POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

The contribution of the Super Resolution to 3D Reconstruction from Satellite Image Pairs

Laurea Magistrale in Space Engineering - Ingegneria Spaziale

Author: Nicola Imperatore

Advisor: Prof. Marco Gianinetto

Co-advisor: Loïc Dumas

Academic year: 2020-2021

1. Introduction

1.1. Context

The *Constellation Optique 3D* (CO3D) mission by the *Centre national d'étude spatiales* (CNES) aims at automatically providing a worldwide accurate Digital Elevation Model (DEM). The 3D photogrammetric reconstruction needs two images of the same scene - taken at a different angles - that will be refered as *stereoscopic couple*. For CO3D mission such acquisitions will be taken at the same time by (at least) 2 satellites, thus minimizing temporal differences. This will allow a boost in the DEM accuracy so that it is conceivable to render smaller scale objects like trees and buildings, moving to Digital Surface Model DSM format.

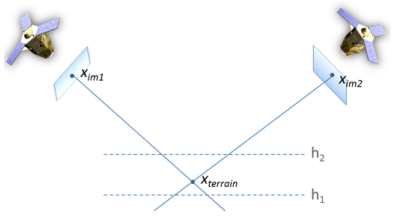


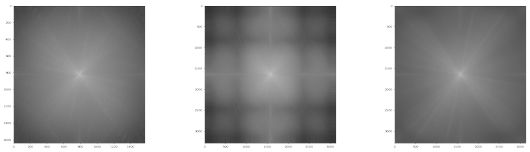
Figure 1: Illustration of the triagulation principle

Such a 3D detailed information is strategic for growing applications in space field downstream, from 3D city mapping to damage assessment, to the incoming smart city market.
CS Group will develop the image ground segment of CO3D mission. The DSM generation pipeline is key and at this purpose, CNES and CS developed two tools: CARS, a multi view stereo pipeline that from a stereo pair generates the corresponding DSM; Pandora [1], which is in charge of the stereo matching step from rectified images.
The shift in the stereo images between homologous pixels, i.e. belonging to the same point on Earth, it's associated at the distance between point and satellite and, by means of a *triangulation* procedure (Fig. 1), we can retrieve the altitude of the point. Thus the problem reduces to find precisely this correspondance from the pixel values. This is called stereo matching and it's the most critical step of the entire DSM chain: it consists of assigning for each pixel its corresponding in the second image, exploiting the radiometric measures of a neighbourhood. The shift between the common pixels in the two images is called *disparity* and can be associated to the depth of the point, i.e. the 3D information.

To estimate a disparity, we slide along image rows in the first image and for each pixel of the second image we compute a cost function that tells the similarity between the two neighbourhoods. Such a function is hereby referred as *cost profile* or *similarity measure* and we can associate at its optimum the estimated disparity.

1.2. Objective

However, DSMs produced with current technology suffer from poor quality in urban areas. To address this issue, one solution may be to increase resolution beyond the sensor limits. One could simply use an interpolation technique (bicubic upsampling is the most used one) but this doesn't introduce any spectral structure that might be used from the matching algorithm to better estimate the disparity. On the other hand, *super resolution* (SR) algorithms are designed to recover high frequencies, introducing significant information in a scene characterized by strong discontinuities such as a city. State-of-the-art methods relying on Deep Neural Networks (DNN) have shown remarkable results in this sense [2] [4]. Fig. 2 shows how the Fourier transform of a neural network super resolved image seems to propagate the spectrum of the image, unlike bicubic upsampling.



LR Bicubic SR

Figure 2: Spectra of an input image, its bicubic interpolation and its super resolved version using deep learning tachniques

The assumption is that this spectral information can enhance the stereo matching step, increasing the confidence we have in the estimation of a the disparity from the similarity measures (Fig. 3). It can be shown that the reliability of a disparity measure can propagate into a stereo pipeline leading to more accuracy in the product [3]. The aim of this work is therefore to assess the contribution of SR Deep Learning techniques to the stereo matching and DSMs generation in space industry. Few similar experiences have been found in literature [5], leaving room for improvement for what concerns both super resolution model training procedure and DSM quality evaluation. 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 model such as deep neural networks.

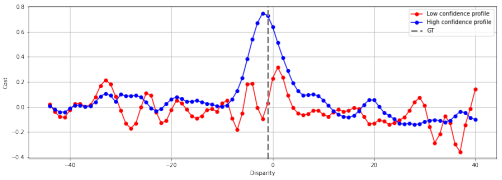


Figure 3: High and low confidence level cost profiles [3]

2. Super resolution via deep neural networks

The general concept of Super-Resolution (SR) refers to those algorithms designed for increasing an image resolution as if a sensor with a higher nominal resolution was used. 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 prior knowledge serves to guide the optimization towards the best achievable solution. DNNs are suitable for such a task as they allow automatic extraction of meaningful highly abstract knowledge, removing the need for identifying case-specific features [2].
Among all possible architectures, residual and Generative Adversarial Networks (GAN) have proved to be interesting solutions. 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. This should add perceptual consistency to the SR image. Together with the Enhanced SR GAN (ESRGAN) [4], the Residual Dense Network (RDN) [2] was implemented, since the latter has an architecture really similar to ESRGAN generator. In this way, we should be able to assess the contribution of a discriminator.

LR Bicubic RDN ESRGAN HR

Figure 4: SR performance on a test image obtained via the CSI, scale factor 4

LR Bicubic RDN ESRGAN

Figure 5: Inference on Pléiades data of Montpellier, scale factor 2

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 doesn't exist, at least for space applications, we usually choose a HR dataset and apply degradation and downsample operations to obtain the LR one. To do so, we assume a sensor model that is applied in the HR-LR transformation.
For this study, the input is represented by very high resolution (VHR) multispectral images (possibly resembling Pléïades products, given their large availability), while the target should have a ground sampling distance closer to aerial sensing ($\leq$ 25 cm). A set of PELICAN acquisitions (multispectral at 10 cm GSD) on urbanized areas in France was kindly provided by CNES; the french space agency also supplied an implementation of the *Chaîne Simulation Image* (CSI). The CSI is a tool that allows to apply any step of a satellite image acquisition pipeline to a given image, producing realistic degradation. It was configured to generate a HR set at 25 cm using a perfect sensor model, and the corresponding LR at 50 cm simulating Pléiades instrument.
The results obtained with these settings are satisfying as the networks outperform bicubic upsampling in standard 2D metrics[1,2] for both scale factor and 4 (Tab. 1) in a test set of the same kind of data of the training set.

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

Table 1: Reference 2D metrics the test set

More importantly, the inference super resolution was successful on real Pléiades acquisitions of Toulouse and Montpellier, as we can see in figures 5 and 4. As a general result, RDN has superior metrics, it renders well the contours. However, it artificially smooths the object interior it doesn't add any meaningful detail 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 6.

LR ESRGAN HR

Figure 6: Test image, scale factor 4, ESRGAN "hallucination": an air-conditioning plant is super resolved as a car transformed into a car

3. Application to DSM generation

The trained networks were used in inference mode for stereo acquisitions of Toulouse and Montpellier to produce SR inputs for the CARS-Pandora pipeline. LR data ad a bicubic upsampled version were also processed.
The error between for the four DSMs and a reference lidar was used as metrics for the possible enhancement. It turns out that in terms of

[1] PSNR=$20log_{10}(\frac{L^2}{RMSE})$
[2] SSIM=$\frac{(2\mu_x\mu_y+c_1)(2\sigma_x\sigma_y+c_2)(cov_{xy}+c_3)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)(\sigma_x\sigma_y+c_3)}$

standard statistics (e.g. mean, root mean square error, etc.) no substantial improvement could be detected. Nonetheless, we could observe up to 34% gain in NMAD, a statistics designed for DEMs which is a sort of median more resistant to outliers. On the other hand, RMSE (Root Mean Square Error) increases for more super resolved input images of zoom 4, suggesting a decrease of measure reliability. However, this also comes with a higher percentage of reconstructed points at this scale. In practice, objects can be better reconstructed but more outliers are present and the noise amplifies, especially when upscaling the input couple 4 times.

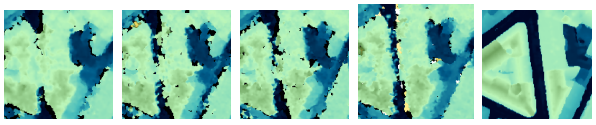


Figure 7: Detail from Montpellier dataset's DSM. Input images upscaled by a factor 2

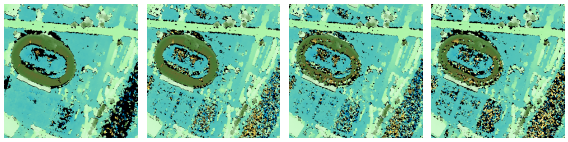


Figure 8: Detail from Toulouse dataset's DSM. Input images upscaled by a factor 2

Input stereo pair type	% valid points	RMSE	NMAD[3]
LR	95.53	4.17	1.27
Bicubic x2	94.29	4.02	0.92
RDN x2	94.99	4.18	0.84
ESRGAN x2	95.03	4.11	0.88
Bicubic x4	98.11	4.43	0.98
RDN x4	98.38	4.65	0.97
ESRGAN x4	98.18	4.71	1.13

Table 2: 3D statistics for Montpellier dataset

Qualitatively, a global modest improvement when upscaling the stereo couple (interpolation or super resolution) is present, although it comes with an increase in noise. Less evident is whether or not SR networks outperform standard bicubic interpolation. 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 renderning of streets (Fig. 7) and edges, whilst they struggle in homogeneous areas and this can be seen in figure 8 as their inputs lead to a failure in well reconstructing the stadium building, and to an amplification noise in correspondence of the river and football terrains.

4. Similarity measure profile analysis

In order to account for these not completely satisfactory results, a further analysis is proposed, this time trying to isolate the contribution of SR images from the rest of the stereo pipeline and thus coming back to the matching step. The idea is to understand whether and how the similarity measures between image patches are influenced by SR. At this purpose, similarity measure profiles are an useful tool. By analyzing locally these curves and the patches corresponding to the match, it is possible find some clues about the influence of radiometric and spectral differences to depth estimation. Two significant examples are proposed. For each one of the considered cases (LR, bicubic upsampling, RDN and ESRGAN super resolution) the plot of the similarity coefficient versus the disparity is presented together with on overview of the surrounding area of the matching pixel, the window (also highlighted with a red square in the bigger crop) used for similarity measure computation, as well as the spectrum of such a window. A ground truth ("GT") estimate of the disparity could be retrieved from lidar data thanks to Beefröst, a tool developed in a collaboration between CNES and CS that produces stereo-rectied images and ground truth disparity maps, from satellite imagery and a 3D reference. We will look for the maxima of the functions, and in particular how much it is accurate and well defined with respect to ground truth. Figure 9 shows how in correspondence of high contrast features, such as the division 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 be-

cause characterized by lower values. The RDN and ESRGAN prediction is more reliable and this introduces stability in the stereo pipeline. On the other hand we have the example in figure 10. The considered pixel is part of a tree which is not consistently rendered by the SR networks. In ESRGAN case, it 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 there might not be needed, adding unhelpful details instead of facilitating disparity estimation.

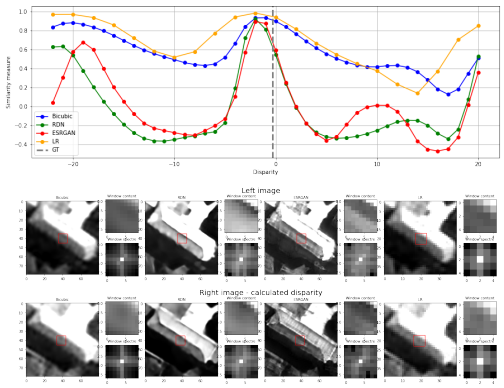


Figure 9: Cost profile sample from Montpellier dataset, red band, scale factor between LR and SR set to 2. SR networks lead to a higher confidence level cost profile

5. Conclusions

With the aim of improving the quality of the DSMs generated by the CARS-Pandora pipeline part of the CO3D mission image ground segment, especially 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 and ESRGAN, were proposed and a bicubic iinterpolation counterexample was taken into account as well. A realistic satellite image dataset was created using the CNES' CSI and this led to remarkable results when applying the networks to real VHR

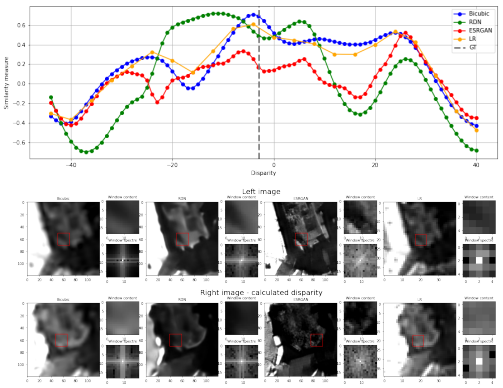


Figure 10: Cost profile sample from Montpellier dataset, scale factor, red band between LR and SR set to 4. SR networks lead to a lower confidence level cost profile and to a wrong disparity estimation

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 upample pairs when it comes to DSM generation, although 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. However, the marks of the different upsampling methods could be highlighted by studying at the stereo matching step, through local analysis of similarity measure profiles. It is confirmed the hypothesis that a denser spectrum can be beneficial for stereo matching when it is performed in correspondence of discontinuities, 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, high frequencies are forced by the networks also where there's no need, thus uniform zones become characterized by artifacts, and textured areas may present inconsistency with respect to the reality. This leads in turn uncertainty 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, that can spread up to the DSM.
Further study should be performed to better understand if these hypothesis are confirmed, and whether it's possible to exploit SR potential with reliability. Additionally, it is possible taht other combinations of data/loss can improve the presented SR networks 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.

References

[1] M Cournet, E Sarrazin, L Dumas, J Michel, J Guinet, D Youssefi, V Defonte, and Q Fardet. Ground truth generation and disparity estimation for optical satellite imagery. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:127–134, 2020.

[2] Juan Mario Haut, Ruben Fernandez-Beltran, Mercedes E Paoletti, Javier Plaza, Antonio Plaza, and Filiberto Pla. A new deep generative network for unsupervised remote sensing single-image super-resolution. *IEEE Transactions on Geoscience and Remote sensing*, 56(11):6792–6810, 2018.

[3] E Sarrazin, M Cournet, L Dumas, V Defonte, Q Fardet, Y Steux, N Jimenez Diaz, E Dubois, D Youssefi, and F Buffe. Ambiguity concept in stereo matching pipeline. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:383–390, 2021.

[4] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[5] Yongjun Zhang, Zhi Zheng, Yimin Luo, Yanfeng Zhang, Jun Wu, and Zhiyong Peng. A cnn-based subpixel level dsm generation approach via single image super-resolution. *Photogrammetric Engineering & Remote Sensing*, 85(10):765–775, 2019.

6. Acknowledgements

I acknowledge my company supervisor, Loïc Dumas, for having guided me in this experience, as well as my academic supervisor at Politecnico di Milano, prof. Marco Gianinetto, for his scientific valuable support throughout this work. Thanks to CS Gruop , the company where the internship valid as the master thesis was host, and to the Centre national d'études spatiales (CNES). The CNES allowed the use of their assets both in term of computational power, and proprietary software such as the CSI for the generation of the dataset, as well as the data used for training. A special mention to my intern and company colleagues for their help in overcoming everyday issues.

POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

The contribution of the super resolution to the 3D reconstruction from satellite image pairs

Tesi di Laurea Magistrale in
Space Engineering - Ingegneria Spaziale

Author: **Nicola Imperatore**

Student ID: 921341
Advisor: Prof. Marco Gianinetto
Co-advisors: Loïc Dumas
Academic Year: 2020-2021

Abstract

CO3D is an Earth observation mission by the Centre national d'étude spatiales (CNES) aiming at providing a worldwide accurate Digital Surface Model (DSM). For this purpose, 3D photogrammetric reconstruction from pairs of satellite images will be employed. CO3D, french acronym for *Constellation Optique 3D*, will be composed by at least four optical satellites that will provide simultaneous acquistions of the same scene. In this way, temporal differences will be minimized, allowing a more accurate stereo matching as well as automatic production of DSMs at a global scale. Such a 3D detailed information in DSM format is strategic for growing applications in space field downstream, from 3D city mapping to damage assessment. The DSM generation is key and the CARS-Pandora pipeline is being developed by the CNES in collaboration with CS Group.

However, DSMs produced with current space technology suffer from poor quality in urban areas, even using very high resolution acquisitions. Indeed, it is not straightforward to render with high accuracy smaller scale objects, such as buildings, by means satellite images that have a limited ground sampling distance capability. To address this issue, one solution may be to increase resolution beyond the sensor limits. One could simply use an interpolation technique (bicubic upsampling is the most used one) but this does not introduce any spectral structure that might be used for a better stereo matching. On the other hand, super resolution (SR) algorithms are designed to recover high frequencies, introducing significant information in a scene characterized by strong discontinuities such as a city. State-of-the-art methods relying on Deep Neural Networks (DNN) have shown remarkable results in this sense. Hence, the aim of this work is to assess the contribution of SR Deep Learning techniques to the stereo matching and DSMs generation in space industry.

