-2 10m/pix B2, B3, B4, B8 and equivalents for PlanetScope.

Based on the publications cited in section 2.2 and on the conducted experiments, a data processing procedure for performing deep learning super-resolution of Sentinel-2 10m/pix to 2.5m/pix based on PlanetScope images was proposed. As PlanetScope images are distributed in a resolution 3m/pix the bicubic interpolation was used for resampling them to 2.5m/pix. Presented general rules could also be applied to any kind of data whose characteristics are similar. The proposed procedure includes three main aspects of data preparation:

1. **Temporal compatibility.** It is very rare that photos from two sensors are taken on the same day. Even if this happens, the acquisition hours also matter for the exposure of the images. It is very important to pay attention to the smallest differences in the acquisition time of images during the data collection process.

2. **Spectral compatibility.** Spectral characteristics differ in each kind of source. Even images from both sources are

similar, some kind of spectral adjustment is required. In this experiment, the method of histogram matching has been applied.

3. **Geometric compatibility.** Data from the same area, but obtained from different satellites, differ slightly in georeferencing. In order to keep the geometry as accurate as possible, methods for sub-pixel co-registration are used. One of them is available in the public package AROSICS (Scheffler et al., 2017).

For robust data preparation, several other factors should be taken into account. One of them is atmospheric corrections, which should be the same for both sources. Another factor that is important and can spoil the data is the presence of clouds and cloud shadows. The solution may be to use a cloud mask and eliminate the patches on which they are located. The next one is gaps in the images, error pixels or places where the photo was not recorded. Finally, random errors could be found, such as planes, unidentified objects or local registration errors. These kinds of errors are the most difficult to eliminate automatically.

This article was based on images that show surface reflectance and both, Sentinel-2 and PlanetScope already have bottom-of-atmosphere correction. Pairs of images could contain small clouds, shadows and random errors. In this article we focuses primarily on measuring the impact of co-registration and histogram matching on the similarity of the HR and LR pairs. During the process, only PlanetScope images have been modified in order to match Sentinel-2 images. To evaluate the impact of modification image similarity assessment metrics such as PSNR (peak signal-to-noise ratio), SSIM (Structural Similarity Index) (Wang et al., 2004), and SCC (Spatial Correlation Coefficient) (Zhou et al., 1998) were used. To compare two images, they should have the same spatial resolution, so Sentinel-2 data were resampled by bicubic interpolation by factor 4 to 2.5m/pix. Three components were checked. Firstly, the temporal compatibility. Images of Sentinel-2 and PlanetScope covering the same area but obtained at different times were compared. Secondly, the impact of method for sub-pixel co-registration was tested. Finally, the spectral resolution was adjusted by histogram matching. It was also checked whether these methods are dependent on the image size: large and small patches were used. Metrics are calculated after each step of HR image processing to verify the impact of each one. It was assumed that if the HR image after the process is more similar to the LR image, the data is better prepared. The proposed data quality improvement procedure may have a positive impact on the model training process as well as on the final results of super-resolution.

3.2 Dataset

For this experiment, we prepared a dataset consisting of Sentinel-2 and compatible with PlanetScope images. Sentinel-2 was downloaded as Level-2A. Images from both sources were atmospherically corrected as a Surface Reflectances product in cartographic geometry. PlanetScope images were downloaded as harmonized to Sentinel-2. The dataset contains images from March to August for the years 2021 and 2022. Images were taken from 30 areas of 1792x1280 pixels for Sentinel-2 images, which translates to 229.38 sq km of data. The images show different types of land cover such as urban, sub-urban, forest, large and geometrical crops, small and chaotic crops, and waterbodies from different parts of the World.

