CONTRIBUTION OF SUPER RESOLUTION TO 3D RECONSTRUCTION FROM PAIRS OF SATELLITE IMAGES

N. Imperatore[1,*], L. Dumas[1]

[1] CS, 5 rue Brindejonc des Moulinaid, Toulouse Cedex 5, France

Commission II, WG II/2

KEY WORDS: Super resolution, Digital Surface Model, Stereo-Matching, Deep Neural Network, Generative Adversarial Network

ABSTRACT:

The photogrammetric 3D stereo reconstruction from pairs of strereo images is rising interest in the past few years in space field downstream. Nowadays, it is conceivable that a large production of DSMs from satellite images can become the primary source of 3D information on a global scale. However, in urban areas, DSMs produced with current technology suffer from poor quality. Indeed, even using very high resolution (VHR) images, there is too little information to generate disparity maps that reproduce very well defined shaped objects such as buildings.
To address this issue, one solution may be to artificially increase image resolution beyond the sensor limits. Super resolution (SR) algorithms are designed to recover high frequencies, introducing significant information in a scene characterized by strong and frequent discontinuities such as a city. State-of-the-art methods relying on Deep Learning have shown remarkable results in this sense. The aim of this work is therefore to assess the contribution of single image SR Deep Learning techniques to the stereo matching and DSMs generation in an urban context, highlighting potential advantages and limitations that can show up when introducing such a technology in a multi-view stereo pipeline. The proposed contributions are: a methodology for super resolution of VHR data that takes into account realistic simulation of a satellite product; a testbed for the evaluation of the impact of super resolution on 3D photogrammetric reconstruction; a local analysis of the consequences of deep learning SR of VHR images on stereo matching.

1. INTRODUCTION

1.1 Context

The 3D photogrammetric reconstruction from pairs of strereo images is a growing application in space field downstream. Thanks to the performance of Very High Resolution (VHR, ground sampling distance less than 1 m) satellites of last generation and a better revisit time, it is possible to render smaller scale objects, such as building and trees so that it is appropriate to talk about Digital Surface Models (DSMs) computed from pairs of satellite images. Data in DSM format are significant in the remote sensing context, and their value is foreseen to raise in the following years, due growing applications in space field downstream, from 3D city mapping to urban fluid mechanics studies, from landcover to glacier studies (Lebègue et al., 2020). With respect to other technologies of 3D geographical representation, e.g. point clouds, DSMs are more convenient because simpler to manipulate with current techniques. Moreover, once a satellite is in orbit, DSMs obtained by pairs of satellite image are more straightforward to be produced than lidar ones, which require ad hoc campaigns and are thus characterized by a lower temporal frequency.

In such a framework, the *Centre national d'étude spatiales* (CNES) is developing the *Constellation Optique 3D* (CO3D) mission (Lebègue et al., 2020). This optical constellation, whose launch is foreseen in 2023, will provide simultaneous VHR pairs. In this way, temporal differences will be minimized, allowing improvements in point cloud generation. CS Group will develop the image processing components of the CO3D ground segment. The DSM generation pipeline is key and at this purpose, CS developed for CNES two tools: CARS (Michel et al., 2020), a multi view stereo pipeline that from a stereo pair generates the corresponding DSM; Pandora (Defonte et al., 2021), which is in charge of the stereo matching step from rectified images.

Nevertheless, when it comes to reconstruct objects at a finer scale, many challenges have to be tackled when using stereo-reconstruction. In particular, DSMs in urban areas may suffer from poor quality. This can be seen as bottleneck for many critical applications. At this subject, the DSM generation pipeline is key and stereo matching is the crucial step of the chain. Most algorithms work by computing, along epipolar lines, a cost function that tells the similarity between the neighborhoods of two points in the left and the right image. Such a function is hereby referred as *cost profile* or *similarity measure* and tells us where it's more likely to find the right disparity.

In order to enhance stereo matching step one solution may be to increase resolution beyond the sensor limits. Besides standard interpolation techniques, more sophisticated methods to upsample an image have been proposed and they're usually referred to as Super-resolution (SR) methods. The general concept of Super-resolution refers to those algorithms designed for increasing an image resolution as if employing a sensor with a higher nominal resolution. In spatial domain it might be seen as the problem of finding the less aliased and blurred interpolation of an image, while in the Fourier space it consists of recovering high frequencies from the low ones. SR is a notoriously ill-posed inverse problem: infinite solutions exist and, typically, prior knowledge is used to guide the optimization towards the best achievable solution. DNNs are then suitable for such a task as they allow automatic extraction of meaningful highly abstract knowledge, removing the need for identifying

case-specific features (Haut et al., 2018).

Given these premises, it is of interest for the development of CO3D DSM pipeline but also for the photogrammetry community to assess whether super-resolved images via innovative deep learning techniques can be beneficial for stereo matching.