Keywords: Super resolution, Digital Surface Model, Stereo-Matching, Deep Neural Networks, Satellite Image Simulation, CO3D

Abstract in lingua italiana

CO3D è una missione di osservazione della Terra del *Centre national d'étude spatiales* (CNES) che ambisce a fornire un modello numerico di superficie (DSM, acronimo dall'inglese *Digital Surface Model*) su scala mondiale. Al fine di ottenere tali prodotti verrà utilizzata una catena di ricostruzione fotogrammetrica 3D da coppie stereografiche di immagini satellitari. CO3D sarà composta da almeno quattro satelliti, operanti nello spettro ottico, che forniranno acquisizioni simultanee della stessa porzione di superficie terrestre. In questo modo, le differenze temporali saranno minimizzate, consentendo una corrispondenza stereo più accurata e una produzione automatica di DSM su scala globale. Questo tipo di informazione dettagliata 3D in formato DSM è strategica per applicazioni crescenti nel downstream del settore spaziale, dalla mappatura 3D delle città alla valutazione di danni urbanistici. La pipeline di generazione di DSM è fondamentale e CARS-Pandora è stata sviluppata dal CNES in collaborazione con CS Group.

Tuttavia, i DSM prodotti con la tecnologia attuale soffrono di scarsa qualità in area urbana, nonostante l'impiego di immagini VHR (acronimo dall'inglese *Very High Resolution*). Infatti, ripordurre con grande accuratezza oggetti di scala inferiore,come gli edifici di una città, comporta serie difficoltà quando si impiegano coppie di immagini satellitari con ovvi vincoli in risoluzione al suolo. Una soluzione potrebbe essere quella di aumentare la risoluzione oltre i limiti del sensore usato, ma una semplice interpolazione non introduce nessuna struttura spettrale che potrebbe essere usata per una migliore corrispondenza stereo. D'altra parte, gli algoritmi di super risoluzione sono ideati per stimare le alte frequenze perse durante il campionamento, introducendo informazioni significative in una scena caratterizzata da forti discontinuità come l'immagine di una città. I metodi allo stato dell'arte, che si basano su reti neurali profonde, hanno recentemente prodotto risultati notevoli in questo senso. Pertanto, lo scopo di questo lavoro è di valutare quale sia il contributo di tali tecniche di Deep Learning sulla corrispondenza stereo nell'ambito della generazione di modelli sumerici di superficie.

Parole chiave: Super Risoluzione, Modello Numerico di Superficie, Stereo-fotogrammetria, Reti Neurali Profonde, Simulazione d'immagini satellitari, CO3D

Contents

1 | Introduction

1.1. Internship topic

The Constellation Optique 3D (CO3D) mission (by the *Centre national d'étude spatiales* hereby refered as CNES) aims at automatically providing a worldwide accurate Digital Elevation Model (DEM). The launch is set in 2023, whereas the global 3D coverage is foreseen by 2025 [25].

Note: A digital elevation model (DEM) is a 3D computer graphics representation of elevation data to represent Earth surface. A digital terrain model (DTM) is a DEM that represents only the bare surface, A digital surface model (DSM) is a DEM which includes objects like trees and buildings.

CO3D places itself within the solid experience of the french space agency for what concerns optical remote sensing with SPOT and Pléiades generations. Indeed, it inherites from Pleiades some design characteristics such as a similar ground sampling distance (GSD) of nearly 50 cm and the spectral bands of the instrument, i.e. red, green, blue and Near Infra-Red (NIR). At least 4 satellites will compose the mission, all of them equipped with the same optical instrument. The orbit altitude is 502 km and the local mean time should be close to eleven o'clock to limit the size of spread shadows and cloud-cover. The satellites will be placed on the same sun-synchronous orbital plane and will work by pair [25].

The reason of this peculiar configuration lies in the principles of photogrammetric reconstruction. In loose words, in order to generate a DEM, one needs at least two images of the scene taken at a different angles. Pléiades products include DEM from stereo pairs, acquired by performing a pitch manoeuvre along an orbit so that a landscape is captured under two different angles. The breakthrough aspect of the CO3D mission is the fact that such stereo captures will be taken at the same time by (at least) 2 satellites, thus minimizing temporal differences. This will allow, on the one hand to improve robustness

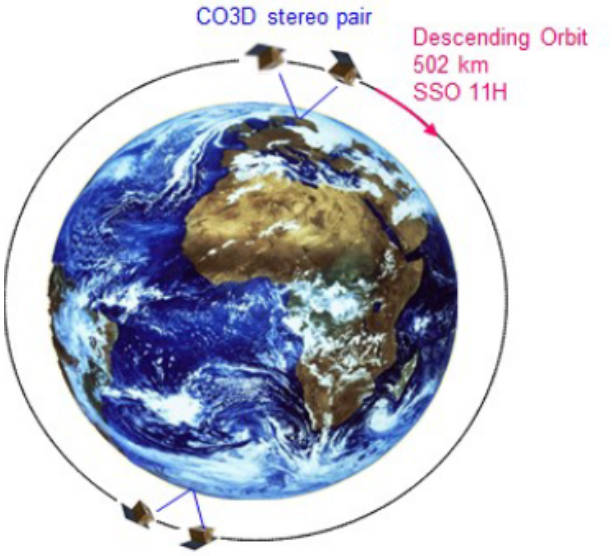


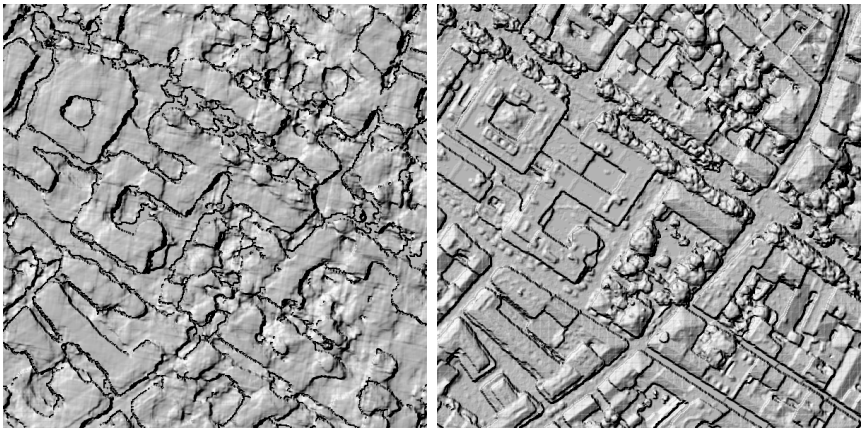
Figure 1.1: Sketch representing CO3D orbital configuration for 4 satellites

of 3D reconstruction passing thus a digital surface model (DSM) format. On the other, it is achievable to render moving objects like such as cars, vehicles, etc. 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 satellites (VHR) of last generation and a better revisit time. CO3D is intended to boost such applications by providing competitive products both in terms of accuracy and speed of generation. DSM format data are significance in the remote sensing context, and their value is foreseen to grow 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 [25]. With respect to other technologies of 3D geographical representation, e.g. point clouds, DSMs are more convenient because simplier to manipulate with current techniques. Moreover, once an adequate satellite is in orbit, DSMs obtained by pairs of satellite image are much easier to be produced than lidar, local acquisitions which require ad hoc campaigns and thus allow less temporal frequency.

CO3D images will be elaborated by the ground segment to deliver high quality final products to French institutions as well as to the public. CS Group is the subcontractor chosen by CNES for the development of the ground segment. The DEM generation pipeline is key since the focus of the mission involves 3D surface models and therefore it is among the main focus for CS Group Department of Payload and Data Applications. For this reason, CNES and CS developed CARS, a new scalable pipeline for stereo reconstruction [35]. The stereo matching step from rectified images is left to Pandora, an independent yet

integrated [5], developed and maintained by CS Group. Such aspects will be adequately detailed in 2. Both frameworks are open source and under improvement[1][2] .

However, when it comes to reconstruct objects at a finer scale, challenges become less easy to overcome. In particular, urban areas renders suffer from poor quality: buildings are not representative of the reality their shape might be unrecognizable. Such defects are a bottleneck for critical applications of such models. For example, when it comes to semantic segmentation of a city image, any detector will struggle in presence of "not squared" buildings, for which is not possible to associate any primitive.



CARS DSM in urban area Lidar DSM in urban area

Figure 1.2: Comparison between photogrammetry DSM and a lidar acquisition on the same area in Montpellier, France

To address this issue and to investigate how to possibly improve CS Group's products and CO3D pipeline, the internship aims at evaluating the contribution of a better resolution of the optical images input of the 3D pipeline, to the final DEM. the basic idea is to "zoom" the input image pair and study downstream the impact this may have. Since sensors have a resolution limit imposed by the characteristics of the instrument, a post processing treatment is needed. One could just "zoom", i.e. interpolate an image on a finer scale grid, but this doesn't add any additional information. Super resolution, on the other hand, refers to the process of recovering a high resolution (HR) image from one or multiple low resolution (LR) versions. This feature could potentially lead to better 3D reconstruction of urban areas, strongly characterized by discontinuities. Furthermore nowadays in super resolution (as in most of computer vision topics) artificial intelligence

[1] CARS documentation
[2] Pandora documentation

in the form of Deep Neural Networks (DNN) is the focus of the related research. Therefore, advantages and shortcomings of the neural networks for such a task will be assessed, meaning that we'll try to investigate how DNN behavior may impact the CARS-Pandora pipeline and the final DSM product.

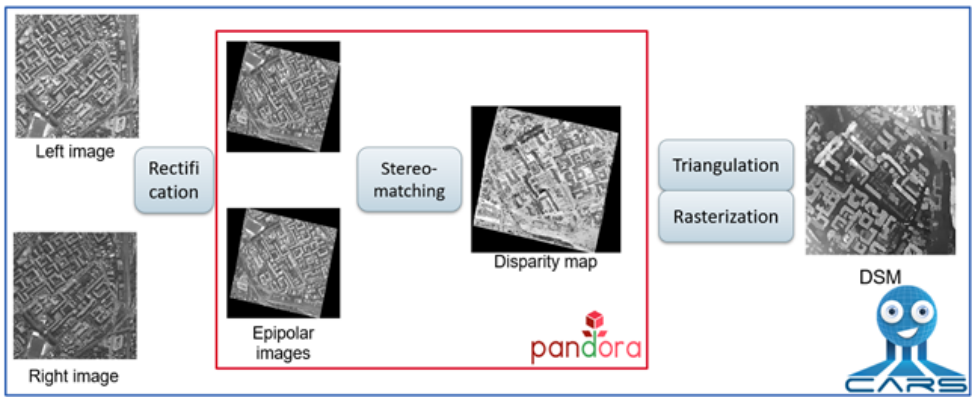


Figure 1.3: Illustration of the CARS-Pandora pipeline

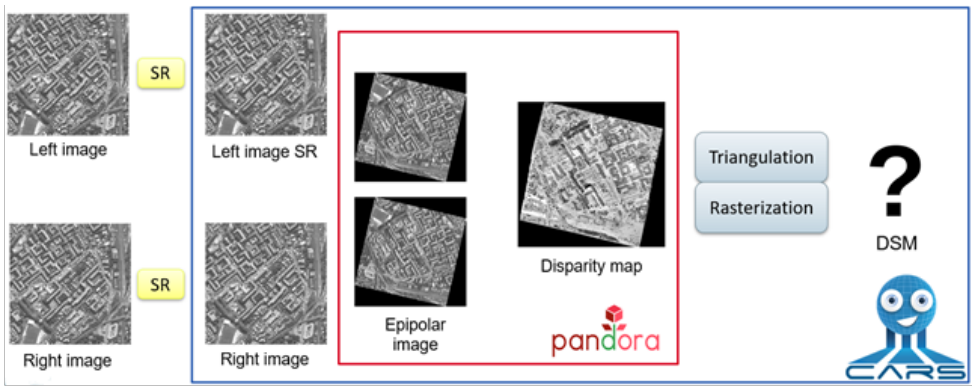


Figure 1.4: Illustration of the proposed contribution to the CARS-Pandora pipeline

This work was mainly done during my intership at CS Group, more precisely at the BU Espace Toulouse, Payload and Data Applications Department. After having defined the context and reported some theoretical justifications (sections 1.2 and 1.3), a state of the art will be presented for what concerns single image super resolution (sections 2.1 and 2.2) and CARS Pandora tool will be presented together with their theoretical fundamentals. In chapter 3 the methodology and results for a single image remote sensing super resolution framework will be illustrated. Chapter 4 will be dedicated to the assessment of the impact of such super resolution technique on the DEM in the CARS-Pandora pipeline. Finally in chap. 6, some conclusions will be drawn together with a final discussion and future work suggestions.

1.2. Company and industrial context

CS Group, initially named CS Communication & Systèmes, is a company active in the domain of design, integration and maintenance of cyber-secured critical systems. With nearly 2000 employees, achieves a turnover of about 200 millions euro. The headquarter is in Paris with other 12 locations in France. 15% of the activities is realised in Germany and Romania, 7% outside Europe. CS works in 5 different industrial sectors: aeronautics, space, energy, defense and industry.

The space Business Unit (BU Espace) is composed of nearly 430 employees and its main clients are CNES and the European Space Agency (ESA). Its activities concern space and ground segments. The BU is further divided into skill centers. This work has been realized within the Payload & Data applications skill center, specialized in ground segment with a solid expertise in image processing and remote sensing. The company took part in some CNES missions such as SPOT or Pléiades. CS Group collaborates with CNES in the realization of open source projects in the photogrammetry domain, such as CARS [35] and Pandora [5]. These softwares are integrated in the ground segment pipeline of the upcoming CO3D mission [25], which will be entirely developed at CS Group. It's the first time that CS is responsible for a CNES ground segment. This allowed a larger financial availability and a good part of the Payload & Data Applications team works on CODIP, the CO3D image processing pipeline.

Moreover, CS Group is part of the AI4GEO consortium, together with Airbus, CNES, ONERA, CLS, IGN and other major actors of the french geospatial industry. It aims to develop a unique solution for the production of automatic 3D geospatial information, lifting the technological barriers to the automatic production of 2D and 3D Geographic Data, exploiting innovative methods such as artificial intelligence algorithms. This internship is therefore also collocated in a set of studies aimed at growing the expertise of CS in Artificial Intelligence (AI) applications within AI4GEO consortium.

1.3. Hypothesis

As stated earlier in par. 1.1, the internship's aim is to implement a super resolution network and apply it to satellite stereo pairs, that are input to a multi-view stereo pipeline (Fig. 1.3), to understand whether this pre-processing step can be beneficial for the pipeline outputs.

However, an artificial increase in resolution is not equivalent to having an instrument capable of a better GSD, so that one may argue that there's no reason to believe that this method can boost stereo reconstruction and that it will be just an increase in resolution for its own sake. In other words, what tells us that such a preprocessing step could be beneficial to DSM production? There are actually two main reasons behind the interest for reducing the sampling interval and in particular for adopting super resolution.

The first argument is linked to stereo matching. Without diving into the details of such a computation (Sec. 2.1.3 is entirely dedicated at this aim), we can say that, in order to estimate the depth of the image objects, we associate at each pixel of the first image a corresponding one in a second image. This correspondence is identified as the one which optimizes a certain function of the distance between the two pixels. Such a distance is referred as *disparity*. Sometimes, it is not as easy to find such an optimum since there are multiple candidates with similar cost values (Fig. 2.6). Increasing resolution means that we have more possible disparities because of a smaller sampling interval, and, consequently, we can better approximate the actual optimum. Figure 1.5 illustrates this idea. A cost function of the disparities for a nominal resolution ("LR") stereo couple and a bicubically ("Bicubic") upsampled version are shown, together with the ground truth ("gt") optimum point, corresponding in this case to the maximum. We see that, thanks to a finer spatial resolution, the bicubic upsample can better define the real maximum and therefore estimate a more accurate disparity. In addition, this can lead to avoid mistakes when two concurrent maximum points are not close to each other and the potential error can be large.