Data preparation was as follows. Sentinel-2 was cropped to selected areas and RGBNir composites were created from them. To make a comparison possible Sentinel-2 images were resampled by bicubic interpolation to 2.5m/pix. PlanetScope were merged in cases when one downloaded order included several pieces, cropped to the selected areas and resampled by bicubic interpolation from 3m/pix to 2.5m/pix. We checked that there are no error ”0” pixels in PlanetScope. Then we divided scenes into small patches of 128x128 pixels for Sentinel-2 and 512x512 pixels for PlanetScope. Two versions of the same datasets were used in the experiment, the first consisted of 30 large scenes and the second of 4200 small patches. Two types of sizes were checked to check the importance of the image size on the processes performed. An example patch for both sources is presented in the figures 1.

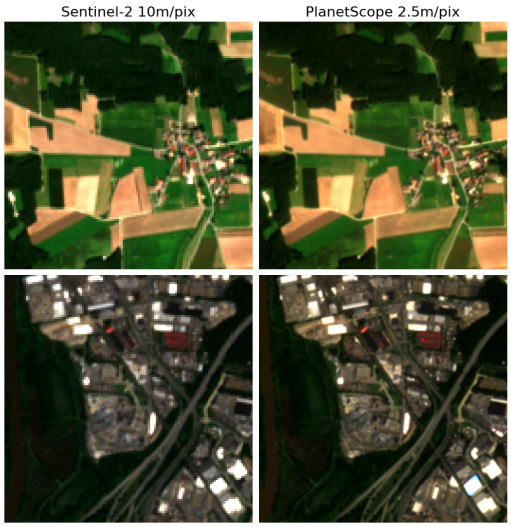

Figure 1. Sentinel-2 10m/pix (left) and PlanetScope 2.5m/pix(right) examples of patches.

3.3 Methods

3.3.1 Co-registration method To co-register images within each pair we used AROSICS (Automated and Robust Open-Source Image Co-Registration Software) (Scheffler et al., 2017). It is a Python-based software for automatic detection and correction of sub-pixel inaccuracy between two remote sensing images. The algorithm relies on the phase correlation technique and the Fourier shift theorem, which allows for accurately calculating the X/Y offsets at a specific geographic location. The authors of this method presented two approaches: local and global. The local co-registration approach correctly detected local geometric misregistration while the global computed one displacement vector within the image subset (matching window) and shifted the whole image based on it. In this article, the global co-registration approach was used. As reference and target images respectively Sentinel-2 10m/pix and PlanetScope 2.5m/pix were used. The matching window covered most of the image area (at least 80%).

3.3.2 Spectral adjustment Images acquired by different satellites will always differ in the range of values, which results from the specificity of the sensors: central wavelength and bandwidth. Planet offers a method of harmonizing the R, G, B and Nir channels for PlanetScope data to Sentinel-2 images, but this does not eliminate all differences between the images. Moreover, PlanetScope provided products could dif-

fer for each year in the range of minimum and maximum values of the images, which was observed in our data sample. In this article, we tested the histogram matching method, which is a frequently used approach for pre-processing steps in computer vision tasks. We match histograms of the PlanetScope 2.5m/pix images to the Sentinel-2 original 10m/pix. The process of histogram matching is divided into 4 steps. First of all, is the calculation of histograms of both images. Second, the calculation of Cumulative Distribution Function (CDF) for both histograms. CDF represents the integral of the probability distribution of pixel values. The third step is the calculation of a mapping function that transforms the pixel values in the target image's histogram to match the distribution of the reference image's histogram. This mapping is usually achieved by finding the corresponding pixel value in the reference image for each pixel value in the target image. Finally, pixel values of the target image are redistributed according to the reference image's histogram by the mapping function. Histogram matching was done both on large and small patches.

3.3.3 Image similarity assessment metrics For images similarity assessment methods we applied widely used metrics: peak signal-to-noise ratio (PSNR) (Wang et al., 2004), Structural similarity (SSIM) (Wang et al., 2004) and Spatial correlation coefficient (SCC) (Zhou et al., 1998). These objective methods are based on a similarity comparison between a super-resolved image and a ground true high-resolution image. Improved results are indicated by higher scores in all three metrics. Moreover, we evaluated the results of processing by visual interpretation. This method is not objective but allows to catch large errors.