1.2 Related work

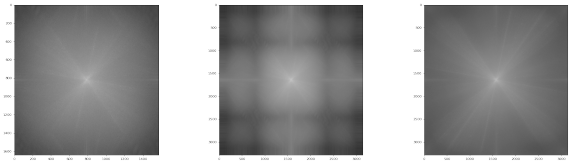
Only few works explored this subject beforehand, suggesting that there might be an interest in using single image super resolution (SISR) for DSM generation, but also highlighting the challenges to overcome in order to rigorously prove it. First, super resolution via deep neural network is a relatively new technology and their usage might involve some challenges. Indeed, in order to train a SR network we would need a set of images in input/low resolution (LR) and the associated target/high resolution (HR) samples, taken with the very same instrument on the very same scene. Since such a dataset is extremely challenging to be obtained, especially for space applications, we usually choose a HR dataset and apply degradation and downsample operations to obtain the LR one. To do so, it is usually assumed a sensor model that is applied in the HR-LR transformation. In SR works we often see that little or no importance is given to the simulation of real satellite images, both in terms of source images and HR-LR degradation. Typically, the methodology relies on widely used datasets (UC-Merced, RSCNN7, AID, etc.), borrowed by image classification studies, whose origin and processing is not always mastered. A simplified LR dataset generation technique (usually bicubic downsampling) is adopted. This is partially justified by the fact in most works the focus is on the model itself and a common benchmark easy to reproduce is needed in order to compare the performance of different models. However, this undermines the credibility of such models when it comes to reliably super resolve real VHR data.

The last few years saw a flourishing of SISR works, in computer vision as well as in remote sensing. Detailing the entire state-of-the-art for SR is beyond the scopes of this paper; nonetheless we consider recent developments for our super resolution method. Among all possible architectures, residual and Generative Adversarial Networks (GAN) are claimed to be the most powerful methods for super resolution (Anwar et al., 2020) (Tsagkatakis et al., 2019). In a GAN two networks are trained: a generator that upsamples the input image, and a discriminator whose task is to recognize which image is real between the ground truth and the generated sample. The networks try to fool each other and the result of this game should be an increase in SR image perceptual value.

An experience aiming at evaluating SISR in the context of stereo matching is proposed in (Zhang et al., 2019), but in their work the model is trained on a dataset created for object detection (DOTA) and assuming a bicubic degradation for the generation of the LR training set. Additionally, the reference used to compute the statistics is also a product of a DSM pipeline and this might introduce a bias in the quantitative results. Moreover, the neural networks used in that work (namely SRCNN (Dong et al., 2014) and VDSR (Kim et al., 2016)) are outdated with respect to recent developments, and many other networks have shown superior performance. Other related works can be found in literature. (Burdziakowski, 2020b) utilizes a dataset composed by UAV images, artificially blurred and then deblurred using a DNN, showing how blurred images lead to worse DSMs. (Pashaei et al., 2020) proves that a denser point cloud can be generated when using super resolved images but it is not clear whether this is due to the upsampling itself or to the information added by the deep neural networks. In another work, (Burdziakowski, 2020a) also addresses point cloud density but the DSM pipeline leads to contradictory results, as it returns worse quality DSMs when having a denser point cloud, even in the case of the reference HR.

1.3 Hypothesis and objectives

It has been illustrated that, already by using standard interpolation techniques, we're able to better characterize the cost function optima thanks to the smaller sampling distance (Szeliski and Scharstein, 2004). On the other hand, this doesn't introduce any spectral structure that might be useful for the matching algorithm to better estimate the disparity. That's why the use of super resolution (SR) methods seems to be justified at this purpose. Fig. 1 shows how the Fourier transform of a neural network super resolved image seems to propagate the spectrum of the image, unlike bicubic upsampling apart from rebound artifacts does not create high frequencies. The hypothesis to verify is that this spectral information can benefit the stereo matching step, increasing the confidence in the estimation of the right disparity from the similarity measures (Fig. 2). It can be shown that the reliability of a disparity measure can propagate into a stereo pipeline leading to more accuracy in the final product (Sarrazin et al., 2021). The aim of this work is therefore to assess whether plugging a SR pre-processing step into a multi-view stereo pipeline can benefit DSM production in urban regions.



LR Bicubic SR

Figure 1. Spectra of an input image, its bicubic interpolation and its super resolved version using deep learning techniques

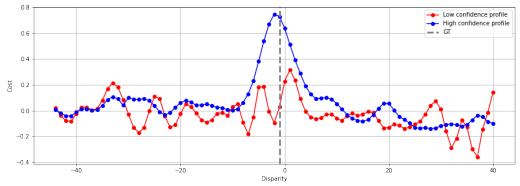


Figure 2. High and low confidence level cost profiles

To do so, we set up an experiment for comparing multiple super resolved image DSMs to a reference. The results of such an experience are presented in section 3. In order to increase the relevancy of this work, we address the issues that have been highlighted in the related works in paragraph 1.2. First, we propose a training set generation procedure that takes into account realistic satellite degradation, targeting a particular sensor model (Pléiades). Training methodology and results are presented in section 2. Second, we'll rely on state-of-the-art neural networks for SR. Together with the Enhanced SR GAN (ESRGAN) (Wang et al., 2018), the Residual Dense Network (RDN) (Zhang et al., 2018) was implemented, since the latter has an architecture really similar to the ESRGAN generator. In this

way, we'll try to isolate the contribution of a discriminator. Furthermore, all the experiments will include a bicubic upsampling counterexample in order to discern the real influence of artificial intelligence the effects that we observe by a mere increase in image sampling, and thus justifying the employment of complex models such as deep neural networks. Finally, we use a lidar reference when measuring DSMs errors to obtain more reliable estimation.