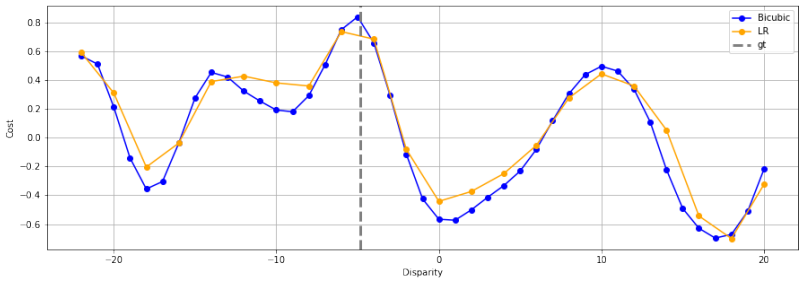
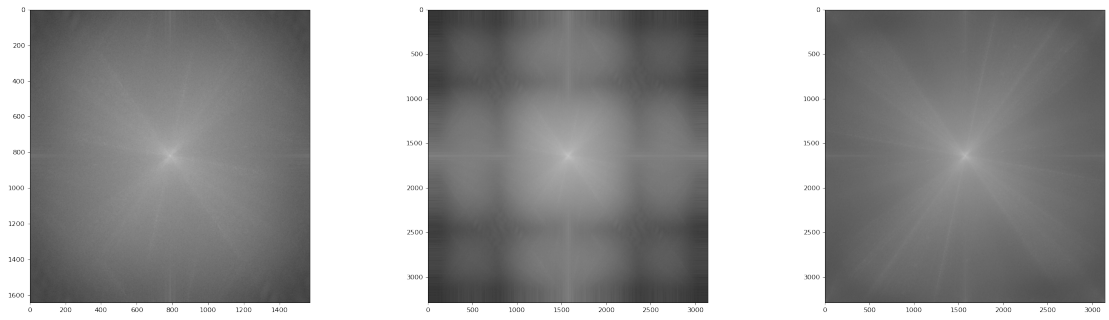


Figure 1.5: Cost functions of a LR image and its bicubically upsampled version. Example of how upsampling an image can lead to find a more accurate cost optimum

This argument might justify a resolution increase but if it was just for it, we could be satisfied with a standard resampling method, without the implementation and training of such a complex object like a neural network. However, if we tackle the problem from a spectral point of view, the perspective changes. Super resolution algorithms are designed

to recover high frequencies from low ones, introducing information, especially in a scene characterized by frequent and strong discontinuities such as buildings in a city. Bicubic or other interpolation methods do not add any high frequency information as the Fourier transform has to be as close as possible to a rectangular 2D function (case of sinus cardinal, perfect interpolator). This is well visible in figure 1.6: the LR spectrum can be found centered in the bicubic spectrum surrounded by a regular pattern with values around zero, whereas the two networks utilised for this study (RDN and ESRGAN, see chap. 3) present a more filled spectrum, that seem to be coherent with the LR values. Such a high frequency information leads to more accurate signal reconstruction especially in the case of discontinuities like building edges. The assumption is that such a more defined spectral information propagates into more accurate matching, as we choose matching according to radiometric similarities/differences. At this purpose, much depends on the consistency of the highlighted high frequencies with respect to ground truth and on the coherence they may have within a couple of images of the same scene.



Low resolution Bicubic upsampling Super resolution (RDN [55])

Figure 1.6: Spectra of 4 versions a satellite image in urban area. Scale factor between LR and the others is 2.

2 | State of the art

2.1. CARS - Pandora pipeline

2.1.1. Principles of stereo-vision

A DSM can be seen as the estimation on a z axis of a given set of coordinates (x, y). Among other techniques to extract such a 3D information from 2D one, stereo-vision is definitely the main one. Indeed, in this work, when we talk about DSM we mean 3D digital models obtained by a stereo-vision processing.

It's impossible to rigorously realize a 3D measure from one single 2D image. Many indexes exist that tell us which objects are in front and which others are behind (e.g. occlusions), yet in certain situations they struggle in being accurate and they're not useful to estimate depth. Stereo-vision is a solution for this problem, it consists of estimating the three-dimensional coordinates of points on an object employing measurements made in two or more photographic images taken from different positions. It's what our brain cortex does in order to assess the depth of the real world, the reason why we have two eyes. Two different images are returned by our two eyes, and they're merged in one single frame together with an estimation of the depth. It is easy to experience this phenomenon by closing on eye: an object close to our face is on the right side of our field of view when we close the left eye, and vice-versa; whereas an object placed far from us apparently remains in the same position when we close on of our eyes. Even if we do not have direct perception, the brain cortex treats the two objects distinctly and the generated information, i.e. the depth of the object, is transmitted to our brain for decision taking.

Formalizing it for an image, we can find common points (i.e. belonging to the same position on the same object) in each image. A line of sight (or ray) can be constructed from the camera location to the point on the object. It is the intersection of these rays that determines the three-dimensional location of the point. Satellite imagery makes no exception and we can use this principle to determine the height of a (x, y) point on Earth surface. The transformation allowing to assign a height measure from a the positions of

common points is called *triangulation* (Fig. 2.1).

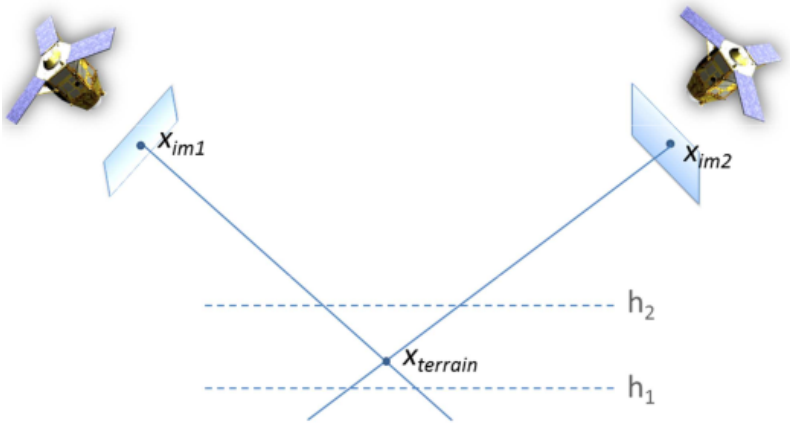
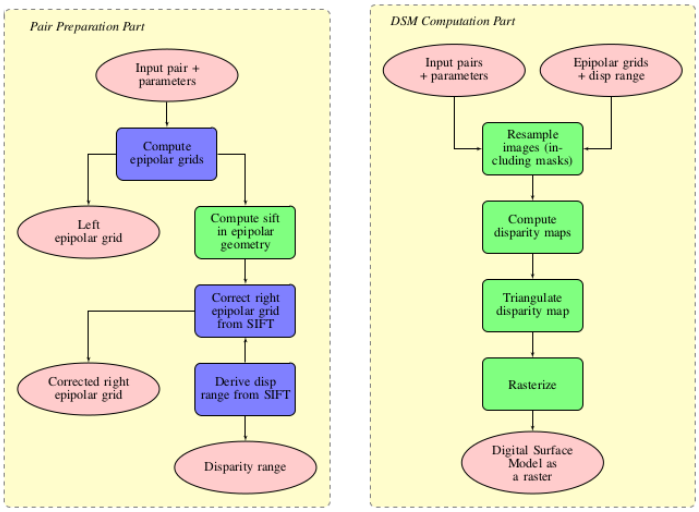


Figure 2.1: Representation of triangulation principle [9]

Now, the question that arises is how the common points are determined. At this point it's important to introduce the notion of image registration. It consists of associating the points of an image to the points of another image in a different reference frame. We call the former the *left image*, the latter the *right image*. Image registration refers to a broad set of methods or applications. Here we'll consider the registration of a *stereo pair*, i.e. two images of the same scene taken at the same instant from different point of views. The distance that separates two homologous points in left and right image is called *disparity* and it's estimated by the registration method, that in this context can be equivalently referred as *stereo-matching* method. In order to find the common points we use the radiometric information present on the images through some sort of algorithm. This is, loosely speaking, the part left to Pandora in the CARS-Pandora pipeline.

2.1.2. CARS

CARS (french acronym for *Chaîne Automatique de Restitution Stéréoscopique*) [35] is a multiview stereo pipeline dedicated to satellite imagery, intended for massive digital surface model production, developed, among other reasons, in sight of the upcoming CO3D CNES mission. It is designed to process stereoscopic acquisitions from existing Very High Resolution optical images such as Worldview3 or Pléïades, on different landscapes (urban environment, mountainous areas, etc.), and to be robust to image defects. It needs in input a stereoscopic pair with the associated geometric model of the acquisition, in the form of Rational Polynomial Coefficients (RPC). The main output is the DSM of the given region of interest (ROI). The main bricks that compose CARS are:



(a) Prepare part (b) Compute DSM part

Figure 2.2: Schematisation of the CARS pipeline

1. **Rectification**: it consists of the transformation of left and right images in epipolar geometry. We say that two images are in *epipolar geometry* when homologous points lay on the same line. It is not a necessary step but it presents a fundamental advantage: once the images are in epipolar geometry, we have to look for corresponding points only along a line, reducing thus matching complexity and possibility of errors. To perform such a rectification, one needs the transformation between a set (x, y, h), i.e. pixel coordinates and altitude, and the global coordinates (λ, ϕ) latitude and longitude, respectively.

$$f : (x, y, h) \rightarrow (\lambda, \phi) \tag{2.1}$$

(Eq. 2.1) denotes this transformation, that is referred as forward localisation function. A representation of such a mapping is always available in the image metadata, for instance in the form of Rational Polynomial Coefficients (RPC), which is a simplified geometrical model of the satellite acquisition that comes with most Very High Resolution data. From these localization functions, f_1 and f_2, respectively, for left and right image, we can derive a colocalisation function that links the coordinates (x_1, y_1, h) of left image to (x_2, y_2) in the right image (Eq. 2.2).

$$f_{1\rightarrow 2}(x_1, y_1, h) = f_2^{-1} \circ f_1(x_1, y_1, h) = (x_2, y_2) \tag{2.2}$$

It is also assumed that we dispose of a coarse DTM that e can use an estimation of h at point (x_1, y_1) and obtain $f_{1\to2}(x_1, y_1, h), h \in [h_{min}, h_{max}]$, the epipolar curves yielded by point (x_1, y_1) from image 1 in image 2 within altitude range $[h_{min}, h_{max}]$

Stereo-rectification operation maps the images into epipolar geometry, i.e. variation of h in the colocalisation function only occur in the horizontal direction. Epipolar geometry does not stricly exist for push-broom images (typical case of VHR data). However, it can be approximated. In CARS, an iterative approximation of a geometry fitting the epipolar constraints is performed. In practice two resampling grids (g_1 and g_2 for image 1 and 2, respectively) are estimated recursively by using the colocalisation function and the fact that disparity in epipolar geometry should be null at h; once determined grid size and starting point by means of an estimation of an average affine transform at full image scale, we move along an epipolar line whose direction is defined as the tangents $\vec{e_1}$,$\vec{e_2}$ to an epipolar curve, ($f_{1\to2}(x_1, y_1, h)$, $f_{2\to1}(x_2, y_2, h)$) in the locations defined by the grids g_1 and g_2. Once completed a line, we move to the next line, along the vector perpendicular to $\vec{e_1}$. To find corresponding point in g_2, we can observe that disparity in epipolar geometry should be null at h. The itaration of this process leads to the completion of the epipolar grids, that will be used along with an interpolation function to resample images to the epipolar geometry, or to convert coordinates between epipolar and sensor geometries.

It follows a sparse matching in epipolar geometry that relies on a SIFT algorithm [31]. SIFT is a well known feature matching method, i.e. that automatically identifies features within an image (e.g. corners, small details) and describes them. It is used for object recognition and image registration based on feature matching, and it combines reliability and good computational performance. In CARS, such a registration exploits the epipolar geometry to look for matches along the same image rows. SIFT matches are sought in tiles generated from the image, since a matching at full scale is not feasible for satellite images. Each tile is defined by an epipolar region $R_1 = [x_m, x_M] \times [y_m, y_M]$ in left image. For finding the corresponding region in the right image R_1, the user sets the ortho-epipolar error upper bound ϵ, as well as δ_m and δ_M, maximum and minimum differences with respect to the low resolution DTM, so that $R_2 = [x_m + \frac{\delta_m}{\alpha}, x_M + \frac{\delta_M}{\alpha}] \times [y_m - \epsilon, y_M + \epsilon]$, where α is the ratio between variations of altitude and disparity in epipolar geometry. The matches are computed for each tile and collected in a global set. They allow us to adjust the epipolar lines, otherwise subject to misalignment because of the imprecision of sensor modeling. This correction is estimated by least square fitting of the error

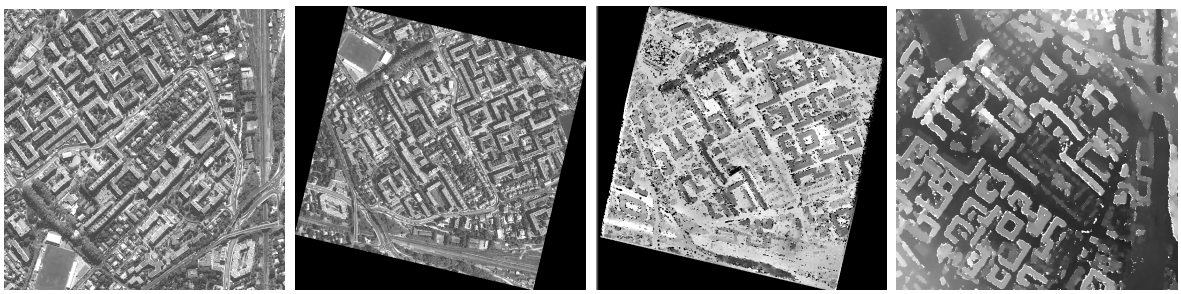
made for the sparse matches. Moreover the SIFT matching allows to estimate a disparity range $[d_{min}, d_{max}]$ that is passed to Pandora in order to limit the research of the disparity to the interval $[x - d_{min}, x + d_{max}]$.

2. **Stereo-matching**: this block represents an execution of Pandora, that takes as input the epipolar images and the range of disparities (as well as, optionally, a configuration file) and outputs the disparity maps. Paragraph 2.1.3 is dedicated to Pandora illustration.

3. **Triangulation**: disparity map is used to find the homologous points in sensor geometry by using g_1 and g_2 functions. Hence, we can project lines of sight from the sensor to the ground points just found and compute their intersection. The 3D point closest to both lines represents the final 3D measurement. Doing it for every pixel, we obtain a 3D point cloud $P = \{(\lambda, \phi, h)_k\}$. Figure 2.1 illustrates this principle.

4. **Rasterization**: with this term we mean the process of converting some data to raster format. To do so, CARS employs a Gaussian weighting algorithm to interpolate the 3D point cloud within a regular terrain grid of user defined resolution. All the points contained in a cell of the grid contribute to the height value that will be given to the cell.

All these legs are in practice grouped in the two CARS steps: `prepare` and `compute_dsm`. The pair preparation step is run pair by pair and mainly produces refined epipolar resampling grids and the estimation of the disparity range. The DSM Computation Step processes the output of the pair preparation step for several pairs and computes a unique DSM from them. CARS has been tested over different cases in terms of landscape (urban, mountainous) and sensor (Pléiades, SPOT7, WorldView3), and it has achieved results comparable to other state-of-the-art stereo pipelines [35].



(a) Input left image (b) Epipolar left image (c) Left disparity map (d) Final DSM

Figure 2.3: Illustration of data transformation through the CARS-Pandora pipeline

2.1.3. Pandora

Pandora [5] is a stereo matching framework that takes in input two images in epipolar geometry and returns the left and right disparity maps. It is inspired by the work of Scharstein & Zelinsky (2002) [43], who proposed a taxonomy of stereo matching algorithms that allows us to breakdown a given algorithm into the following steps: matching cost computation, cost aggregation, cost optimization, disparity computation, subpixel disparity refinement, disparity filtering and validation. Although we won't detail each step, we will spend some words on how a cost volume can be computed and optimized.

Matching cost: Even by restricting our research of common pixel to 1 dimension, i.e to the epipolar line, looking for homologous pixel by pixel would be a main source of error, especially in uniform zones where pixel values are very similar to each other. An immediate improvement is to perform the associations for blocks of pixels (*block-matching*). A neighbourhood surely contains more information or even features that are more discriminant in finding correspondences.

In order to determine how much two blocks resemble each other, we define a similarity measure, that is an operation performed at block level on the pixel values and tells us whether the two blocks are similar. Multiple types of similarity measure exist, from basic differences to neural networks specifically designed at this purpose [53]. The two metrics considered in this work are:

- Census [52]:

$$Census(I_L(x,y), I_R(x+d,y) = \sum_{(i,j)\in w} HAMMING(\hat{I}_{LW}(i,j), \hat{I}_{Rw}(i+d,j) \tag{2.3}$$

where

$$\hat{I}_w(i,j) = \begin{cases} 1, & \text{if } I_w(i,j) < I_w(x,y) \\ 0, & \text{otherwise} \end{cases}$$

and $HAMMING$ represents the Hamming distance operator. The Hamming distance indicates in how many pixels the two windows differ. It's a non parametric measure, it gives less importance to patch radiometry than to the order with respect to the central pixel, characteristic that allows it better results in presence of discontinuities.

– ZNCC :

$$ZNCC(I_L(x,y), I_R(x+d,y)) = \sum_{(i,j)\in W} \frac{(I_L(i,j) - \mu_{LW})(I_R(i+d,j) - \mu_{Rw})}{\sqrt{\sigma_L . \sigma_R}} \tag{2.4}$$

ZNCC stands for Zero-mean Normalized Cross Correlation. $\mu_{R,Lw}$ and $\sigma_{R,Lw}$ being, respectively, the mean and the standard deviation calculated on the window. It is a crossed correlation, centered and normalized, what makes it robust to gains and offsets between the windows. On the other hand, it doesn't perform very well in presence of discontinuities since it's subject to adhesion effects [9].