PSNR is the ratio between the maximum possible power and the power of corrupting noise and is based on mean squared error.

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX^2}{MSE} \right) \tag{1}$$

where MAX = the maximum possible pixel value of the image, for this case is 2^{16}.
MSE = mean squared error

SSIM takes into account the influence of three components: structural information, illumination and contrast.

$$SSIM = \frac{(2\mu_{LR}\mu_{HR} + c_1)(2\sigma_{LRHR} + c_2)}{(\mu_{LR}^2 + \mu_{HR}^2 + c_1)(\sigma_{HR}^2 + \sigma_{HR}^2 + c_2)} \tag{2}$$

where μ_{HR} = the pixel sample mean of HR
μ_{LR} = the pixel sample mean of LR
σ_{HR}^2 = the variance of HR
σ_{LR}^2 = the variance of LR
σ_{LRHR} = the covariance of LR and HR
$c_1 = (k_1 L)^2$
$c_2 = (k_2 L)^2$
where L is the dynamic range of pixel value (in this case it is $2^{16} - 1$)
$k_1 = 0.01$ by default
$k_2 = 0.03$ by default

SCC is a correlation coefficient between a high-pass filtered SR image and a high-pass filtered HR original image. This metric is especially important for images or part of the images concentrated in the high-frequency domain. It is primarily used to evaluate the spatial component. The mask of the used high-pass filter is equal:

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix} \tag{3}$$

3.4 Results

3.4.1 Temporal comparison As a first step, we checked how the influence of the time discrepancy affects the similarity of the pairs of HR-LR images. For comparison, fragments of three areas (T18SUJ, T32UPV, T33UYR) and three dates were selected. The chosen scenes contain mostly agricultural land, sub-urban areas and vegetation. The first date was the closest one to the Sentinel-2 images, the second was close to that day of the month but a year earlier or later, and the third was from a different time of year. In this case, images were not pre-processed, only bicubic interpolation was done to maintain a consistent resolution. Figure 2 presents the example of one area with different dates. The image shows part of the crops which is the area where the biggest changes can be observed during the year. Sentinel-2 and PlanetScope images from the same day look very similar, while others show different periods of plant growth. PSNR, SSIM and SCC were calculated on three large scenes (7168x5120 pix) for each PlanetScope date to Sentinel-2 reference scene. The results presented in the tables 1, 2 and 3 indicate the best metrics scores for the closest date.

Figure 2. Images present one patch from the T32UPV scene for different days. The upper left is the Sentinel-2 image, then PlanetScope. Changes in the fields could be observed.

Sentinel-2 T18SUJ: **13.08.2022**			
PlanetScope dates	PSNR	SSIM	SCC
13.08.2022	35.8532	0.6388	0.2012
13.08.2021	34.6130	0.6012	0.1346
07.10.2022	33.6626	0.6374	0.1350

Table 1. PSNR, SSIM and SCC scores for the T18SUJ scene for three different dates.

3.4.2 Impact of co-registration and histogram matching To examine the impact of co-registration and histogram match-

Sentinel-2 T32UPV: **14.08.2021**			
PlanetScope dates	PSNR	SSIM	SCC
14.08.2021	49.2090	0.9923	0.1244
14.08.2022	41.3019	0.9327	0.0841
24.03.2022	40.9284	0.9436	0.0511

Table 2. PSNR, SSIM and SCC scores for the T32UPV scene for three different dates.

Sentinel-2 T33UYR: **05.08.2022**			
PlanetScope dates	PSNR	SSIM	SCC
04.08.2022	35.9045	0.8036	0.0936
14.08.2021	33.9733	0.7510	0.0421
05.09.2022	34.6701	0.7633	0.0570

Table 3. PSNR, SSIM and SCC scores for the T33UYR scene for three different dates.

ing we prepared two versions of samples. For both versions, we performed two processes directly on PlanetScope images, not one after the other. The intention was to check the impact of each process separately. PlanetScope were bicubicly resampled to 2.5m/pix images and Sentinel-2 10m/pix was a reference.