On the top of that, none of the works presented in section 1.2, isolates the contribution of SR to the stereo matching itself. In fact, DSM pipelines are composed by multiple steps and the propagation of errors through them is a subject which is not yet mastered in photogrammetry. Yet, the actual contribution of super resolution must lie in this step as it is where radiometric information is transformed into a first depth estimation. Therefore, in section 4 we'll present a qualitative analysis of matching when the stereo couple is upsampled using super resolution networks, by means of similarity (or cost) profiles, whose definition has been supplied in paragraph 1.1

2. CREATION OF A VHR SATELLITE SUPER-RESOLUTION SET, TRAINING AND EVALUATION

2.1 Data and methods

For this study, the objective is to dispose of a deep neural network capable of super resolving very high resolution (VHR) pan-sharpened images of Pléiades type (resolution 50 cm), to a target ground sampling distance closer to aerial sensing ($\leq$ 30 cm). At this purpose, a set of multi-spectral aerial acquisitions at 10 cm GSD totaling 1.8 GB was kindly provided by the CNES and used as source data. They consist of twenty 4096x4096 PELICAN (Deliot et al., 2006) images on urbanized areas in France and they have been used to generate both the LR and HR datasets. Even if available and used for pan-sharpening, the NIR band was discarded from the training and result analysis. The CNES also provided the means for the generation of the dataset. They are represented by an implementation and the configurations of the *Chane Simulation Image* (CSI), a tool that allows to apply any step of a satellite image acquisition pipeline to a given image, producing realistic degradation. The source images are converted into luminance values so that acquisition through an imaging system can be simulated taking into account, at least, the modulation transfer function (MTF) and the sensor noise. Hence, the product is resampled at the desired resolution. The following sets are generated: the LR set, at 50 cm taking into account on board and on ground treatments of a real VHR satellite (in this case, Pléiades) product, thus including the addition of noise through the definition of a signal to noise ratio, compression and decompression, denoising, deconvolution and resampling and pan-sharpening operations; the HR set, for which only a dezoom and a quantification in 12 bits are applied. We considered in this work both scale factors 2 and 4 for the networks, fixing thus the GSD of the HR set to 25 cm and 12.5 cm, respectively.

With these settings, training was performed on a a GPU node reserved with 1 GPU NVIDIA Tesla T4, 4 CPUs Skylake 2.2GHz 92 Gb RAM. Moreover, a test portion of these data was kept out of the training. Network hyperparameters were roughly fine-tuned by means of grid and random searches.

Once the training succeeded, the models could be used in inference mode for evaluation. Peak Signal-to-Noise Ratio (PSNR) (Eq. 1) was considered as in most of other super resolution works. It is an inverse measure of the Root Mean Square Error (RMSE) that takes into account the maximum value that a pixel can have L.

$$PSNR = 10\,log_{10}\left(\frac{L^2}{RMSE}\right) \tag{1}$$

Structure SIMilirarity Index (SSIM) is also taken into account s. In (Eq. 2), $\mu_x, \mu_y, \sigma_x^2, \sigma_y^2$ are averages and variances along the x and y direction, the constants c_1, c_2, c_3 are, respectively, $(k_1 L)^2$, $(k_2 L)^2$, $\frac{c_1}{2}$, with k_1 set to 0.01 and k_2 to 0.03.

$$SSIM = \frac{(2\mu_x\mu_y + c_1)(2\sigma_x\sigma_y + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{2}$$

2.2 Results

LR Bicubic RDN ESRGAN HR

Figure 3. Inference on a test image, scale factor 2

LR Bicubic RDN ESRGAN HR

Figure 4. Inference on a test image, scale factor 4

Figure 3 and 4 shows some samples from the test set, which was composed by three images of the same type of the training set but not used for training. RDN and ESRGAN results are compared to LR, HR and bicubic upsampling. Table 1 resumes the quantitative evaluation of the test set images super resolved via the two networks for both scale factors. The results are rather encouraging as the networks outperform bicubic upsampling both in terms of PSNR and SSIM. Superior metrics can be observed in the case of RDN, but perceptually ESRGAN can achieve more sharpness.

	PSNR [dB]	**SSIM**
Bicubic x2	18.54	0.4705
RDN x2	23.88	0.7753
ESRGAN x2	20.97	0.6335
Bicubic x4	17.47	0.3191
RDN x4	22.94	0.5898
ESRGAN x4	19.31	0.4032

Table 1. 2D statistics of the test set

More importantly, the inference super resolution was successful on real Pléiades acquisitions of Montpellier, as we can see in figures 5 and 6. It is noteworthy that in the training set there are no images from the site shown in these two samples. This demonstrates how, in spite of the limited amount of data in terms of variety, the network could generalize a sensor model irrespectively of a specific landscape. Although not measurable, the impression here is that the LR sample is improved in a

perceptual sense. Edges are sharper and objects more detailed, even if not necessarily in a physical manner.