Adhesion effects are observed in correspondence of depth discontinuities and they can be associated with occlusions, i.e. when a point in an image is not visible in the other image, because of the different angle of view. This is the typical case of a tall building that can hide partially the surrounding ground for off-nadir views. Figure 2.4 illustrates this phenomenon. Part of the ground is occluded by the building in the left frame. Let Q be a point whose distance to the building is less than half of the matching window. We look in the right image for the best correspondent for Q. If the grey level difference between the ground and the building is larger than the intensity variations in the textured areas, a blockmatching method will probably choose P, which means that the disparity accorded to Q will be the same as the one of the building. As a consequence, the reconstructed building will be dilated by the size of a half window. The application of a census matching cost is an efficient way to limit this effect because the elevated radiometry differences in the presence of discontinuities are as important as more limited variations on the non-occulted ground, that in turn may help the matching algorithm to choose the right disparity. For this work, both ZNCC and census have been used and their utilization is specified and justified in section 4.1.

By repeating the measure for each pixel of the epipolar line, we'll obtain a series of values in function of the candidate disparity (Fig. 2.5). We call such a curve the *cost profile*. The algorithm can then select the optimum (maximum or minimum according to the used measure) by means of the *winner takes all* strategy, meaning

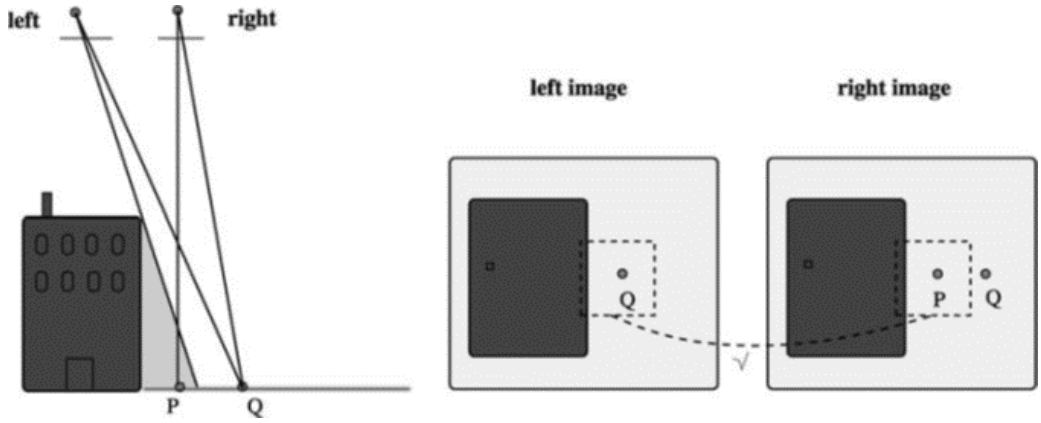


Figure 2.4: Illustration of adhesion effects [7]

that only the best candidate accounts for the chosen disparity.

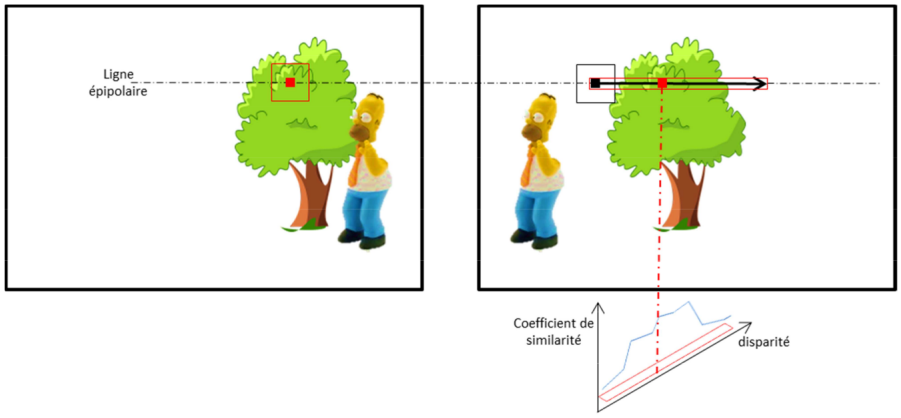


Figure 2.5: Block matching in epipolar geometry [9]

The cost profile is an useful graphic tool to reveal us what has happened during a matching operation. We can use it to understand algorithm errors, robustness and reliability. It can be simply read as a discretized function of one variable, for which we will take its optimum (maximum or minimum depending on the matching cost function used). A critical situation is when two (or more) optima concur, meaning that there are two blocks on the right image that can match the left block, one of which is wrong but accidentally resemble the considered patch. In these case, the algorithm may mistake the wrong optimum for the good one, introducing a strong error. Another unpleasant condition occurs when many similarity coefficients have similar values around the real optimum . When this is the case, we do not have confidence that the chosen measure coincides actually with the correct one. In other words, the ideal cost profile configuration is a single Dirac placed at the correct disparity, Fig. 2.6 exemplifies these considerations by comparing a similarity

function in which we have large confidence and therefore preferable, and another one less reliable as optimum (maximum in this case) is not clearly distinguishable. The additional information contained in such curves allows to better characterize a stereo matching pipeline and in turn to improve its performance. For example, in [42] an ambiguity measure is defined from cost profiles in order to quantify the confidence we have in a disparity measure, and it is shown how this parameter can be used to enhance disparity maps in Pandora.

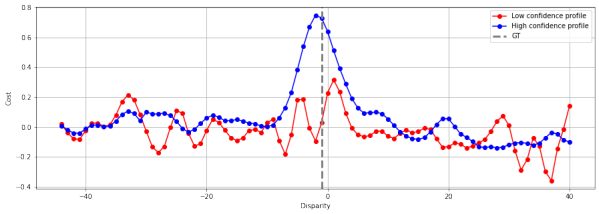


Figure 2.6: Example of high confidence and low confidence similarity measures

Cost optimization: By calculating the cost profile for each pixel along its epipolar line, between an arbitrary range $[d_{min}, d_{max}]$, we will eventually come up with a 3D array called Disparity Space Image (DSI), that stores for each possible disparity d in the considered interval, and pixel coordinates (x, y), a similarity coefficient. We can consider this volume globally and apply thus some sort of global optimization. The approach is to find the disparity map which minimizes an energy defined as in (Eq. 2.5),

$$E(D) = E_{data}(D) + \lambda E_{smooth}(D) \tag{2.5}$$

where E_{data} is the sum of similarity measures and E_{smooth} penalizes the less regular solutions. λ is a coefficient that allows the two measures to be on the right order of magnitude, while D is the DSI of an image.

It is an optimization problem with exponential complexity with respect to the image size and thus it cannot be solved easily. A wide variety of optimization algorithm might be used at this purpose. Pandora implements a semi-global matching (SGM) algorithm proposed by Hirschmuller et al (2008) [14]. It consists of defining a regularization term as in (Eq. 2.6).

$$E_{smooth}(D) = \sum_{p} \sum_{q \in N_p} P_1 \times T[|D_p - D_q| = 1] + P_2 \times T[|D_p - D_q| > 1] \tag{2.6}$$

with N_p the is neighbourhood of a pixel p, T an operator whose value is 1 if the expression is true and 0 otherwise, P_1 and P_2 two arbitrary penalty values. D_p and D_q are the Disparity Space Image image. The first term penalizes small disparities differences with weight P_1. The second term penalizes larger disparities steps, i.e. discontinuities, with value P_2 that has to be larger than P_1.

For each point of the DSI, this energy is calculated along 8 (or 16) 1D paths and summed. The disparity assigned for each point is then the one for which the energy is optimal, i.e. a "winner takes all" strategy. This approach allows a 2D approximation while keeping the complexity of the problem 1D. It is halfway between a 2D approach, which would be a NP-complete problem and hence with exponential complexity with respect to the image size, and a 1D solution that are prone to artefacts due to the difficulty of relating the 1D optimizations of individual image rows to each other.

An appropriate similarity measure together with a SGM-optimization already supply a reliable disparity map in most cases. Other steps of the Pandora pipeline are essentially post-processing and consist of classical techniques of noise reduction, outlier removal, etc. Their contribution is marginal with respect to matching cost computation and optimization and therefore will not be treated.

2.2. Super resolution

The general concept of Super-Resolution (SR) refers to those algorithms aimed at increasing the image resolution, that is, increasing the number of pixels but providing fine details in the resulting image as if a sensor with a higher nominal resolution would have been used [10]. One could zoom, i.e. resample one image, in order to increase the number of pixels, but this does not introduce new information; more precisely, it does not recover the high frequencies lost during downsampling and the result it is just a blurred larger version of the original image. Moreover, some aliasing effects may occur in high frequencies textures. In other words, SR can be associated with the upsampling of an image limiting (ideally, removing) blur, aliasing and noise amplification effects [1]. SR is meant to overcome the limits imposed by the acquisition instrument through data processing.

The interest of the scientific community for SR in imaging dates as back as 1984, when Tsai & Huang et al (1984) [45] proposed a frequency domain method to artificially increase an image resolution. Since, SR has been object of researches and numerous image processing tools have been tested in order to solve the SR problem. Recently, the global

diffusion smartphones allowed billions of people to dispose of a camera. These cameras must be limited in size, hence they're provided with a restricted CMOS (Complementary Metal-Oxide-Semiconductor) sensors that in turn means poorer quality images. This hardware are reaching physical limits that are difficult to overcome, so it is more and more interesting to artificially augment image resolution through post-processing. Thus, related research is boosted in order to retrieve higher quality, eye pleasant, images from the size limited sensors. In remote sensing, large availability of LR data and the ascent of recent applications with high resolution requirements (i.e. object detection) are playing a similar role in encouraging funding for SR research. This, together with the rise of deep learning which is opening new frontiers in image processing, lead to a significant interest in the scientific community for super resolution.

An adequate study of the theoretical basis and the contemporary literature is essential gain confidence in the results that will be proposed in this work. Therefore, some mathematical fundamentals will be illustrated and a state-of-the-art will be presented.

2.2.1. The super resolution problem

Let's consider (Eq. 2.2.1)

$$\boldsymbol{y} = \Phi(\boldsymbol{x}) \tag{2.7}$$

where $\boldsymbol{y}$ is an observed variable $\boldsymbol{x}$ the corresponding model variable, and Φ the unknown operator that links them. The inverse problem consists in retrieving $\boldsymbol{x}$ from $\boldsymbol{y}$.

Although the notion of inverse problem is important numerous different fields, in imaging it represents a fundamental concept. Typically we have one (or multiple) observed image(s) from which we want to extract the pictured information overcoming the limitations of the instrument used [1]. Main applications include denoising, deblurring, fusion as well as super resolution.

The operator Φ is commonly modeled by the contribution of a downsampling operator $\mathcal{D}$, a blur induced by a convolution kernel $\boldsymbol{h}$ and noise (usually additive) $\mathcal{N}$. Therefore, (Eq. 2.2.1) can be rewritten (Eq. 2.2.1) [10].

$$\boldsymbol{y} = \mathcal{D}(\boldsymbol{x} * \boldsymbol{k}) + \mathcal{N} \tag{2.8}$$

The resolution of such a problem consists in the minimisation of a potential J defined by (Eq. 2.2.1). The first contribution is a data fidelity term, while the second one is a regularizer in which we may introduce some a priori knowledge of the model in question. A scale factor α is required in order to be able to compare the two contributions which might be very different in value.

$$J = ||\boldsymbol{y} - \mathcal{D}(\boldsymbol{x} * \boldsymbol{k})|| + \alpha\psi(\boldsymbol{x}) \tag{2.9}$$

The super resolution problem (Eq. 2.2.1) is a notoriously ill-posed problem: infinite solutions exist and the priors term serves to guide the optimization towards the best achievable solution [40].

2.2.2. Super resolution in Remote Sensing

Among other areas relying on the visual information, such as computer vision, microscopy, astrophysics, also remote sensing may benefit from super resolution and therefore there is a strong interest in the research community. In particular, LR data are widely available (e.g. Sentinel-2 data) while HR or Very High Resolution (VHR) data are generally own by companies and in some cases not even available in the market. Yet, almost every application would gain from higher resolution data [49]. On the one hand, LR open source data have generally poor ground sampling (in the order of 10 meters for the visible range, even more for other spectral bands), and this is a bottleneck for many situations. Increasing the resolution would allow to generate perform scientific analysis at a finer scale and to be more precise when generating products from these data. On the other hand, even for HR or VHR data would be beneficial. For example, in some current relevant space downstream use cases like map updating, road extraction and target identification where, high requirements in terms of ground sampling distance, closer to aerial imagery, have to be met. Therefore, there is a natural concern for super resolution, amplified during the past few years by the success of object detection and classification algorithms, which strongly benefit from an increase in image definition.

Note: The notation of HR and LR might have two different meanings within this work: in this paragraph, we adopted the vocabulary of satellite imagery, where data with a GSD lower than 1 m are generally referred as HR, while the others as LR. In the context of super resolution, though, LR is associated to the input image, while HR to the target one, independently on the actual GSD of the data.

The most dynamical application of SR is in computer vision for enhancing natural images. Natural images is the term used in literature to indicate photos of a scene in its natural environment, taken on ground by standard RGB camera, like anyone of us does with his smartphone. These are the typical images used as benchmark for computer vision algorithms. This denomination is often used in contrast to aerial or satellite images, that are taken onto the Earth surface from above with specific instruments. Remote sensing presents some specificities that have to be taken into account when designing methods for SR [44]. We point out some of them:

- Larger size of the image to treat. Remote sensing data are usually supplied in the form of very large acquisitions over a given range of latitude and longitude. They can be up to several hundreds of MPs while in computer vision we usually are in the order of some megapixels. Thus, we typically need to set some criteria to divide space images into tiles so that this amount of information can be processed in reasonable times.
- Big amount of information encoded in one observation: cities, vegetation and land may be pictured in one image. This means that it is difficult to introduce prior knowledge valid in the totality of the image.
- Various object orientation. Even objects of the same type can be oriented in different ways, depending on the scene and on the satellite direction. This, again, introduces concerns when modelling an a priori that can work in as many cases as possible.
- Satellite acquisition chain peculiarities: CCD (Coupled Charge Device) sensors, satellite movement, n-bit encoding, compression, etc. The trouble we might encounter are multiple, e.g. the complexity of modelling all these different contributions, the exact knowledge of all the parameters composing the considered camera model.
- Challenging conditions of the scene, like atmospheric contribution, presence of clouds, etc. The scene might not be uniform and well visible.

Nevertheless, the use the results of computer vision SR methods in remote sensing (or

vice-versa) is motivated by the fact that the literature of the two fields is closely correlated in direct or indirect means [40]. On the one hand, this is justified theoretically by the fact that a robust model should be able to inverse different kinds of sensor model [48]. On the other, such an approach is motivated empirically by numerous successful experiences in the dedicated literature [29] [38] [48]. Especially in the past decade, that saw computer vision as one of the most growing fields in the scientific research, remote sensing applications often follow the main trends of computer vision. What is typically done is to borrow methods proven for natural images, and adapt them in the context of overhead imagery. For instance, by addressing the multi-scale aspect of such data [13] [54] [17].

2.2.3. Super resolution algorithm classification

A notable amount of methods have been proposed to solve the ill-posed super resolution problem (Eq. 2.2.1), involving classic tools of the signal, estimation and probability theories. Several works proposed a review of SR algorithms, both in remote sensing [10] [40] [44] and in computer vision [49] [1].

A first distinction can be drawn between single image super resolution (SISR), which attempts to recover a HR version of a LR input image, and multiple image super resolution (MISR), where typically sub pixel differences between photos of the same scene are exploited to extract the subpixellic information. In the case of remote sensing, MISR can be further broken down based on the sequence of used images, whether the latter is multispectral, multiview or multitemporal. Although multispectral and multitemporal SR methods are receiving greater attention due to the increasing of revisit time and spectral bands of the publicly available data [36], in the context of the CO3D mission only multiview approaches might be interesting, when the information redundancy included in a stereo pair (or n-views) can be exploited for further refinement of image products.

However, in this work we'll mostly focus on single image approach. Many criteria can be identified to classify the different algorithms (domain, theoretical tools employed, etc.). Based on the reviews proposed in [10] [40], we can observe a main distinction between reconstruction and learning based algorithms. The former arises from traditional signal treatment techniques in both frequency and spatial domains by attempting to produce features appearing in the LR image to a higher resolution level (Fig. 2.8). The latter relies on the automatic estimation of a LR to HR transformation thanks to examples elaborated by the algorithm (Fig. 2.7). In other words, we try to extract the mapping between LR and HR from external examples and then apply this transformation to our image [10]. The main shift of interest we can remark in remote sensing in the last decade

is indeed the passage from reconstruction based to learning based methods.