Variant 1 - co-registration and histogram matching were performed on 30 scenes, and then scenes were divided into small patches.

Variant 2 - at first, scenes were divided into small patches and then co-registration and histogram matching were performed on each small patch.

The division after the processing into small patches in version 1 made it possible to compare results between both versions. The example of the histogram matching for one patch is shown in Figure 3.

Distributions of the metric scores of PSNR and SCC for variant 1 are shown in Figure 4. Both processing parts change the shape of distributions. Co-registration and histogram matching are processes that change images for two separate components: spatial and spectral. Upon analyzing the distributions, it becomes evident that both procedures enhanced the metrics scores, but for the spectral component, PSNR is functional, while for the spatial component SCC. Compared to distribution based on original samples PSNR is much better after HM and SCC improved after co-registration. Although on the Planet website, the products were downloaded in the same mode, two peaks can be seen in the PSNR distribution. This is caused by the different spectral ranges of the data for 2021 and 2022. Histogram matching eliminates such differences in the data.

Figures 5 and 6 present a comparison between two variants which differ in the order of data division and processes performed. Figure 5 shows the changes in SCC scores for co-registration for variant 1, variant 2 and for data without processing. Figure 6 shows the changes in PSNR scores for histogram matching with the same variants configuration. For both cases, variant 2 (processing on small patches) gives slightly better results. To test that, we compared the means of the respected metrics for samples of two variants by calculating a dependent t-test (Student, 1908), (David and Gunnink, 1997). In both cases, the results are statistically significant with a confidence level of 99%.

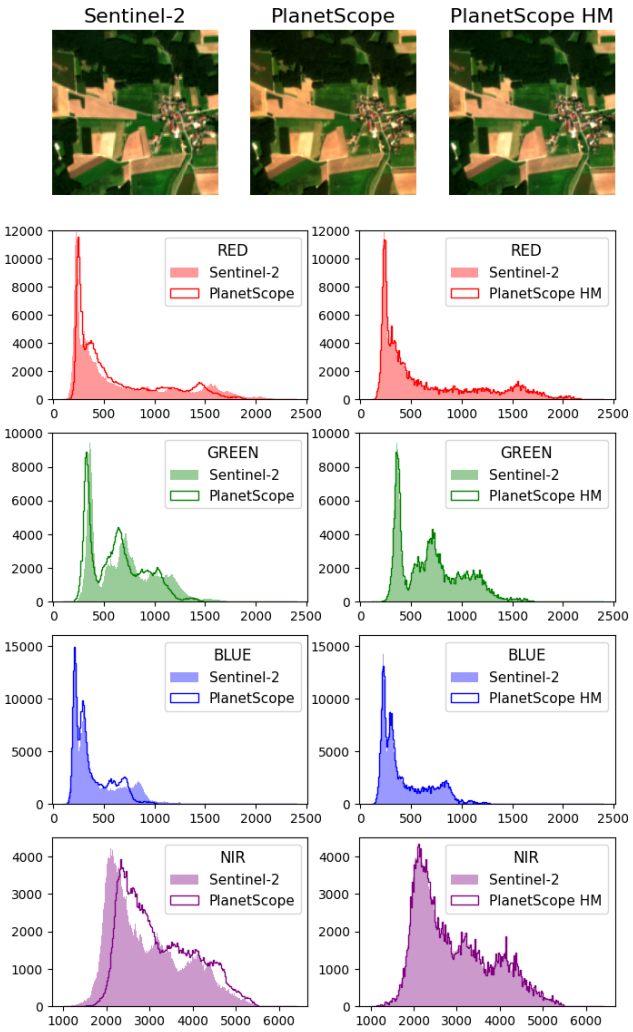

Figure 3. Example of histogram matching for one patch from T32UPV scene. Upper part: Sentinel-2, PlanetScope and PlanetScope after histogram matching. Bottom part: histograms for each channel R, G, B and NIR before (left) and after (right) histogram matching.