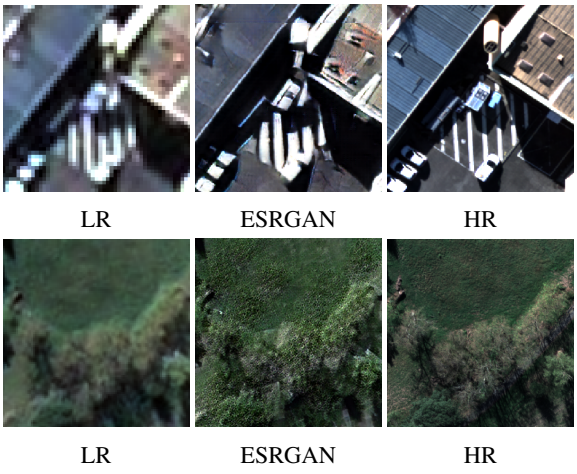
LR Bicubic RDN ESRGAN

Figure 5. Inference on Pléiades data, Montpellier dataset, scale factor 2

Bicubic Bicubic RDN ESRGAN

Figure 6. Inference on Pléiades data, Montpellier dataset, scale factor 4

2.3 Discussion



LR ESRGAN HR

LR ESRGAN HR

Figure 7. Test image, scale factor 4, two cases of ESRGAN hallucinations

By observing figures 3, 4, 5, 6 we can compare the results of the networks. For what concerns RDN, the object contours are well rendered, yet the interiors are artificially smoothened and no meaningful detail is added when passing from scale factor 2 to 4. On the other hand, ESRGAN shows an impressive sharpening capability when pushed to zoom 4, but this comes at the price of evident artifact generation: uniform regions are inconsistently textured and some objects can be even mistaken and resolved as different entities that have been better learned, as in figure 7. In the top example, an air conditioning plant is mistaken for a car. Less common objects are more prone to cause hallucinations, because there are very few (or even absent) samples in the training set and therefore the network upscales them as the known object that more closely resembles their LR version, (i.e. a car). The bottom example shows a texture learned by the network for a tree propagated into a field. These evident artifacts suggest that the utilization of ESRGAN in a 3D pipeline it's potentially critical, because nothing can guarantee that these hallucinations are coherent between left and right images, leading to mismatches.

3. APPLICATION TO DSM GENERATION

3.1 Data and methods

The trained networks were used in inference mode for Pléiades stereo acquisitions of the cities of Toulouse and Montpellier. Two region of interest, one for each site, covering around 3.33 km^2 were considered for this experiment. Left and right pansharpenend images are passed through the networks independently. Then the super resolved images are used as inputs in CARS, CNES' multi-view stereo pipeline (Michel et al., 2020), and the final outputs are compared to the results obtained using the original and the bicubically upsampled data. In other words, a pre-processing super resolution step is added prior to a DSM generation process, identifying 4 cases that will be compared this section: no treatment to the inputs (LR), bicubic upsampling, RDN and ESRGAN inference. Since in order to do stereo vision we need two gray scale images, only one band was extracted from the RGB images for the experiments (green and red were both tested). CARS configuration was kept as default, apart from epipolar error upper and lower bounds that were set to, respectively, 80 and -80. Unlike, Pandora's (Defonte et al., 2021) configuration, the stereo matching core of CARS, was modified to use Zero-mean Normalized Cross Correlation (ZNCC) (Eq 3).

$$ZNCC(I_L(x,y), I_R(x+d,y)) = \sum_{(i,j)\in W} \frac{(I_L(i,j)-\mu_{LW})(I_R(i+d,j)-\mu_{Rw})}{\sqrt{\sigma_L.\sigma_R}} \quad (3)$$

$\mu_{R,Lw}$ and $\sigma_{R,Lw}$ are, respectively, the mean and the standard deviation calculated on the windows used for computing the similarity measure. The window size (dimension in pixels of the side of the square defining the window) was set to 5 for the LR inputs, and scaled with respect to zoom factor: as window size can only be odd, for a scale factor 2 we set this parameter to 9, for a scale factor 4 to 19. In this way, approximately the same content is present for all scales, yet at different resolutions. Furthermore, the penalties utilised for the semi-global matching (SGM) optimization (Hirschmuller, 2007) in Pandora were adapted to the ZNCC measure, by setting P_1 and P_2, respectively, to 2 and 4.

Once generated, the different DSMs needed to be evaluated in order to asses the SR contribution. This was possible thanks to a reference lidar that served as ground truth for the Montpellier dataset. At this stage, a necessary premise has to be put in advance. When it comes to measure the global quality of a DSM, there are not many options a part from the mere calculation of the error with respect to a reference and of the associated standard statistics (Höhle and Höhle, 2009). When relying on a lidar as reference, there might be some temporal differences (e.g. new buildings) in comparison to the utilized data. In other words, 3D metrics is not as developed and reliable as the 2D one when it comes to global analysis. We tried to address this issue by considering, together with the Root Mean Square Error (RMSE) the Normalized Median Absolute Deviation (NMAD) (Eq. 4), a statistic, first proposed in (Höhle and Höhle, 2009), designed for DSMs that is a sort of median more resistant to outliers,

$$NMAD = 1.4826\, median_j(|\Delta h_j - m_{\Delta h}|) \quad (4)$$

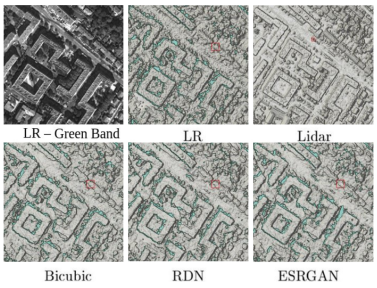
Δh_j being the individual errors and $m_{\Delta h}$ median of the errors.