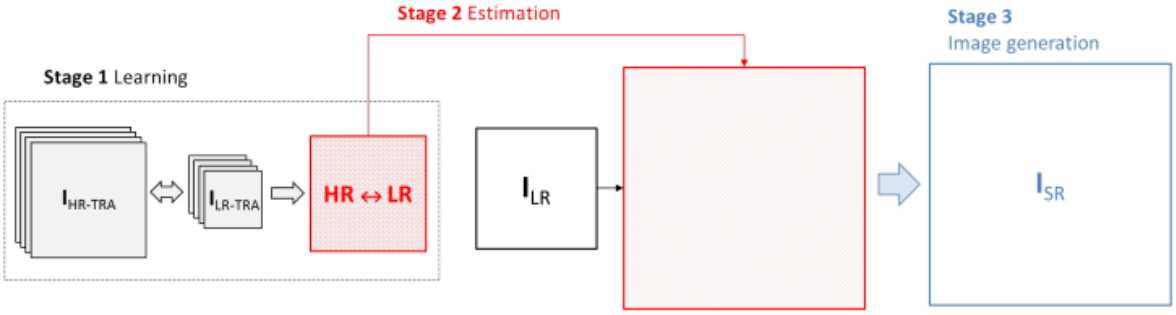


Figure 2.7: Representation of single image resolution via learning, Ref. [10]

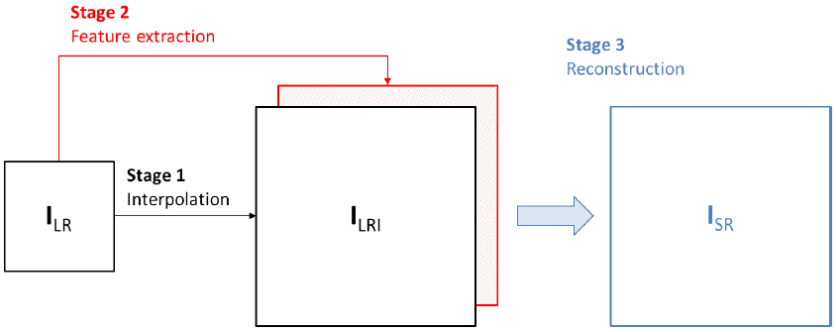


Figure 2.8: Representation of single image resolution via image reconstruction, Ref. [10]

Based on the reviews in [10] and [40], it is possible to delineate the following brief, non complete, overview.

- Interpolation

  Resampling at higher resolution doesn't mean necessarily super resolving. This is the case of interpolation: some literatures [40] include it in SR methods while some other [10] exclude it because, in fact, there's no real attempt to recover high frequencies. Indeed, the perfect interpolator is a cardinal sinus whose spectrum, by definition, is empty in correspondance of frequencies that have values larger than the starting image. Moreover, it cannot be considered a research topic as interpolation is generally included in every image processing formation and well-known interpolation methods (bilinear, bicubic) are commonly implemented in any image related software. In this state-of-the-art bicubic interpolation will be briefly described because any SISR paper compares its results to a standard bicubic interpolation, and so it will be done in this work. Such a benchmark status is due to the fact that bicubic interpolation is the in practice the best cost-effective method when it comes

to resample an image. Indeed, it preserves fine detail better than the common bilinear algorithm. However, due to the negative lobes on the kernel (Fig.1.6), it can cause clipping, which is an artifact. On the other hand but it increases apparent sharpness, and thus it can be desirable.

As the name may suggest, the bicubic interpolation is the 2D extension of the cubic interplation. In digital imaging it is achieved by applying a convolution with the kernel in(Eq. 2.10) in both dimensions. a is usually set to -0.5 or -0.75. The approach was first proposed in [22] and sixteen points (a 4x4 grid) are considered for the operation.

$$W(x) = \begin{cases} (a+2)|x|^3 - (a+3)|x|^2 + 1 & \text{for } |x| \leq 1, \\ a|x|^3 - 5a|x|^2 + 8a|x| - 4a & \text{for } |x|1 < |x| < 2, \\ 0 & \text{otherwise.} \end{cases} \tag{2.10}$$

- Reconstruction based
  - Frequency domain: these methods concern transformation-based operation that perform mathematical operations on the frequency components [40]. It naturally comes to find a way to link high and low frequencies, whether using Fourier or wavelet transforms. The first image super resolution work is usually attributed to Tsai & Huang et al (1984) [45], in which they derived a system equation that describes the relationship between LR images and a desired HR image by using the relative motion between LR images. The frequency domain approach is based on the shifting property of the Fourier transform, and the aliasing relationship between the continuous Fourier transform of an original HR image and the discrete Fourier transform (DFT) of observed LR images. These properties make it possible to formulate the system equation relating the aliased DFT coefficients of the observed LR images to a sample of the CFT of an unknown image.

    Although the analysis in a frequency domain can be powerful, these approaches suffer from the proper modeling of the motion with real-world problems [40].
  - Space domain: the focus progressively shifted towards spatial approaches, that allow to overcome frequency domain limitations. For example, the iteration back projection (IBP), introduced by Irani et al. et al (1991) [18] is one of the most cited interpolation methods in the dedicated literature. The SR image is estimated by back projecting the difference between simulated LR images via

imaging blur and the observed LR images. The reconstruction process is realized by minimizing the energy of the error iteratively. A foremost experience in super resolution is the SPOT-5 super mode, that combined hardware and image processing to overcome the constraints of the flying optical system. This system shifts half a sampling interval in the horizontal and vertical directions by a double CCD linear array, which obtains two panchromatic 5 m resolution images, and then produces an approximately 2.5 m resolution high-resolution image through super-resolution reconstruction processing [24].

Space domain approaches pushed forward super resolution limits and lead and to conceive super resolution using a single frame, i.e. the SISR, that sees its more recent developments in learning based techniques.

- Learning based: SR research is nowadays mostly focused on this set of methods. SISR can be more easily achieved when priors can be automatically modelled by the algorithm. Yang et al. et al (2010) [50] introduced sparse coding (SC) for super resolution and their result is considered among the best non deep learning techniques to achieve SISR. SC takes advantage of the fact that natural images tend to be intrinsically sparse when they are characterised as a linear combination of small patches [10]. Firstly, an overcomplete dictionary from the training patches is learnt by forcing the high resolution training images and their low-resolution counterparts to share the same sparse codes. Then, each test LR patch is expressed in the dictionary with sparse coefficient. Finally, the HR image is reconstructed with the weighted coefficient computed in the previous step [49].

Over the last years, learning based techniques relying on DNN became definitely the main trend in SISR. Sec. 2.2.4 is entirely dedicated at this subject which is in turn the main focus of this work. Data driven approach seem to overcome the limitations of the previous generation methods, when adequately supported by proper databases and computational power. However, many other methods exist, each one with its advantages and shortcomings. Bicubic interpolation usually represents the benchmark to be outperformed. Although it produces blurred images as it doesn't add any high frequency, it represents the best cost-effective technique when it comes to upsampling an image.

2.2.4. Deep learning in super resolution for remote sensing

CNNs (Convolutional Neural Networks) are specific DNN architectures which involve the use of convolutions in several layers, and are an ideal tool for processing regularly sampled

data such as 2D and 3D imagery [44]. Almost every field of research which deals with visual data is nowadays trying to understand the potential of such artificial intelligence (AI) architecture, including super resolution in remote sensing. Deep Learning success boosted research for computer vision tasks, such that sophisticated DNN techniques for SR showed to be able to outperform past methods.

In SR context, they are appealing because of their ability to find an optimum solution by inferring effective high level abstractions that bridge the LR and HR spaces [51]. In other words, DL allows automatic extractions of meaningful a priori knowledge, removing the need for identifying case-specific features [44].

Architectures

The first work which proposed Deep Learning for SR is Dong et al (2014) [8], where they demonstrate the equivalence between the contemporary state-of-the-art learning base methods (such as sparse coding) and a deep convolutional network. The proposed network, called Super Resolution Convolutional Network (SRCNN), is the first of its kind and thus utilized as base reference for any other SR via Deep Learning work.

The fundamental difference in comparison to the classical CNN tasks such as object recognition or classification, is that the mapping is performed from a high dimensional space to another high dimensional space (unlike classification where a MxN image is converted into a limited number of classes). Therefore particular architectures (e.g. Autoencoders, Generative Adversial Networks) have to be designed which may differ from the traditional deep neural networks that map an image into a set of output features. This is common to other image processing tasks like image restoration, image generation.

In remote sensing, the first application of CNNs for super resolution is commonly attributed at Liebel et al (2016) [29]. They show how re-training a CNN designed for single-image super resolution using an appropriate dataset for training can enhance multispectral images of Sentinel-2 database. The idea is to train a proven CNN (in fact, SRCNN) with a different dataset, assuming that the properties of a remote sensing image only affect the parameters of a network and not the structure itself.

Such an approach was successful enough to encourage the remote sensing community utilizing this strategy: train state-of-the-art networks for natural images with appropriate datasets (or even just fine tuning natural images trained networks with an ad-hoc database, see paragraph 2.2.4). In other words, architectures are taken from computer vision with marginal modifications. On the other hand, large scale datasets for remote

sensing super resolution were lacking and researchers proposed different solutions.

SRCNN [8] proposes a very basic structure in which at each layer corresponds a fundamental step: the first convolutional layer extracts a set of feature maps, the second layer maps these feature maps nonlinearly to high-resolution patch representations, and the last layer transforms the predictions to produce the final HR image.

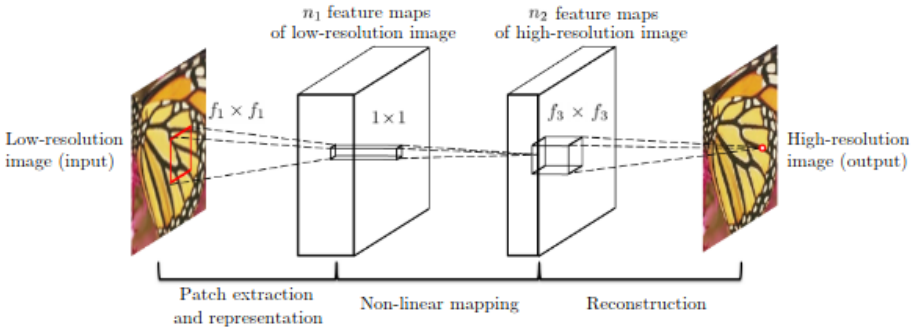


Figure 2.9: Schematics of the first CNN for single image super resolution [8]

For its simplicity and its effectiveness, SRCNN is an important milestone in SR literature. Nonetheless, it left space for improvement. For instance[51]: the input of the network is the bicubically interpolated version of the LR image (time consuming and the kernel used may introduced undesired information); the structure is only 3 layer deep, while it is believed that deeper networks can harness more sophisticated transformations; the loss function (Mean Square Error) is trivial, while it would be interesting to use it in order to inject some prior knowledge in the model.

In fact, first CNN for SR [8] [29] barely outperformed bicubic interpolation while increasing very much the model complexity and the amount of data to handle. Therefore, much work has been done to propose more satisfying architectures: different DL approaches can be identified in the literature. Here we propose some examples of architectures used for enhancing the resolution of remotely sensing data, inferred from the reviews [40] [51] [44] [1]. Although all these architectures are based on convolution operations, we recognize GANs in a different category because of their peculiar framework, where two networks try to "fool" each other: the first has to generate a high resolution image from the low resolution one, while the second network will try to understand which one between the SR and HR image is real. Some examples of CNN architectures are hereby reported, but it is far from being a complete list of what can be found in literature.

- Convolutional Neural Networks (CNN)
  - **Linear**: it is the straightforward way of building a CNN, i.e. the image is

passed through a series of layers progressively; in fact, nothing conceptually different from SRCNN depicted in figure 2.9. The main way for improving the results is to add layers, to make the network deeper to improve the learning of LR to HR mapping. That's the basis on which the Very Deep Residual Network (VDSR) [23] was created in natural images context, and then used as benchmark in several remote sensing SR works. It represented an improvement with respect to the SRCNN but it was inevitably outperformed by more elaborated networks.

- **Residual**: skip connections in the network design avoid gradients vanishing and make it feasible to design very deep networks. In this approach, algorithms learn residuals i.e., the high-frequencies between the input and ground-truth, [1], making them interesting for super resolution since the goal is to retrieve the missing high frequencies. Haut et al (2019) [13] proposed residual learning in SR for remote sensing. Specifically, residual units and skip connections were adopted to uncover more relevant features on both local and global image area.
- **Recursive**: as the name indicates, some layers/ units are recursively connected. This should facilitating the laerning process breaking down the main task into a set of smaller problems [1]. An interesting implementation of such networks in remotely sensed data is Ma et al (2019) [34]: the network operates in the wavelet domain, where low to high frequencies relations may be more naturally described. Nevertheless and in spite of the network complexity, such a method couldn't make a big leap forward in terms of results.
- **Multibranch**: the feature mapping is divided into different paths at multiple scales. This way the model should be able to learn multilevel representations, which in remote sensing is of paramount importance given the diverse range of feature scales in Earth Observation data. A relevant experience often reported in the literature is Lei et al (2017) [34],' Local global combined network (LGCNet): a "multifork" structure is introduced in the non-linear mapping step (Fig. 2.9) in order to treat both local details and global environmental priors.
- **Attention based**: a relatively novel approach in which one considers that not all the features (channel, spatial locations) are essential for super-resolution but have varying importance [1]. Several works have been proposed by the aerial/space imaging community. We may consider the work from Zhang et al (2020) [54], characterized by a feature extraction network and a feature

refinement network with high-order attention mechanism for detail restoration, because of its ability to restore fine and "straight" building edges.

- **Combined**: the approaches just depicted (residual, recursive, etc.) are often combined to build more elaborated and remote sensing oriented networks.In the benchmarks [40] [44], the multi-scale residual neural network (MRNN) by Lu et al (2019) [32] attains results comparable to GANs. MRNN addresses the multi-scale nature of satellite images to reconstruct high-frequency information accurately. Different sizes of patches from LR satellite images are extracted; large-, middle-, and small-scale deep residual neural networks are designed to simulate differently sized receptive fields for acquiring relative global, contextual, and local information for prior representation. Then, a fusion network is used to refine different scales of information.

However, improvements are accomplished at the cost of a further ramification and deepening of the network. The impression is that such a complexification is not always balanced by breakthrough performances. It seems that such an approach is reaching a sort of "plateau" and it is hard to distinguish an imposing solution, while the focus is shifting towards GAN.

- Generative Adversarial Networks (GAN)

They consist of a generator network and a discriminator network that produces high quality and realistic reconstruction of super-resolved images. It is acknowledged to Ledig et al (2017) [27] the creation of SRGAN,the first GAN for super resolution. The adversarial strategy pushes the generative result towards more natural outputs which please the human eye. This is achieved by the introduction of a perceptual loss function (for the generator) which consists of a content loss, i.e. pixel values similarly to standard CNNs, and an adversarial loss that involves the response of the discriminator (Eq. 2.2.4). In this way, prior knowledge is synthetized by the discriminator and taken into account by means of the generator loss function. In SR, a variant of the original GAN is in fact employed, namely conditional GAN, where the input is the LR image (or a version of it) instead of random noise. Haut et al (2018) [12] proposed a GAN for training in an unsupervised way as first application in remote sensing. Their work is more similar to "traditional" GANs, the generator starts from random noise to create a the LR image at different scales, and the transformation learned is then applied to the LR data to obtain super resolved data. In such manner, there's no need for supervision and of an external training set composed by HR data. However, this approach didn't impose among

others, whereas most of state-of-the-art methods use the LR-HR pairs to generate the network knowledge.

$$\mathcal{L} = \mathcal{L}_{cont} + \alpha \mathcal{L}_{adv} \tag{2.11}$$

GANs generally improve perceptual quality but may introduce large number of artifacts and indeterminate details. However, well-tuned architectures and good regularizations have allowed GANs to become the current reference when it comes to SR.

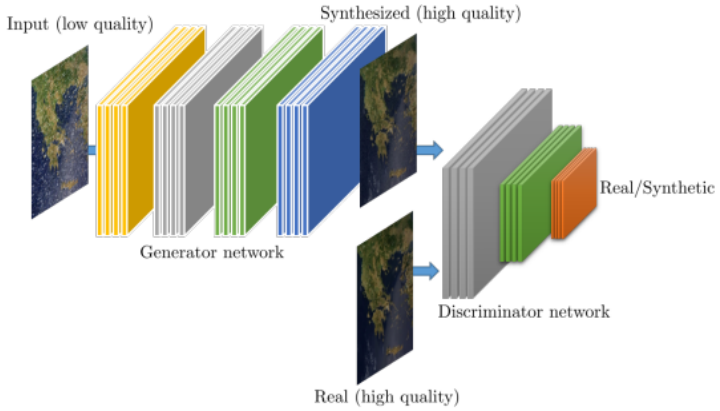


Figure 2.10: GAN representation for super resolution [44]

The first GAN for super resolution[27] saw soon the rise the rise of other concurrents. A relatively simple yet effective approach was presented by Ma et al (2018) [33]. They propose a simplified version of the prototypical SRGAN [27] capable to improve the quantitative results. One of the main dinding of the work is that the removal of batch normalization layers boost both accuracy and speed of architecture. Batch normalization is considered suitable in the field of target classification, rather than in SR. They also adopt transfer learning from computer vision, where larger datasets exist, as an effective way to perform training of SR network in remote sensing. More specifically, they performed pre-training on the DIV2K, a natural image dataset often used in for training SR networks. Generally speaking, the features learned from the former part of the network are in low level and can be shared across different tasks, while the features learned in the later part are specific to the target task. Thus they fix the parameters in the former part of the network and only finetune the last three convolutional layers. A weak point of the paper is the use of a limited dataset (UC merced) [40].