Finally, we performed processing sequentially for one option: division to the patches, co-registration and histogram matching. Figure 7 presents a pair plot of PSNR and SCC metric scores for results of each process separately for small patches (variant2) and a combination of both processes. The higher the metric value, the greater the similarity of the images. The results show that each method improves the image in a different aspect and the combination gives the best result.

A summary of the metric scores results for each step and variants is provided in Table 4. The table shows the mean and standard deviation of PSNR, SCC and SSIM for pairs of images. The order of results is as follows: original data (without processing); variant 1 co-registration, histogram matching and combined methods were done on big scenes; variant 2 co-registration, histogram matching and combined method were done on small patches; finally mixed variants in which co-registration was made on big scenes (variant 1), then are divided to small patches and then histogram matching was done (variant 2).

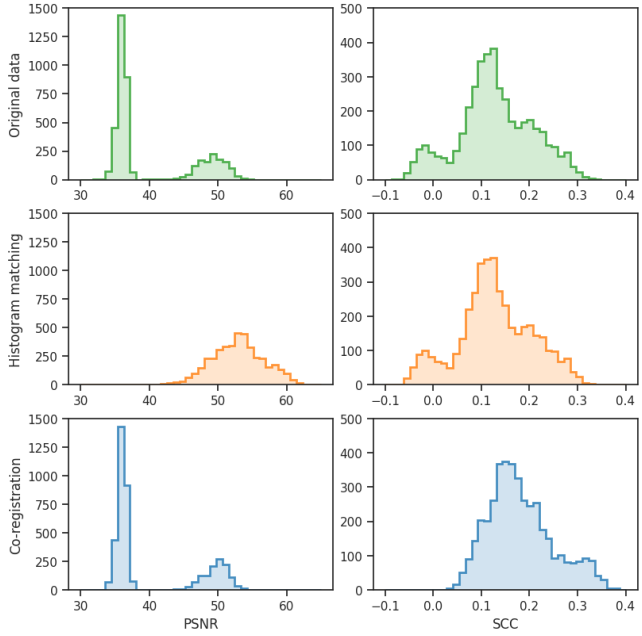

Figure 4. Distribution of metrics scores: PSNR (left) and SCC (right) for all pairs of image patches for variant 1. From the top: data without processing, after histogram matching and after co-registration.

	PSNR		SCC		SSIM	
processess	mean	std	mean	std	mean	std
original	39.88	6.13	0.128	0.078	0.804	0.134
	V1					
coReg	40.05	6.34	0.181	0.068	0.805	0.134
HM	52.81	3.82	0.127	0.078	0.996	0.004
coReg+HM	53.82	4.06	0.180	0.068	0.997	0.004
	V2					
coReg	40.06	6.36	0.186	0.071	0.805	0.134
HM	53.57	3.97	0.126	0.078	0.996	0.005
coReg+HM	54.90	4.20	0.183	0.071	0.997	0.004
	coReg V1 and HM V2					
coReg+HM	54.78	4.20	0.178	0.068	1.00	0.004

Table 4. Mean metrics scores and standard deviation for each process and variant combinations.

4. DISCUSSION

In preparing a large set of real data for super-resolution based on satellite images collected from two or more different sources many issues must be taken into consideration. Most of them are atmospheric correction, presence of clouds and shadows, temporal, geometric and spectral compatibility, and moreover incorrect registrations, and random errors. This article focused on the examination of the differences regarding acquisition time, co-registration and spectral characteristics.