3.2 Results

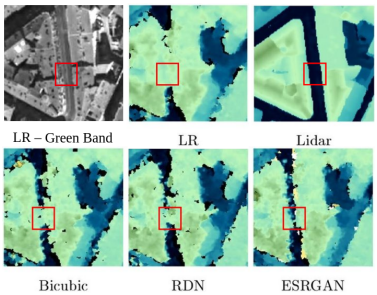
Input stereo pair	% valid points	RMSE	σ [m]	NMAD
LR	95.53	4.17	4.58	1.27
Bicubic x2	94.29	4.02	4.63	0.92
RDN x2	94.99	4.18	4.62	0.84
ESRGAN x2	95.03	4.11	4.64	0.88
Bicubic x4	98.11	4.43	4.66	0.98
RDN x4	98.38	4.65	4.71	0.97
ESRGAN x4	98.18	4.71	4.94	1.13

Table 2. 3D statistics for Montpellier dataset. Measures in meters

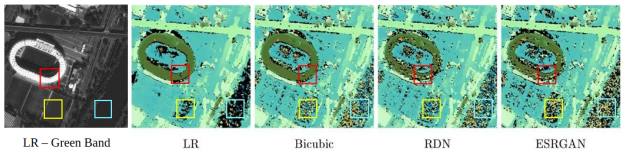
Table 2 resumes the quantitative results on a large ROI for both zoom 2 and 4. In terms of RMSE no improvement could be detected. Moreover, we see an increase when passing to zoom 4 with neural network cases performing worse than the bicubic one. On the other hand, the percentage of valid points grows for a stronger upscale. In practice, objects can be better reconstructed but more outliers are present and the noise amount amplifies, especially when upscaling the input couple 4 times. Nonetheless, we could observe up to 34% gain in NMAD for the RDN case for a scale factor 2. NMAD measure highlights an enhancement when zooming the input images by a factor two, using the networks but also for a bicubic interpolation. However, for a zoom 4 such a gain seems to vanish.



(a) Montpellier, hillshade



(b) Montpellier



(c) Toulouse

Figure 8. Details from experiment result DSMs. Input stereo couple is upscaled by a factor 2 by means of the method in the subcaptions. Lidar sample is shown for Montpellier dataset

3.3 Discussion

Qualitatively, a global modest improvement when upscaling the stereo couple (interpolation or super resolution) is present, as we can observe in figure 8a. Less evident is whether or not SR networks outperform standard bicubic interpolation for this application. In general, deep learning models have the property of well sharpening edges in a 2D image and this can be found in 3D with an enhanced rendering of streets and edges. In Fig. 8b the buildings are better distinguished when applying super resolution and bicubic interpolation, while only in the ESRGAN case they are totally separated. On the other hand, deep neural networks seem not to be advantageous in homogeneous areas and this can be seen in figure 8c as their inputs lead to a failure in well reconstructing the stadium building (red box), and to an amplification noise in correspondence of the river (light blue box) and football terrain (yellow box).

The results shown in this section 3 are not totally satisfactory as: (a) there's no evidence that the use of SR networks outperforms standard bicubic interpolation for the 3D task and (b) forcing a factor 4 zoom doesn't improve the DSM quality. However, the used 3D photogrammetry pipeline might be too complex to have total control on the path of the injected information. In other words, it might be too ambitious to rely on a mere input-output comparison for our analysis and contributions of different upscaling techniques might be flattened in some steps during the process. In order to account for that, a further analysis is proposed is section 4, trying to isolate the SR image contributions. The critical step in this sense is the matching cost computation where a similarity measure is performed in order to associate patches of the left and right images along an epipolar line.