A significant work in recent literature is Jiang et al (2019) [20], who proposed edge enhanced GAN (EEGAN) (Fig. 2.11). As the name may suggest, such a network was designed with the purpose of improving the sharpness of edges, thus leading to superior edge detection results. A generator network composed by several dense layers is firstly used to obtain an intermediate HR result I_{base}. Since the latter may suffer from noise, a second structure is included, whose task is to enhance the target edges extracted from the intermediate SR image by cleaning up the noises and artifacts. I_base is passed through a laplacian operator to extract edges, hence a dense subnetwork is utilized to infer fine edge maps. Such maps are noisy and will cause difficulties for subsequent discrimination. Therefore, a mask branch simultaneously learns the noise mask through the attention mechanism. The combination of these two representations produces an enhanced edge information I^*_{Edge}. By replacing these purified edges in I_{base}, the final image is obtained. Finally, a basic discriminator network will assess the quality of the result by comparison with the ground truth. In the loss function, in addition to the GAN loss (Eq. 2.2.4), we remark the presence of a consistency term whose purpose is to reduce potential artifact generation, weak point of GAN architectures.

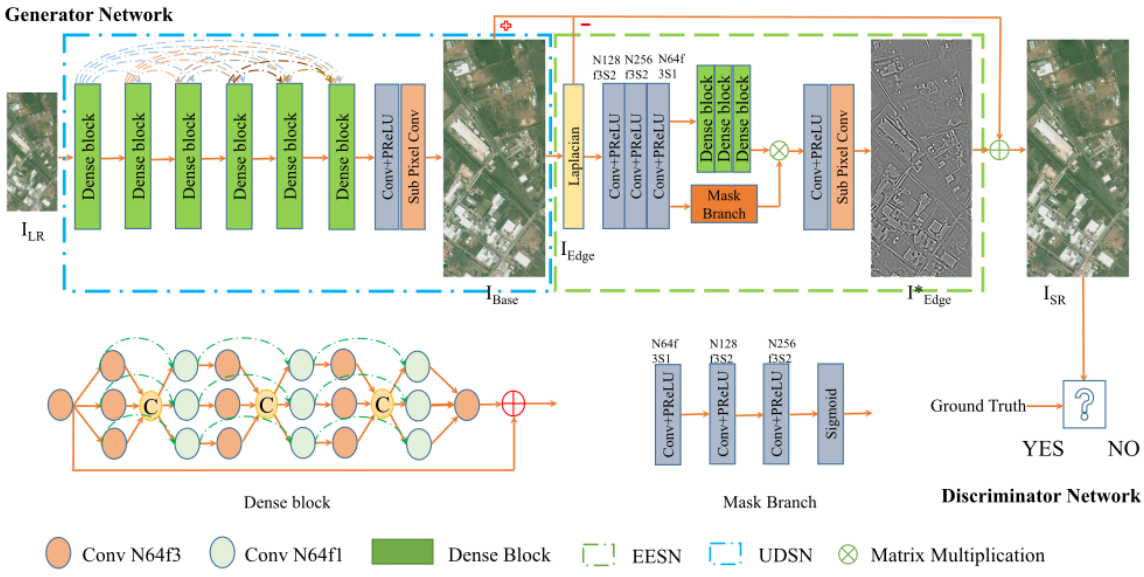


Figure 2.11: Schematics of EEGAN [20]

One may argue that TGAN and EEGAN have been tested on a limited dataset, making the results presented in the respective articles less reliable. However, a comparison between the two networks and other state-of-the-art methods exist in ref. [40] and they show to be taken as reference frameworks for remote sensing SR. On the other hand, CGAN is too recent to have been reviewed but the robustness

of the neural network is somehow proved in the article by the utilization of multiple datasets.

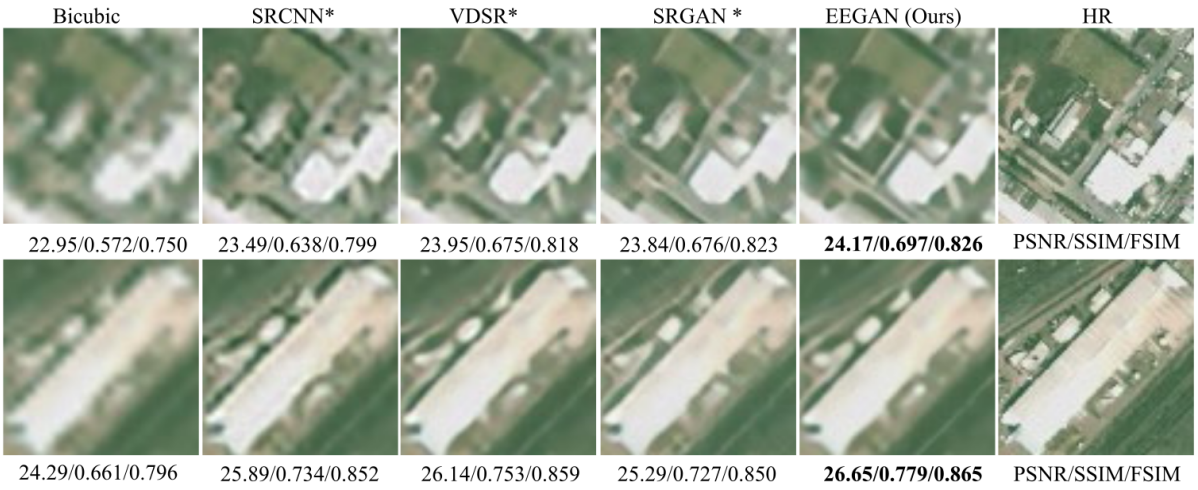


Figure 2.12: Comparison of EEGAN and other networks, building edge rendering detail [20]

- Hybrid approaches
  Combination CNNs and more traditional approaches (sec. 2.2.3) also showed to be a valuable solution especially in computer vision. It is an hybrid technique one recent work of Google for SR [41].

Datasets

Datasets represents a very critical element when it comes to remote sensing learning based methods, especially for what concerns super resolution. Indeed, it is well known that large and representative datasets are essential for the success of deep neural networks. In overhead imagery, it is not as easy to collect a sufficiently large amount of images well suited for a network. While in object detection/landcover the most relevant problem is the labeling of the photos, in super resolution there's an issue is even more complex to overcome. In fact, in digital imaging, single image super resolution means finding the version of an image as if it was captured in higher resolution. Thus, we should have two images of the very same scene from the very same angle (in the very same conditions !). Even if we took a picture of a scene and then we zoomed, they may suffer of misalignment making it inconvenient to evaluate the performance of trained models on such a dataset [4]. Considering the dynamics of a aerial or space frame acquisition where the capturing device moves at high speed, it's unrealistic to be able to capture a perfect LR-HR pair.

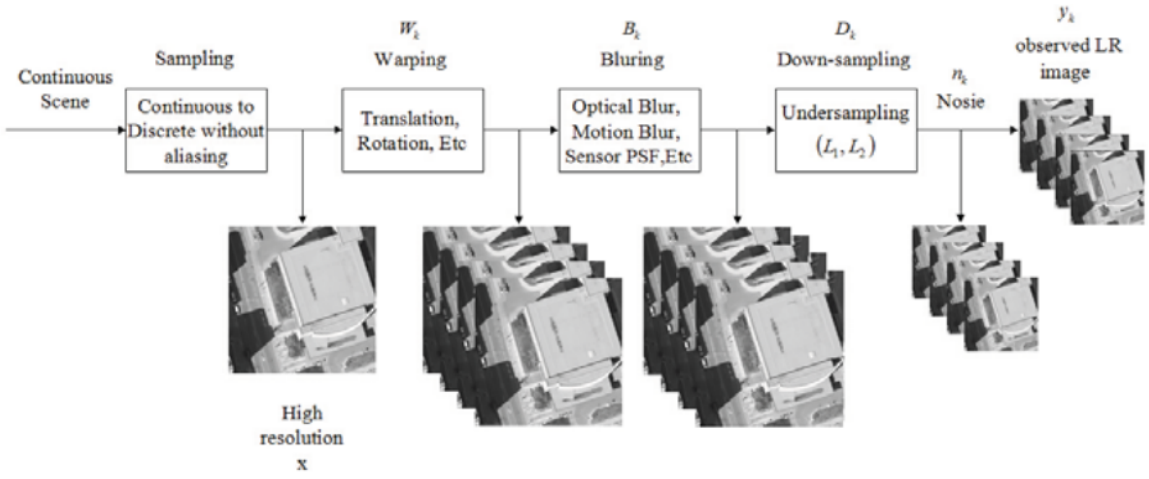


Figure 2.13: Representation of a sensor model taken from Ref. [49]

In practice, authors of remote sensing SR usually utilize sets of HR data as ground truth, and degrade them in order to simulate a corresponding LR set, then train the model to find back the HR images. However, the realistic degradation of an image passes through the definition of an imaging model (Eq. 2.2.1). The latter might also be quite complicated (Fig. 2.13) in the case of remote sensing imagery, where some factors like the translation of the instrument and the atmospheric distortions are added to other natural images contributions. A solution which is often employed in literature is to apply a bicubic degradation to simulate the blur and then to downsample the data by the desired factor. Such a procedure enables generation of training data without further complications, but it is far from a real simulation of remotely sensed data. Typically the lack of noise addition and therefore, noise influence in single image SR reconstruction is rarely assessed. Moreover, one can argue that a mapping solely composed by a bicubic blurring plus downsampling is too simple to justify the employment of such complex models as CNNs, and that a network capable of super resolving images obtained in this procedure might fail in the case of more composite degradations. Nonetheless,the use of Deep Learning in single image SR for remote sensing despite the lack of proper data seems to be justified. For example, in [48], a realistic imaging model is adopted to generate the LR data and yet a deep convolutional network pre-trained on natural images is a able to super resolve remote sensing frames for independently from the geographical context.

Hence, the problem of creating a LR/HR dataset is in turn transformed in collecting a representative set of HR (at the desired ground sampling distance) views. This is again not trivial in overhead imagery. In fact, single image SR in remote sensing is mainly interesting because HR or VHR images are often not available (at least, not free), whereas

LR open data are abundant [10]. This causes issues in learning based SR, where a large amount of varied HR ground truth data are needed to effectively train a model.

Recently, the growing demand for object detection aerial/space systems, a task which has high requirements in terms of GSD, pushed the creation of larger HR datasets, that in turn are borrowed by SR authors for the training of their network. Among the most used datasets, we can name UC-Merced, RSCNN7, AID [11], Spacenet and Spacenet 4 [40].

Dataset	Resolution[m]	No. images	Size [GB]	Off-nadir	Remarks
UC Merced	0.3	2100	0.317	No	
RSCNN7	Google Earth[1]	2800	0.377	Yes	
AID	0.5 - 8	10000	0.740	No	
Spacenet	0.31	rasters	3.4	No	
Spacenet 4	0.5 - 1.5	rasters	186	Yes	US cities
INRIA	0.3	rasters	11.6	No	US & EU cities
ISPRS (Vaihingen)	0.08	rasters	17.1	Yes	IR-R-G
ISPRS (Toronto)	0.15	rasters	5.5	Yes	

Table 2.1: Main datasets used in single image super resolution for remote sensing HR/VHR data

Among all, SpaceNet4 seems to be a good candidate in the context of CO3D mission because it features an important off-nadir range of angles, which is indeed the purpose of its creation [47].

Transfer Learning

Due to the lack of very large datasets, it is a very common practice to pre-train a network on natural image data (where bigger databases are common) and then use remote sensing ones to perform fine-tuning [33]. As justified in sec. 2.2.2, there's no reason to believe that some features of an upsampling and deblurring mapping can't be common between natural and remote sensing images. Even though today larger HR datasets are available, transfer learning can be used to speed up the convergence of any network. Indeed, it is a common practice in remote sensing SR to use systematically transfer learning. In some cases [3] even fine tuning on remotely sensed data is not done.

2.2.5. Single image super resolution based DEM generation

Some attempts of the employment of single image SR for improvement of DEM are proposed in the recent literature. In fact, a more common approach to improve DSM quality is to super resolve or refine a stereo DSM, often by means of DNNs in the recent works. A problem of such an approach is the lack of high resolution DSMs, that are challenging to obtain.

Here, we will stick to image super resolution, i.e. improve stereo pair resolution in order to get better 3D models at the end of the pipeline. In Zhang et al (2019) [56] the feasibility of such a methodology is explored. Several SR models are compared as well as a basic bicubic interpolation for image SR. Furthermore, directly upscale DSMs are added to the analysis, where LR DSMs are firstly generated with the low resolution stereo pairs and then interpolated through a coarse-to-fine pyramidal approach. The architectures considered are: the basic SRCNN [8], its deeper version VDSR and the super resolution network for multiple degradations (SRMD). The latter takes into account more complex imaging models by treating both LR images and degradation maps. Hence, it might be more robust to real world images that suffer from noise (suitable for satellite imagery) . The training is performed on 300 HR and VHR WorldView GeoEye images.

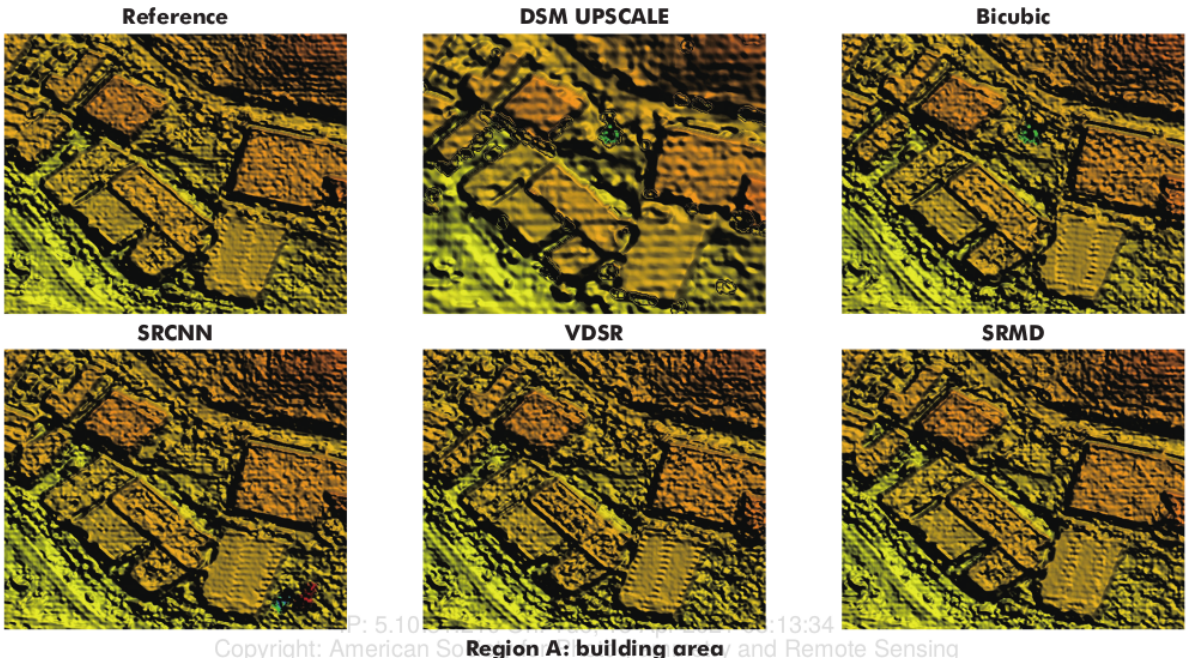


Figure 2.14: Comparison of different SR methods for DSM improvement [56]

Results show an actual enhancement of DSM models thanks to CNN based SR compared to direct DSM upscale and bicubic upscale SR. In building regions the proposed method

seems to be satisfactory while in difficult terrains with vegetation DSM are improved but the error remains high. On real images data, neural networks do not achieve as good results as on the simulated data, but it might also be due at the difference of information in the panchromatic images which are taken as reference.

These results are encouraging but several aspects can be improved. First of all, the network architecture can be changed to obtain a better SR. GANs and other kind of architectures have proven to outperform the models considered in the paper. Additionally, the objective function used is RMSE, that it is not very well suited for single image SR and it doesn't introduce any prior knowledge. Again, the GAN framework can be useful thanks to its adversarial loss. Secondly, the reference in all experiments is better to be replaced by ground truth because DSM reconstruction, especially in some kind of terrains, still suffers from important errors. Finally, a more complex degradation process of images might be taken into account to make the technique more robust with respect to real data. Besides, it is noteworthy that RPC files also need to be regenerated once the image is super resolved (the authors propose a simple procedure).

The same approach has been also proposed in aerial imagery, which is likely more coherent within CO3D where VHR images will be already available and the objective is to get closer to aerial data. Burdziakowski et al (2020) [3] assess improving of UAV photogrammetric products by means of image SR. They enhance data taken at 110 meters in altitude by a factor of two and compare the output dense clouds with the ones originated by original 110 m and 55 m (half the altitude) frames. This means a GSD of, respectively, about 4 and 2 cm. The network employed is a SRGAN, pre-trained on ImageNet and fine-tuned on NITRE. This means that no remote sensing dataset is used for SR training and yet the model is able to improve no-reference image quality score in comparison with the bipolar interpolation. As expected, the density of the point clouds is increased hence its resolution and its capacity to solve finer details. However, photogrammetric accuracy is degraded when compared to the not super resolved input. By a comparison with 55 m data, it turns out that lower ground sampling distance results in artifacts, deficiencies in the structure that in turn cause reconstruction errors. Thus, the reduction in precision should be due to some issues of the used softwares with views at lower altitude.