The analysis of differences due to the date was performed for 3 areas. The metrics scores present in tables: 1, 2, 3 vary by scene which depends mostly on PlanetScope registration. For each case, the result of the closest date to the Sentinel-2 image is the highest. Data from the same day but from different years may also show high similarity due to the similar phase of plant growth. Vegetation areas and agricultural fields are the areas

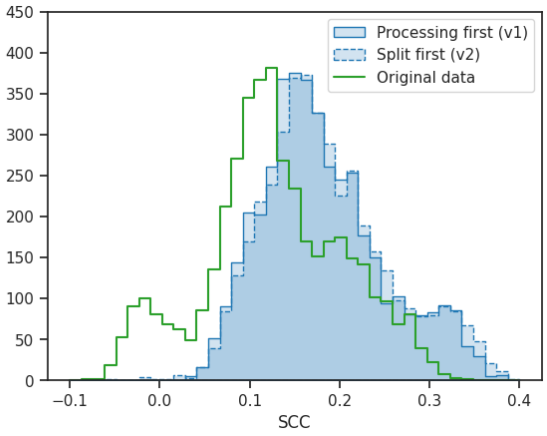

Figure 5. Distributions of SCC scores for co-registration method for original data, variant1 (processing on big patches), and variant2 (processing on small patches).

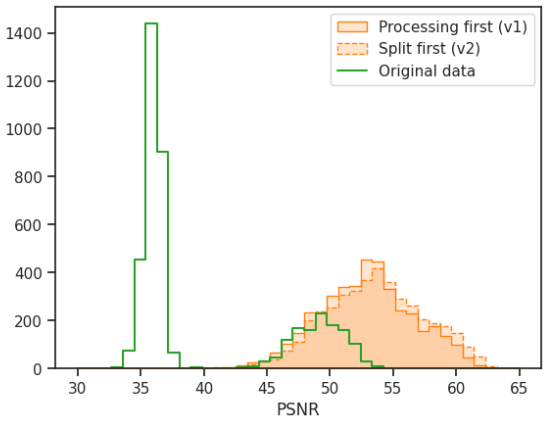

Figure 6. Distributions of PSNR scores for histogram matching for original data, variant1 (processing on big patches), and variant2 (processing on small patches).

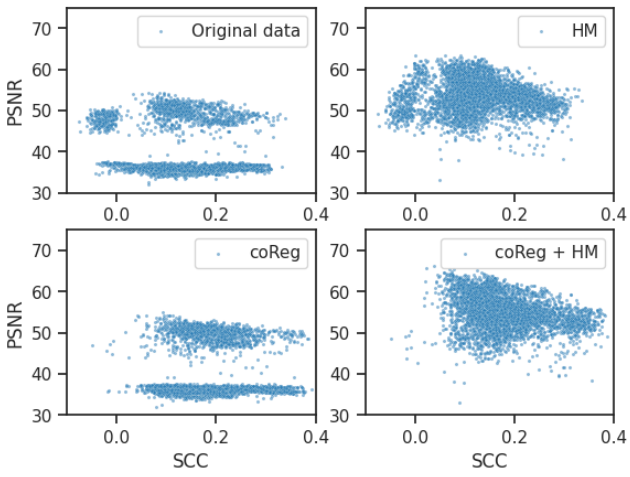

Figure 7. Pair plots for metric scores PSNR and SCC for original data, results of co-registration, results of histogram matching and both processing sequentially. Scores for variant 2 (processing on small patches).

where the greatest changes occur during the year, while built-up areas can be described as more constant.

PSNR and SCC metric scores calculated for PlanetScope images without processing, after co-registration and after histogram matching show that both of the processes improve similarities between within image pairs. Figure 4 indicate that the SCC metric is more sensitive to changes in the spatial component, while PSNR is for the spectral component. These two metrics can be used as complementary in the process of comparing the similarity of images.

Comparison of the results of processes performed on different sizes of images (Figure 5 and 6 leads to a conclusion that performing both processing on small patches gives slightly better results. This is also visible in Table 4 which presents means values for metrics for each combination. Despite the statistical significance of this result the moment of dividing the scenes is not a key aspect of obtaining better image similarity. In some cases it is more convenient to perform co-registration on large photos first and, histogram matching on smaller ones. A convincing argument may be the processing time, unnecessary creation of more data or co-registration on the original scenes (and not, as in this case, on PS scenes after interpolation using the bicubic method). The mixed method could be also useful, in that co-registration is performed before dividing scenes into patches and histogram matching afterwards. In Figure 7 we can observe an overall trend of improving image similarity by performing both processes as measured by SCC and PSNR metrics.