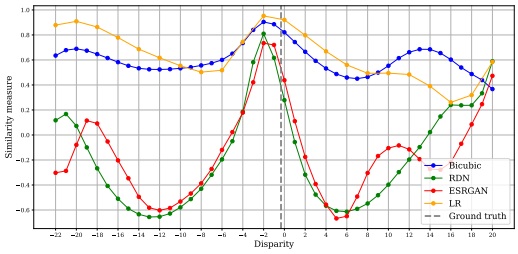
4. SIMILARITY MEASURE PROFILE ANALYSIS

4.1 Data and methods

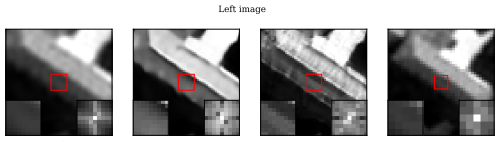
With the purpose of understanding whether and how the similarity measures between image patches are influenced by SR, cost profiles are an useful tool. By locally analyzing these curves and the patches involved in the match, it was possible to find some clues about the influence of radiometric and spectral differences to depth estimation. To do, we set up a tool to visualize, for a given pixel, the plot of the similarity coefficient versus the disparity of the considered cases (LR, bicubic upsampling, RDN and ESRGAN super resolution), the area surrounding the pixel, the windows (also highlighted with a red square in the bigger crop) used in the matching and the corresponding spectra, for both left right and left images. A ground truth ("GT") estimate of the disparity could be retrieved from lidar data following the approach depicted in (Cournet et al., 2020). Since the lidar was only available in Montpellier dataset, the analysis was performed with these data. The red band was considered for this analysis, being the region of interest characterized by tiled roofs. The measure used for its computation is Zero-Mean Normalized Cross Correlation (Eq. 3), with the same window size parameters used in section 3.1 (i.e. 5 for LR samples, 9 for scale factor 2 and 19 for scale factor 4). To illustrate the results of this analysis, we'll present some significant examples, each one corresponding to the matching operation on a single pixel. Although this kind of analysis suffers from a local character, and depends on the specific pixel considered, the samples proposed are representative of a large array of studied matching and they were chosen in order to highlight some

features that have been found multiple times in the experiment set.

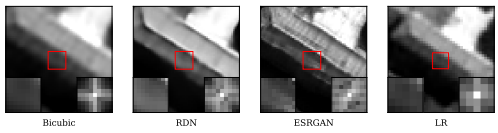
4.2 Results



(a) Cost profile, ZNCC

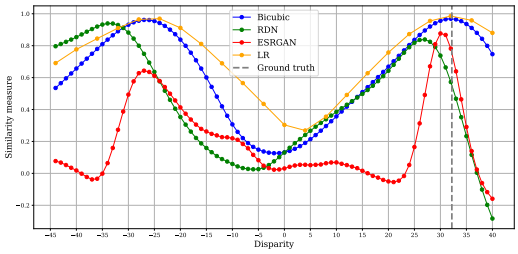


(b) Left image

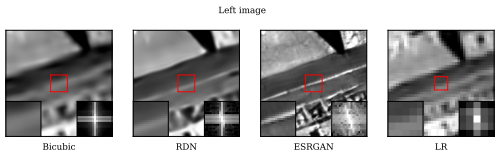


(c) Right image

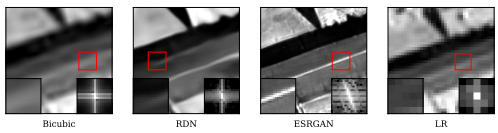
Figure 9. High contrast, illumination change. Cost profile and stereo matching detail, scale factor set to 2.



(a) Cost profile, ZNCC



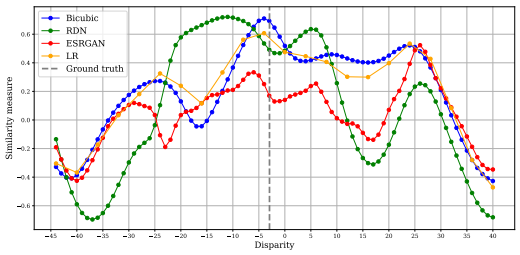
(b) Left image



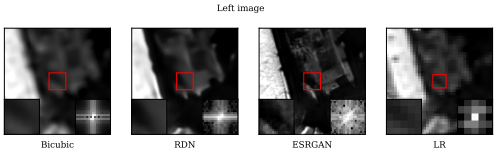
(c) Right image

Figure 10. High contrast, traffic line. Cost profile and stereo matching detail, scale factor set to 4.

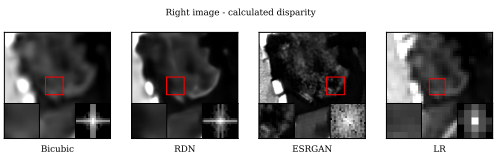
The reported examples of figures 9, 10, 11, 12 will be commented in paragraph 4.3. We remark that for the plotted functions, being correlations, their maxima correspond to the most probable disparities. We are interested in how accurate and well defined the maximum of these functions is with respect to ground truth. Images, windows and spectra help us in giving an interpretation to these curves.



(a) Cost profile, ZNCC

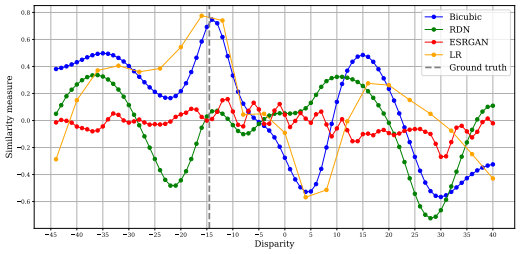


(b) Left image

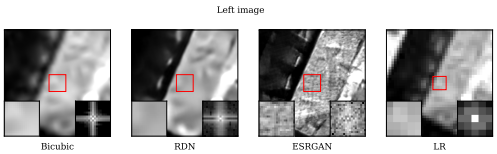


(c) Right image

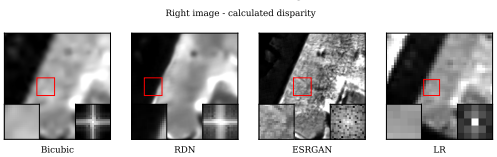
Figure 11. Incoherent artifact, tree. Cost profile and stereo matching detail, scale factor set to 4.



(a) Cost profile, ZNCC



(b) Left image



(c) Right image

Figure 12. Homogenous area, roof, cost profile and stereo matching detail, scale factor set to 4.