In Pashaei et al (2020) [38] an the contribution of an ESRGAN (enhanced super resolution GAN, [46]) to dense scene reconstruction for UAS (unmanned aerial systems) photogrammetry is evaluated. Findings of such a work concern: the possibility of improving results with a smaller dataset, the investigation of noise removal capability of SR GANs, the employment of a task-based image quality metrics (IQM).

With the respect to data, experiments are carried out on a limited number of original HR and virtually-generated LR UAS images by downsampling the original HR images using a bicubic kernel with a factor 4 and the addition of white gaussian noise. Unlike many other works, patches of resolution 1000 x 1000 pixels are used for fine tuning, assuming that larger patches can feed the network with higher scale feature information. The flight altitude was designed to achieve a GSD of 2.5 cm. The network was pre-trained on very large computer vision datasets and only fine tuned on the UAS images. Results show that this is a valuable approach and fine-tuning significantly increases model precision. The measures are task-oriented: camera parameters (interior and exterior) estimation by Structure-from-Motion (Sfm) photogrammetry and DSM reconstruction. We'll focus only on the latter: the ESRGAN model performs much better over man-made objects and natural objects with definite boundaries than other targets. For example, DSM generation fails in case of natural and man-made water bodies with lack of texture and along the suffers from offsets in correspondence of edges of tall natural and man-made structures.

2.2.6. Final observations

SISR literature is thriving in the recent years. Results shown by deep learning methods are impressive, yet it looks that we're from a a reliable and large scale introduction in space data applications. As we saw in section 2.2, in order to train a SR network we need a target and an input datasets generated from the former. The design of the degradation model is crucial, as the network will de fact learn how to inverse it (Eq. 2.2.1). In literature (Par. 2.2.4), a bicubic downsampling is usally used to operate this transformation. Although when the focus is the performance of the model and a benchmark is needed to have a comparison with other works, bicubic downsampling is not representative of real remotely sensed data, and thus these results lack of realistic benchmarks, what reduces their reliability and doesn't exclude biases linked to the particular kernel used in the models learned by a network.

Generally speaking, several other architectures can be found in the literature, each one claiming to achieve better results in terms of quality metrics and/or computational performance. Additionally, the tendency in remote sensing is to extend the results of computer vision rather than proposing brand new approaches. Thus, we remark that numerous networks born for natural images have been directly utilized in remote sensing and their outcomes are still comparable to architectures born for aerospace imagery. Such a methodology may seem naive, nevertheless the close connection between the two fields of application is consolidated in the literature, and such a feature naturally remains when it comes

to DNNs, models whose purpose is to learn tasks.

Furthermore, in literature DNNs are often tested on new samples coming from the same database used for training. That is, images taken out from the training dataset, but still of the same kind of data, as much as the contents might be different. Yet, the underlying idea of these technologies is to be added to a ground segment pipeline, hence to super resolve images coming from a satellite and not ad hoc processed for a deep learning training. This makes less reliable the results proposed by the referenced bibliography. In this work we will attempt to super resolve real sensor data after a training on simulated ones.

Finally, we may identify the following factors that can influence a SR deep model:

- **Degradation model**: whether it is possible that after an appropriate training has generalized features which belong to every sensor model, or a given network is closely connected to the sensor considered and it's impossible to use it to infer other type of acquisitions.
- **Network tuning**: determine how much are deep learning intrinsic aspects relevant for such a super resolution problem. Loss, data augmentation, network hyperparameters etc.
- **Transfer learning**: from the literature it seems that transfer learning from computer vision large datasets is pretty much straightforward. For example, in [3] the network is not even fine tuned on a remote sensing dataset, yet is considered to have enough generalized to super resolve an aerial set of images.
- **Ground Sampling Distance**: We do not expect the model to necessarily be able to resolve at different scales. Nevertheless, if the learned features are more linked to signal processing and less to physical objects, we could perhaps a network capable of super resolving images across different scales.
- **Radiometry and geometry**: in deep learning, the tendency is always to create bigger and bigger datasets in order to allow the network to generalize. But utilizing data from different acquisition chains implies the contribution of different sensor models, which could lead the network to be more confusedt.

3 | Implementation of super resolution networks

Super resolving images image quality's sake is not the main objective of this work, as we rather want to study how this impacts an important space downstream application such us DSM production. However, in order to add a step to such a complex and optimized process as the CARS-Pandora pipeline (Fig. 1.3), we need to make sure that such a step is adequate and optimized as well. Indeed, there's no point in launching a large DSM production if the results of our preprocess are of the same or lower quality of the original inputs. In addition, the implementation of a SR method might not be straightforward when we want to take into account realistic image degradation (instead of bicubic downsampling). Therefore, much efforts served in order to accomplish reliable single image super resolution before running even once the stereo pipeline, and hence the strictly 2D part (Chap. 3) and the 3D one (Chap. 4) were almost treated separately.

Some objectives were set before designing the SR pipeline, necessary from the very beginning in order to choose models and data, and driven by the company interest around certain types of data and their availability. The proposed SR method should provide reliable results:

1. VHR images: in order to do DSMs, we need fine details to be rendered on the 2D image, hence only HR or VHR data are suitable. Moreover, post processing is already applied to such images in order to improve resolution (e.g. PAN-sharpening). A VHR PAN image is a sort of benchmark, because it's what we usually do today. In the context of this project the idea is to overcome this technological limit.

2. RGB images: Since SR learning methods are supposed to extract low-high frequencies relations from samples, the underlying idea is to feed them with as much valuable information as possible. Working on single bands appears then as limiting the amount of knowledge that can be derived within a training. Perhaps from one band the network can infer some geometrical information that in turn helps it in super resolving in another bands. Moreover, in literature RGB data represent

predominant choice 2.2.

3. In urban areas: this is where we have interest in improving DSMs. To implement a method capable of super resolving images in any environment would be as finding the holy grail. For the moment, we can be satisfied if the networks are successfull on french cities, where the greatest amount of data is available and that will be among the first targets in CO3D mission [25].

Among VHR choices, Pléiades images represent definitively the priority given the large amount of available data for a collaborator of CNES such as CS Group. Furthermore, typically we expect CO3D images to share some of their characteristics with Pléiades ones. VHR data (as for Pléiades) often are in 4 multispectral bands plus panchromatic one. Since we also want to work in RGB, we will be oriented to data that are or resemble Pléiades PAN-sharpened images. We recall that the latter have a GSD of around 50 cm. There's no specific reason in ignoring NIR band, but it will be discarded from training for sake of simplicity and to work with a format suitable for most datastes.

3.1. Methodology

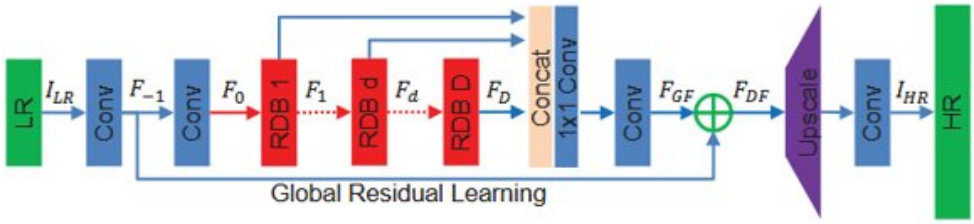
One of the first design choices regarded the model, i.e. the actual neural network to use. Among all the possibilities presented in the chapter 2, we chose ESRGAN because it is somehow a safe choice in GAN domain: it overcomes some limitations of the prototypical SRGAN [27] but it avoids to add "too innovative" (hence, not extensively tested) modules, like the Laplacian module of EEGAN [20]. However, GANs are notoriously more difficult to train and more prone to artifact generation. This led to the decision of keeping a second non-GAN network. It has to be remarked that this choice was undertaken after an extensive literature review as well some trials with different networks. In addition, it's interesting to study the contribution of discriminator to SR and its applications, since it introduces remarkable differences during a training. That's why the second network was chosen to be as close as possible to ESRGAN generator. It is the case of the Residual Dense Network (RDN) [12]. The only remarkable difference between it and ESRGAN's generator is the main block, Residual Dense for the former and Residual in Residual Dense Block for the latter, as illustrated in figures 3.1 and 3.2. Indeed, remarkable differences can be identified between the two cases, as it will be shown in chapter4.

RDN [12]

The Residual Dense Network [12] combines the features of residual and dense net-

works. In practice, connections are skipped (residual) and outputs of a layer are passed to multiple successive layers (dense). This allows on the one hand the learning of the residuals, i.e. the difference between inputs and ground truth (high frequencies) [51]. On the other, to learn features a different scales avoiding the vanishing gradient problem, and speeding up the training process [51]. The residual architecture has its origin in the well known *ResNet*. Any loss could be used, most of them relying on an error measure between prediction and ground truth. For this project implementation a L_1 loss was used (Eq. 3.1), which quantifies the error between the elements $n \in 1, ..., N$ of the predicted array y_{pred} and the ground truth one y_{true}.

$$L_1 = \frac{1}{N} \sum_{n=1}^{N} |y_{true} - y_{pred}| \tag{3.1}$$



Global architecture

Residual Dense Block

Figure 3.1: Illustration of Residual Dense Network (RDN)

ESRGAN [46]

Such a network is an evolution of the prototypical GAN for SR [27] and indeed it stands for Enhanced Super Resolution GAN. It proposes a generator composed by Residual in Residual Dense Blocks (RRDB), thus not very much different from RDN (Residual Dense Block, RDB). Batch normalization layers are removed. The discriminator is of probabilistic kind: it returns the probability of an image to be true or synthetic. The loss 3.2 is composed by three contributions: a perceptual loss L_{percep} that compares the errors on features extracted by a VGG generator (taken pre trained on DIV2K for this study), an adversarial loss L_G^{Ra} which is the discriminator

response, and a content loss L_1, that consists of the mere pixel difference between prediction and target, added in order to reinforce attachment to the data. λ and η are two coefficients set to harmonize the order of magnitude between losses that do not measure the same features.

$$L_G = L_{percep} + \lambda L_G^{Ra} + \eta L_1 \tag{3.2}$$

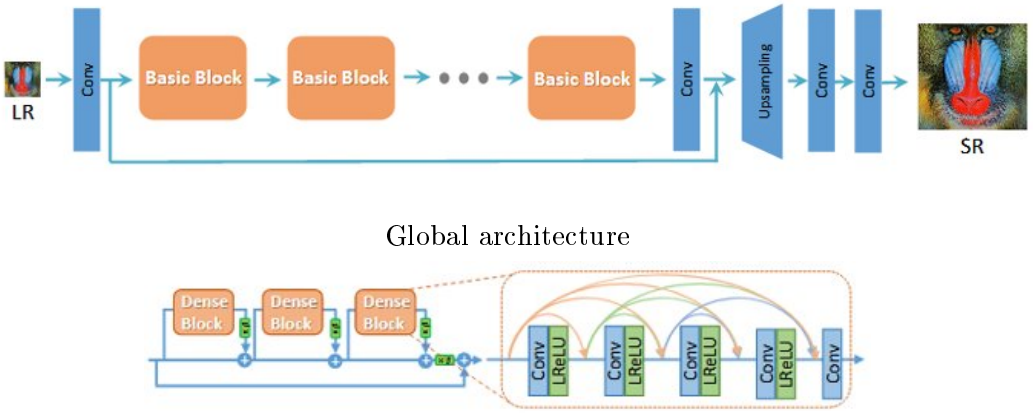


Figure 3.2: Illustration of Enhanced Super Resolution Generative Adversarial Network (RDN)

The experimental framework used for this phase of the project was largely inspired by the SR literature features, in particular the taxonomies proposed in [1] [40]. It consists of training multiple networks (two in our case) with a HR-LR train dataset where LR dataset is generated by imposing degradation and downsampling to the HR set. A test set is obtained in the same way and serve to compare network performances by means of a coherent metrics as well as visual inspection. The details on the metrics considered for this work can be found in 3.4.2. On the top of that, a counterexample is usually included as a support of the proposed methods. Bicubic upsampling is the usal choice in this regard and this work will be no different. This is likely because it offers the best deal between performances in term of 2D metrics and cost of execution

Since it is difficult to predict the effects of SR single image on DSM production, the implemented framework allows flexibility and modularity: the objective is to test the behavior of different SR deep networks in terms of architecture, databases as well as training methodology.

The enormous amount of open source deep learning software and the increasing appealing that SR has in the last years, makes it easy to find open source software and support. The developed codes were adapted from MHAN Github repository [54] and PyTorch ESRGAN implementation. The deep learning framework utilised is *Pytorch* which allows medium level operations and because it is slightly easier to find SR online code and support in *Pytorch* with respect to *Tensorflow*.

The experiments run on a a GPU node reserved with 1 GPU NVIDIA Tesla T4, 4 CPUs Skylake 2.2GHz 92 Gb RAM, courtesy of CNES who allowed the utilization of the HAL cluster facility.

The approach essentially consisted of modifying training algorithm and data handling, in order to make it compliant with the used type of data and desired metrics/training settings, while less attention was given to DNN architectures, assuming that these models are robust enough and can be transferred to any application without influencing the results.

As remarked in section 2.2.2, remote sensing data are more difficult to handle then the the digital images we're used to deal with. In particular three main challenges had to be tackled: data size (acquisitions larger than 10 MP), radiometric values (continuous measure of luminosity and not encoded in 8 bits), data format (georeferenced GTiff format). Some of the datasets widely used in literature (RSSCN7, UC-Merced, AID, etc.) consists of various source images preprocessed in order to have a format that resembles everyday pictures (8-bit, less than 1 MP, jpg or png format). Indeed, very often the open source implementations of SR networks are note capable of handling satellite images. For this project, though, the choice was to develop a methodology to handle remotely sensed acquisitions rather than impose a preprocessing step. This was due to the intention of conserving original radiometry and georeferencing through the networks for using super resolved images in CARS as we would use standard data.

A first constraint to be taken into account is the limited amount of memory when operating with GPUs. When using very large DNN models with milions of coefficients the available memory may run out very soon. Even an image of normal size, say 512x512 pixels, can turn out to be too large to be fed into a SR network. What is usually done is to crop the image into a smaller *patch*. Typical sizes for patches ranges from 48 to 128 pixels. This might sound odd as it is in contradiction with that we want the network to learn spatial features that can range from few pixels to some hundreds. Nonetheless it is coherent with what one can find in other deep learning applications such image classification. In

addition, it turns out that files larger than 1 MP are too big to even be loaded in a such deep learning routine. The strategy is then to partition an input large image in square tiles of medium size (typical values for tile size are 256, 512 pixels), and from each tile operate a random crop of patch size. This way we make sure that within a reasonable number of epochs, the network will have seen most of one image's extent in the form of random details. Most of all, it will process the data in the same way independently on the actual size of the available acquisitions.

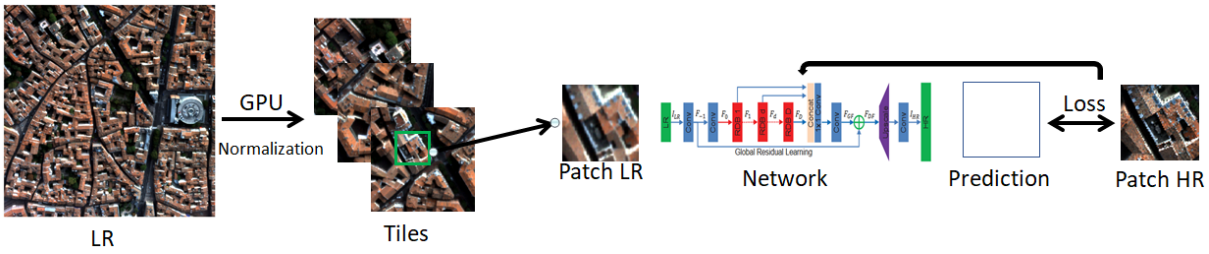


Figure 3.3: Illustration of the training methodology

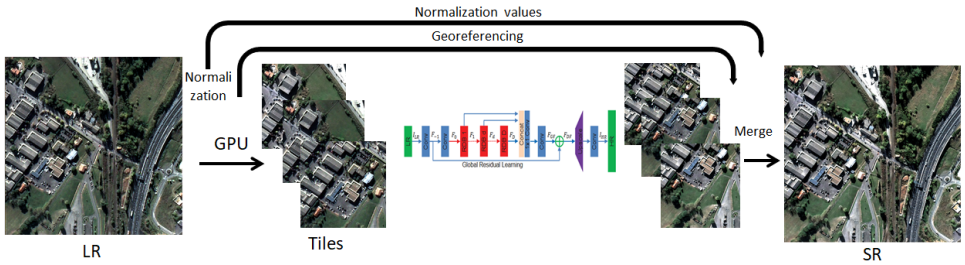
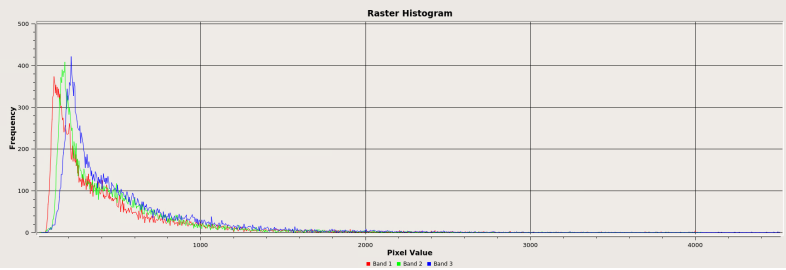


Figure 3.4: Illustration of the inference methodology

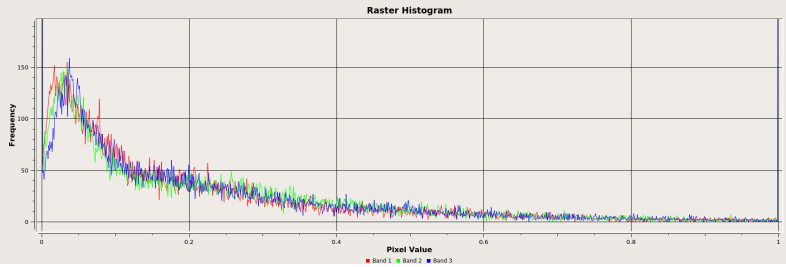
In the inference pipeline, where we take a satellite image and ask the network to super resolve it, tile partition is performed as well but we don't to further extract patches because of the absence of backpropagation and the fact that, unlike the training phase, we don't need batches of data. On the other hand, we would like to get back the acquisition at the original extent and not a set of tiles, and possibly that can be superimposed on a reference frame to the original image for visual comparison. This is achieved with the following steps. We store the geotransform of each tile (coordinates of upper left corner and pixel size in meters), update it by dividing the pixel size by the scale factor, and reassign back this new transform to the SR sample. Then, the georeferenced tiles are reassembled together by means of a merge algorithm.