The result of the performed tests is the proposal of the following procedure for preparing data for SR on satellite data from various sources:

1. Downloading pairs of images corresponding to each other in time
2. Atmospheric corrections if needed
3. Divide into smaller patches
4. Co-registration
5. Histogram Matching
6. Calculation of metric scores for possible rejection of weak samples

There is still the issue of other inaccuracies in the data, such as clouds and cloud shadows that are difficult to detect, missing data or appearing objects like aeroplanes. Moreover, in the case when the missing data is a few pixels with the value ”0”, a method to consider may be filling these pixels with the median or average from the kernel, e.g. 5x5 adjacent pixels. Distorted data may cause poorer model training results, but it is important to discard as small pieces as possible so as not to lose good ones. Verification of small patches metric scores may allow to the rejection of erroneous samples. Regarding the results presented in Figure 7 for the combined method (coReg + HM), we can observe outliers, samples that achieved poor metrics scores. Probably these pairs are so inconsistent that it is difficult to correct them, so they can be discarded to obtain a cleaner dataset.

Obtaining and preparing images from various sources is a demanding task, and in training neural network models both the amount of data and their purity count. The example prepared in this paper shows a positive impact of data consistency and improvement of the spatial and spectral aspects on the similarity values of image pairs. This suggests that the conclusion should be the same for data with similar characteristics. To check how strongly the presented processes influence the SR model training results, it would be necessary to prepare data in each of these variants and retrain. It is beyond the scope of this article and could be assigned for future work.

5. CONCLUSION

The research presented in this paper contributes insights into the development of benchmark datasets tailored for super-resolution satellite imagery. We addressed the problem of approximate data for real-world cases. We focused on the preparation of datasets for single-image super-resolution in satellite images. Experiments were conducted using Sentinel-2 as low-resolution and PlanetScope as high-resolution data, illustrating a methodology that could be adapted for data from various satellites. The proposed procedures ensure consistency in time, location, and spectral values between HR and LR image pairs.

Performing co-registration by the AROSCIS algorithm and histogram matching improved significantly spatial and spectral components respectively. Image similarity assessment was done by calculating metrics such as PSNR, SSIM and SCC. The analyses conducted proved that PSNR is more sensitive to spectral changes, while SCC to spatial changes. Metrics are complementary and using both gives full insight into the impact of processes carried out to improve the quality of images pairs. It was also checked what effect the size of the photos on which co-registration and histogram matching were performed had. Smaller patches received slightly higher scores.

The procedure for preparing satellite data for deep learning super-resolution presented in this paper can also be used for data from other sensors. On a data set with different characteristics, the presented processes may have different ranges for improving the similarity of pairs of images but maintaining temporal, spatial and spectral coherence is essential for preparing images for super-resolution.

6. ACKNOWLEDGEMENT

The authors would like to acknowledge Planet Labs for providing the Planet images through their research and education program “Education and Research Standard Plan”.

REFERENCES

Agustsson, E., Timofte, R., 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1122–1131.

Ali Ahmadi, S., 2021. Awesome Satellite Benchmark Datasets. The GitHub repository for Satellite Benchmark Datasets. https://github.com/Seyed-Ali-Ahmadi/Awesome$_{S}atellite_{B}enchmark_{D}atasets$.

Bakuła, K., Mills, J. P., Remondino, F., 2019. A REVIEW OF BENCHMARKING IN PHOTOGRAMMETRY AND REMOTE SENSING. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-1/W2, 1–8. https://isprs-archives.copernicus.org/articles/XLII-1-W2/1/2019/.

Bevilacqua, M., Roumy, A., Guillemot, C., Morel, M. L. A., 2012. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. British Machine Vision Association, BMVA.

Cole, R., 2022. Datasets for deep learning applied to satellite and aerial imagery. The GitHub repository for Satellite Images Deep Learning . https://github.com/satellite-image-deep-learning/datasets.