4.3 Discussion

Figure 9 shows that with high contrast features, such as the discontinuity between illuminated and dark side of a roof, the confidence in the measure strongly benefits from super resolution. In fact, well sharpened edges and a more precise spectrum lead to correct the disparity estimation and to totally exclude a vast portion of the disparity range because characterized by lower values. The RDN and ESRGAN prediction is more reliable and this introduces stability in the stereo pipeline. This intuition is confirmed in figure 10. This example covers the very same area shown in Fig. 8b. We remark a better resolution of the street when upsampling the inputs to the DSM pipeline. Furthermore, ESRGAN is the only one capable of completely separating the two buildings. Indeed, the impressive definition of this super resolved image leads to a similarity profile in which we are very confident. This sample also supports the hypothesis that, when zoom is forced, ESRGAN is capable of adding significant (yet not necessarily reliable) information, while bicubic and by some degree RDN convey more or less the same information which in these cases cannot guarantee a confident match.

On the other hand, when it comes to uniform or textured areas, the use of super resolution turns out to be disadvantageous. In figure 11, the considered pixel is part of a tree which is not consistently rendered by the SR networks. In ESRGAN case, it is likely to be mistaken for a building in the left image while fairly returned as a tree in right image. This leads inevitably to confusion in the matching step and indeed the ESRGAN similarity function is essentially wrong, whilst the LR original image and the bicubic upscale, although not presenting a selective profile, manage to guess the real disparity with discrete accuracy. By looking at the spectra, we see how RDN and especially ESRGAN force high frequencies even where they're not needed, adding unhelpful details instead of facilitating disparity estimation. As pointed out when illustrating figure 8c, in uniform zones the networks cause noisier DSMs. This might be due to the fact that RDN makes such areas even more homogeneous, while ESRGAN textures them inconsistently. This is to say, stereo matching always needs some sort of contrast, so uniform zones are notoriously complicated areas for stereo matching. Nevertheless, in figure 12 enough information is present originally to guess the actual optimum in LR and bicubic interpolation cases. The considered roof, which presents slight radiometric oscillations in the original data, looks smoothened by RDN in the left and right images. As a consequence, the matching algorithm struggles in finding the disparity. ESRGAN, in turn, strongly textures such a surface, totally filling up the spectrum. However, this texture is something that the network associates to roofs in general, and not specific to the roof in question, leading to a very flat profile that doesn't contain any information on the real disparity.

5. DISCUSSION AND CONCLUSIONS

5.1 Final considerations

Super resolution is one of the most prominent image enhancement techniques for satellite applications being studied in the last years. Nonetheless, the usefulness of such a technology for satellite data based services and products has yet to be proven. DSM generation from VHR data is a significant example that was examined in this study. With respect to previous works, in our testbed we tried to remove sources of bias such as non mastered stereo pipelines, poor quality reference data, inaccurate degradation model for SR dataset generation as well as gaps with respect to state-of-the-art for what concerns SR. Looking at the setup of the experiment and the data used, the most related work is (Zhang et al., 2019), where the authors carefully suggest that there might be an interest in this application. Our findings do not refute this idea, but highlight how, looking only at the final output, results can appear counterintuitive and it might not be clear what is due to super resolution and what to some side effects produced by the DSM pipeline. Therefore, we dug further into the topic by taking into account the essence of 3D stereo reconstruction, i.e. the stereo matching step. In this work, we show how SR spectral contribution in terms of sharpness can be beneficial in presence of high contrast features. However, high frequencies forced by the networks can be non reliable or even unneeded, thus uniform zones become characterized by artifacts, and textured areas may present inconsistency with respect to the reality. This leads inevitably to mismatching and in turn allows uncertainties to propagate through the stereo pipeline canceling out the favorable effects that can be seen in presence of strong contrast. As a matter of fact, the overhead image of city is essentially composed by uniform or textured objects (roofs, parks, squares), divided by discontinuities (building edges, traffic lines), so it might not be worth to be more precise in stereo matching on edges and lines while introducing errors elsewhere.

5.2 Resume

With the aim of improving the quality of the DSMs generated by CARS multiview stereo pipeline (Michel et al., 2020) in urban context, a large scale experiment was carried out, to assess whether deep learning based single image super resolution could be beneficial to this process. Two neural networks, RDN (Zhang et al., 2018) and ESRGAN (Wang et al., 2018), were utilized and a bicubic interpolation counterexample was taken into account as well. A realistic satellite image dataset for super resolution was created, using CNES' image simulation tool to apply the model of Pléiades image pipeline and this led to remarkable results when applying the networks to real VHR data. Visual and quantitative analysis showed how SR was successfully implemented, but this comes at the price of more synthetic images and flagrant local artifacts. Later, we could learn that better 2D metrics doesn't automatically propagate into better 3D models, as SR input pairs do not outperform standard bicubic upsampled pairs when it comes to DSM generation. Moreover, an increase in noise can be observed when forcing a zoom 4, mostly in uniform or textured regions, and in ESRGAN input pair upscaling case. A further local analysis of similarity measures during stereo matching step could give more insight into the contribution of SR to 3D reconstruction from satellite pairs. The hypothesis that a denser spectrum can be beneficial for stereo matching when carried out in correspondence of discontinuities was confirmed, as less errors and more selective similarity functions could be observed where the matching is performed in the presence of high contrast features. On the other hand, uncontrolled artifact generation and inconsistent patterns in super resolved images lead to poor matching in uniform and textured areas. Hence, without addressing these shortcomings, it is not clear whether it's possible to exploit SR potential with reliability in 3D photogrammetry.

Being a relatively new branch of application, further study should be performed to better understand whether these new hypothesis are correct. The results shown in section 4 lack of quantitative insight, although representative of numerous tests.

In order to support the intuitions proposed, it could be useful to project a left image using the disparity map resulting by a matching, and compare it to the right image. In this way, we should be able to highlight on a larger scale the areas where matching was more or less successful. Additionally, one could generate confidence maps by assigning a measure of the reliance at every pixel of a stereo matching as in (Sarrazin et al., 2021). Furthermore, it is possible that other combinations of data/loss can improve the results of the presented and hence supply images better suited for any application, including DSM production. For instance, one could enlarge the data base or perform monochromatic SR, instead of RGB as in this work. Finally, enforcing coherency between left and right images could potentially limit the mismatches caused by uncontrolled artifact generation. Finally, enforcing coherency between left and right images could potentially limit the mismatches caused by uncontrolled artifact generation.

ACKNOWLEDGEMENTS

We acknowledge the CNES which allowed the use of their assets in term of computational power, i.e. the access to the HPC cluster, the use of the CSI for the generation of the dataset and its configuration to simulate Pléiades data, as well as the data used for training (BD Merou PELICAN database) and tests (Pléiades acquisitions): all of these have been necessary elements for the realization of this study.

REFERENCES

Anwar, S., Khan, S., Barnes, N., 2020. A deep journey into super-resolution: A survey. *ACM Computing Surveys (CSUR)*, 53(3), 1–34.

Burdziakowski, P., 2020a. Increasing the geometrical and interpretation quality of unmanned aerial vehicle photogrammetry products using super-resolution algorithms. *Remote Sensing*, 12(5), 810.

Burdziakowski, P., 2020b. A Novel Method for the Deblurring of Photogrammetric Images Using Conditional Generative Adversarial Networks. *Remote Sensing*, 12(16), 2586.

Cournet, M., Sarrazin, E., Dumas, L., Michel, J., Guinet, J., Youssefi, D., Defonte, V., Fardet, Q., 2020. Ground Truth Generation and Disparity Estimation for Optical Satellite Imagery. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 127–134.

Defonte, V., Dumas, L., Cournet, M., Sarrazin, E., 2021. Evaluation of mc-cnn based stereo matching pipeline for the co3d earth observation program. *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, IEEE, 7670–7673.

Deliot, P., Duffaut, J., Lacan, A., 2006. Characterization and calibration of a high-resolution multi-spectral airborne digital camera. *ICO20: Remote Sensing and Infrared Devices and Systems*, 6031, International Society for Optics and Photonics, 603104.

Dong, C., Loy, C. C., He, K., Tang, X., 2014. Learning a deep convolutional network for image super-resolution. *European conference on computer vision*, Springer, 184–199.

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(11), 6792–6810.

Hirschmuller, H., 2007. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2), 328–341.

Höhle, J., Höhle, M., 2009. Accuracy assessment of digital elevation models by means of robust statistical methods. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(4), 398–406.

Kim, J., Lee, J. K., Lee, K. M., 2016. Accurate image super-resolution using very deep convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1646–1654.

Lebègue, L., Cazala-Hourcade, E., Languille, F., Artigues, S., Melet, O., 2020. CO3D, a Worldwide One One-Meter Accuracy dem for 2025. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 299–304.

Michel, J., Sarrazin, E., Youssefi, D., Cournet, M., Buffe, F., Delvit, J., Emilien, A., Bosman, J., Melet, O., LHelguen, C., 2020. a New Satellite Imagery Stereo Pipeline Designed for Scalability, Robustness and Performance. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2, 171–178.

Pashaei, M., Starek, M. J., Kamangir, H., Berryhill, J., 2020. Deep Learning-Based Single Image Super-Resolution: An Investigation for Dense Scene Reconstruction with UAS Photogrammetry. *Remote Sensing*, 12(11), 1757.

Sarrazin, E., Cournet, M., Dumas, L., Defonte, V., Fardet, Q., Steux, Y., Jimenez Diaz, N., Dubois, E., Youssefi, D., Buffe, F., 2021. Ambiguity Concept in Stereo Matching Pipeline. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 383–390.

Szeliski, R., Scharstein, D., 2004. Sampling the disparity space image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(3), 419–425.

Tsagkatakis, G., Aidini, A., Fotiadou, K., Giannopoulos, M., Pentari, A., Tsakalides, P., 2019. Survey of deep-learning approaches for remote sensing observation enhancement. *Sensors*, 19(18), 3929.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C., 2018. Esrgan: Enhanced super-resolution generative adversarial networks. *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 0–0.

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y., 2018. Residual dense network for image super-resolution. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2472–2481.

Zhang, Y., Zheng, Z., Luo, Y., Zhang, Y., Wu, J., Peng, Z., 2019. A CNN-Based Subpixel Level DSM Generation Approach via Single Image Super-Resolution. *Photogrammetric Engineering & Remote Sensing*, 85(10), 765–775.