Cornebise, J., Oršolić, I., Kalaitzis, F., 2022. The WorldStrat Dataset: Open High-Resolution Satellite Imagery With Paired Multi-Temporal Low-Resolution.

David, H. A., Gunnink, J. L., 1997. The Paired t Test Under Artificial Pairing. *The American Statistician*, 51(1), 9-12.

Galar, M., Sesma, R., Ayala, C., Albizua, L., Aranda, C., 2020. LEARNING SUPER-RESOLUTION FOR SENTINEL-2 IMAGES WITH REAL GROUND TRUTH DATA FROM A REFERENCE SATELLITE. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-1-2020, 9–16. https://isprs-annals.copernicus.org/articles/V-1-2020/9/2020/.

Huang, J.-B., Singh, A., Ahuja, N., 2015. Single image super-resolution from transformed self-exemplars. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5197–5206.

IADF TC and IEEE GRSS, n.d. IEEE Earth Observation Database. The Image Analysis and Data Fusion Technical Committee (IADF TC) and the IEEE Geoscience and Remote Sensing Society (IEEE GRSS). https://eod-grss-ieee.com/dataset-search.

ISPRS Team, 2021. ISPRS Benchmarks. ISPRS scientific initiatives project team. https://www.isprs.org/education/benchmarks.aspx.

Liebel, L., Körner, M., 2016. SINGLE-IMAGE SUPER RESOLUTION FOR MULTISPECTRAL REMOTE SENSING DATA USING CONVOLUTIONAL NEURAL NETWORKS. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B3, 883–890. https://isprs-archives.copernicus.org/articles/XLI-B3/883/2016/.

Märtens, M., Izzo, D., Krzic, A., Cox, D., 2019. Super-resolution of PROBA-V images using convolutional neural networks. *Astrodynamics*, 3(4), 387–402.

Michel, J., Vinasco-Salinas, J., Inglada, J., Hagolle, O., 2022. SEN2VENμS, a Dataset for the Training of Sentinel-2 Super-Resolution Algorithms. *Data 2022, Vol. 7, Page 96*, 7(7), 96.

Pouliot, D., Latifovic, R., Pasher, J., Duffe, J., 2018. Landsat Super-Resolution Enhancement Using Convolution Neural Networks and Sentinel-2 for Training. *Remote Sensing*, 10(3).

Scheffler, D., Hollstein, A., Diedrich, H., Segl, K., Hostert, P., 2017. AROSICS: An Automated and Robust Open-Source Image Co-Registration Software for Multi-Sensor Satellite Data. *Remote Sensing 2017, Vol. 9, Page 676*, 9(7), 676.

Student, 1908. The Probable Error of a Mean. *Biometrika*, 6(1), 1–25. http://www.jstor.org/stable/2331554.

TorchGeo Team, Microsoft AI for Good program and the PyTorch Team, 2021. torchgeo.datasets Geospatial and Non-geospatial Datasets. The PyTorch Foundation. https://torchgeo.readthedocs.io/en/latest/api/datasets.html.

Wang, P., Bayram, B., Sertel, E., 2022a. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Science Reviews*, 232, 104110.

Wang, X., Yi, J., Guo, J., Song, Y., Lyu, J., Xu, J., Yan, W., Zhao, J., Cai, Q., Min, H., 2022b. A Review of Image Super-Resolution Approaches Based on Deep Learning and Applications in Remote Sensing. *Remote Sensing 2022, Vol. 14, Page 5423*, 14(21), 5423.

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(4), 600–612.

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

Zeyde, R., Elad, M., Protter, M., 2012. On single image scale-up using sparse-representations. *Curves and Surfaces*, Springer Berlin Heidelberg, Berlin, Heidelberg, 711–730.

Zhou, J., Civco, D. L., Silander, J. A., 1998. A wavelet transform method to merge Landsat TM and SPOT panchromatic data. *International Journal of Remote Sensing*, 19(4), 743-757.