Finally, data in training and inference phase are subject to some radiometric manipulations. Two conflicting evidences have driven this step. On the one hand, it is not desirable to modify data radiometry as it correspond to physical measures of object radiation. In

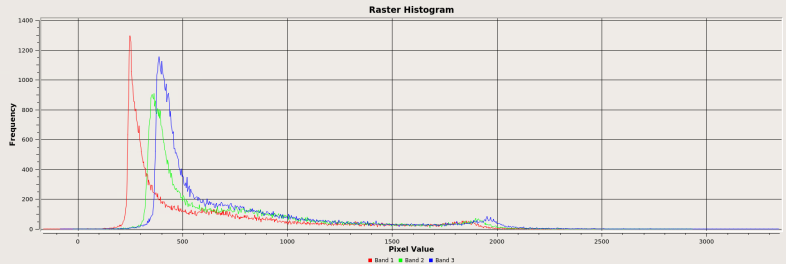
remote sensing, these values are used to physically measure Earth's surface properties. Therefore, if modified, the data become less reliable from a scientific point of view. On the other hand, in deep learning data are typically normalized and it is well known that it's beneficial for training when data are well distributed in the interval $[0, 1]$. Now, looking at the histogram of a typical Pléiades image, this is not the case (Fig. 3.5). The proposed compromise is a band-wise "stretch" of the data histogram.



Original data histogram



Histogram after normalization before inference



Histogram denormalized after inference

Figure 3.5: Stretch and unstretch operations

In practice, before the passage through the network, the 2nd and 98th percentile are removed and the remaining values are linearly brought back to the $[0, 1]$ interval. In inference mode, the same is done just before forwarding an image in the DNN, and the limit values (corresponding to 2nd and 98th percentile) are stored in order to reverse the stretch once inference is performed. This way, the output pixel values will be compatible with the input data.

3.2. Dataset generation

In data science, data can be very often more important than models. Data shape what is learned by the algorithm and we can teach different tasks to the very same model by feeding it with diverse nature data. Usually, we want a neural network to *generalize*, i.e. to be able to repeat a task learned during a training with new, challenging, data. In the context of this project, the objective is to super resolve 50 cm GSD RGB images of, at least, Pléiades type. It is a quiet specific task, yet challenging because during training the model doesn't receive any actual Pléiades or even space data, but only simulated ones.

In paragraph 2.2.6 we justified the choice of looking for realistic data degradations of our study. For this scope, the CNES kindly allow the utilizatin of its *Chaîne Simulation Image* (CSI), illustrated in paragraph 3.2.1. The CSI can apply any step of a satellite acquisition chain as well as an arbitrary sampling. Thus it is the perfect tool for generating our dataset and it was used for both HR and LR datasets. Furthermore, the CNES made available a set of PELICAN aerial images that are described in paragraph 3.2.2, as well as a configuration file for simulating Pléiades images from PELICAN data.

3.2.1. CNES' image simulation chain

The *Chaîne Simulation Image* is an internal CNES tool that allows the simulation of an entire satellite acquisition pipeline. A source image is processed according to a set of modules, that are defined by the user in a configuration file. Any step can be removed and modified in order to assess its effects on the simulated image. The working principle can be roughly explained thorough the following process. The input image is loaded in the memory, a radiometric correction is added if needed, and the luminance values of the image are converted into reflectance. Then the imaging system is simulated, taking into account, at least, the modulation transfer function (MTF) and the sensor noise, and the product is resampled at the desired resolution. Afterwards, other steps associated to a particular pipeline can be simulated as well, like, for example, compression and denoising. This is done band-wise, while a fusion operation can also be applied to get a RGB output. The source image resolution fixes the maximum resolution achievable. Ideally, it should be at least two times higher than the simulated resolution, in order to be able to neglect the MTF of the instrument that took the source acquisition. The sensor model can be precisely defined through numerous parameters, e.g. MTF frequencies, maximum luminance for each band, etc.

Two configurations must be prepared, one for LR and the other for HR dataset. The LR dataset should simulate Pléiades images at 50 cm GSD. Therefore, for a scale factor of 2, the HR dataset is set to 25 cm GSD. This ensures that the MTF requirement is largely satisfied, since the source BD Merou data are at 10 cm. For the HR dataset, a perfect sensor is assumed. Loosely speaking, no degradation are applied to the inputs. This is because we want the network to learn super resolution at the highest quality achievable. Hence, just a dezoom at a factor of 2.5 and a quantification in 12 bits are applied. The quantification is the operation of writing pixel values in memory into a defined number of bits.

Regarding the LR dataset, the set up was more complex in order to be representative of real Pléiades data. A complete illustration of what happens inside such simulations would require an additional chapter and it's out of the scopes of this report. Here we present a brief overview of the main manipulations that doesn't mean to be a complete explaination of how the models used by the CSI for the dataset generation.

1. **Resampling**: or downsampling in this case. It consists of sampling the image on a lower scale grid, coarser by a scale factor. In order to respect Shannon criterion, a low pass filter is usually applied. This potentially causes a modification of the Modulation Transfer Function of the image. The MTF is defined as the module of the transfer function of an imaging system. i.e. that characterizes the system behavior in the frequency domain. It is a quiet central notion when modeling acquisition systems and it can be associated to the resolving power of an instrument. Many different physical phenomena contribute to the definition of this function, from atmosphere effects on light to the speed of a satellite with respect to the scene. The most important one is the contribution of the instrument diffraction. In our case, the global MTF is approximated with a diffraction (Eq. 3.3).

$$H(f_x, f_y) = \frac{2}{\pi}[arccos(\frac{f}{f_c}) - (\frac{f}{f_c})\sqrt{1 - (\frac{f}{f_c})^2}] \tag{3.3}$$

   f_c is the cut-off frequency of the instrument.

   In the CSI, the source image MTF can be defined. While resampling, we apply the point spread function (PSF) of the simulated instrument and update the MTF of the source image

2. **Radiometric simulation**: at this stage it is realized the conversion between the input luminance and the numerical values read by the detector by adding radiometric

noise, accordingly to the specified signal to noise ratio (SNR) as well as the noise potentially present in the source image.

3. **Quantification**: described a bit earlier in the paragraph.

4. **Denoising**: a Non Linear (NL) Bayes method is applied in order to remove noise from the resulting image. In a nutshell, NL-Bayes is an improved variant of NL-means. In the NL-means algorithm, each patch is replaced by a weighted mean of the most similar patches present in a neighborhood [26]. As most denoising methods, it relies on the assumption of additive white Gaussian noise of constant standard deviation. However, this is not the case of optical whose noise is usually modeled as the contribution of a constant plus a Poisson terms, i.e. the standard deviation can be written as $\sigma_{noise}(x, y) = \sqrt{A + B \cdot s(x, y)}$, $s(x, y)$ being the signal. Therefore, an Anscombe transform is beforehand applied. In loose words, it's a variance-stabilizing transformation that transforms a random variable with a Poisson distribution into one with an approximately standard Gaussian distribution. After denoising, an inverse Anscombe transform returns the image in its natural radiometry.

5. **Deconvolution**: when the sampling of an image is correctly applied, the MTF of the optical instrument tends to attenuate high frequencies and hence the raw image is characterized by blur and needs a contrast enhancement treatment which takes the name of deconvolution. It consists of multiplying the blurred image spectrum by a function that depends on the target MTF. In order to limit artifacts due to high frequency oscillations when we convolute a MTF with a finite support, target MTF can be obtained by the convolution of a rectangular and a gaussian functions [30]. The definition of the target transfer function, together with the model of the instrument MTF, allow us to determine deconvolution filter in the case of Pléiades images.

3.2.2. BD Merou

PELICAN is the French acronym for "*Plateforme Et Logiciels Informatiques de Cameras Aeroportees Numeriques*". The system, developed in 2003 as an IGN-ONERA-CNES cooperation project consists of a set of four optical heads each equipped with a 4096 x 4096 pixels CCD detector working in the visible and near IR spectral range (from 400 nm to 950 nm). This sensor provides synchronous images achieved with a TDI-like control of the CCD to avoid smearing, thanks to GPS data [6]. CNES provided a collection of acquisitions in different environments across France at 10 cm called BD Merou (french

acronym for "Base de Données", i.e. database, Merou). For the main runs, only the images containing buildings and hence in residential, semi-residential and industrial areas were selected. Examples of this images are shown in figure 3.6. The total size of the used HR and LR datasets are, respectively, 627 MB and 157 MB.

Figure 3.6: Samples from BD Merou

3.3. Hyperparameters fine-tuning

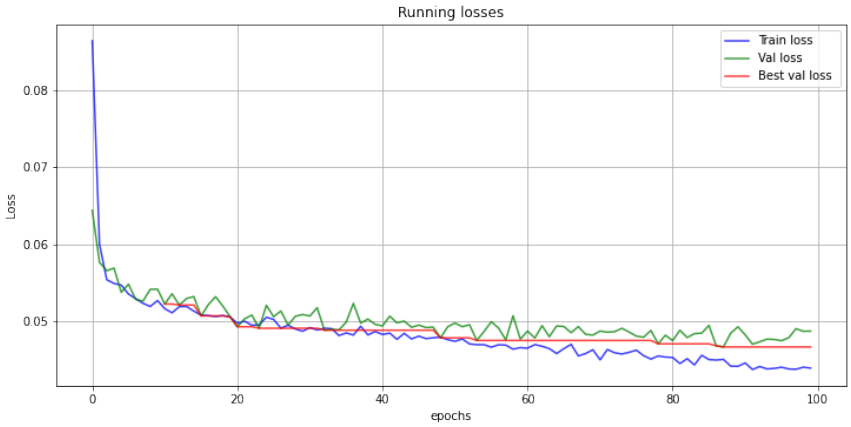
Once set up a training pipeline and prepared suited data for training, one further step is often useful. A neural network presents a bunch of hyperparameters that can be changed in order to well adapt the model functioning and/or find the optimal working for the given data. Sometimes a wrong configuration can lead to failure of the training. For this purpose, an ad-hoc tool was used. *RayTune* is a platform for fine tuning of large size models hyperparameters. It allows to spare lots of time in manual tailoring of the network sizing values. It proposes different modes, from basic grid search to complex multivariable optimization algorithms. During this project its usage was limited to grid and random searches. The search space is defined in table 3.1. Grid search consists of trying all the combinations defined in the search space. For this mode continuous intervals, as for batch size and learning, were arbitrarily discretized. In a random search, a random combination of values taken from continuous or discretized interval is tested. In both cases, a trial is performed for a certain number of epchs that won't be bigger than a user defined maximum of epochs, and it's recorded. Afterwards, the configuration that provides the best performance can be chosen. The great advantage of this way of using RayTune is the early stopping of the training, i.e. training is stopped if the algorithm understands that the current configuration doesn't converge or converges worse than the best configuration.

The search space values were defined after some considerations of network nature and constraint. It is a rule of thumb that larger batches enhance training, so a minimum of 10 batches is set. On the other hand, a limited amount of memory is available in GPU, we cannot account for more than 40 batches for the models used. Patch size and tile size

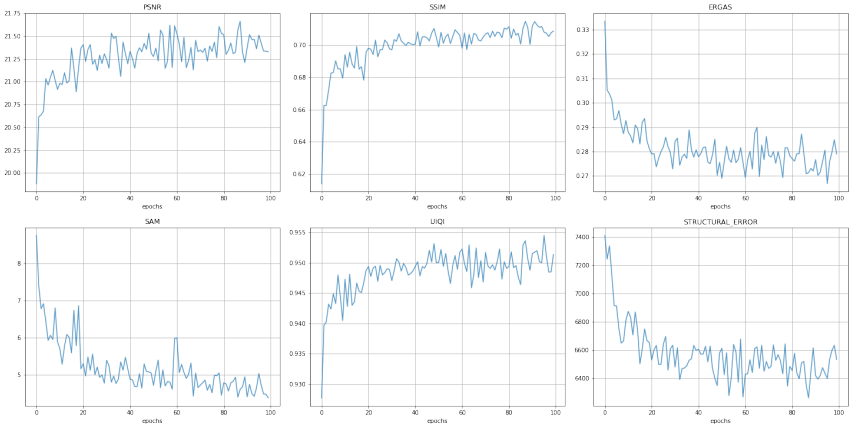
	Batch size	Patch size	Tile size	Learning Rate
Search space	$[10, 40]$	$48, 64, 96, 128, 160$	$256, 512, 768$	$[1e-4, 1e-2]$
Best config. RDN	30	96	512	1e-4
Best config. ESR-GAN	20	96	512	1.4e-4

Table 3.1: Search space definition for Ray Tune search of hypermaters combination

delineate the amount of information: larger patches mean larger scale details included in a passage through the network; the bigger the tile means the larger extent for taking the random crop, the less numerous the samples are available for one epoch. The learning impacts dramatically a network's training. If it's too small, convergence can be so slow to become impossible. If too large, convergence is achieved rapidly but it's more likely to get stuck on a local minimum.



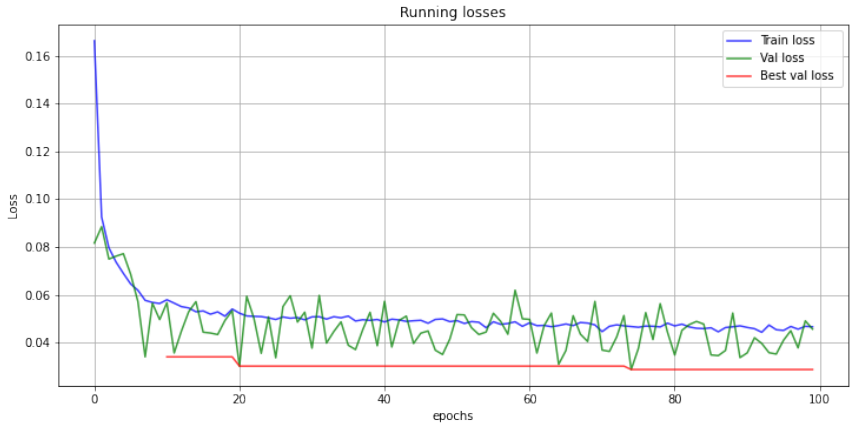
Loss as function of the epoch number



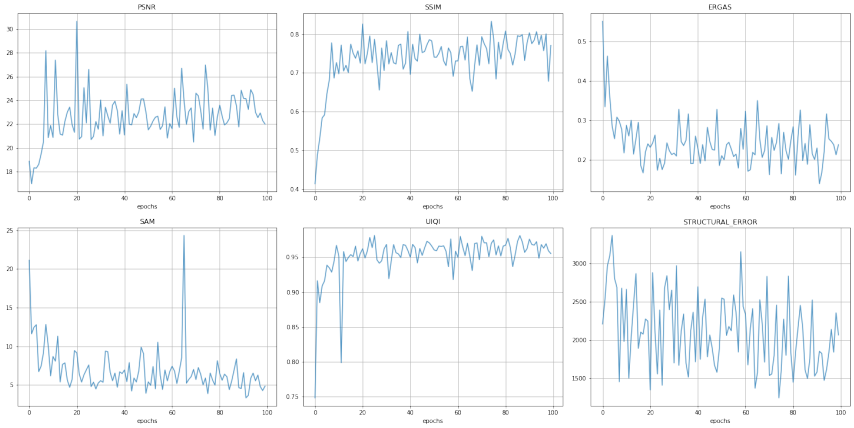
Metrics as function of the epoch number

Figure 3.7: Training figures for RDN before fine tuning network hyperparameters

A part from supplying a reliable combination of roughly optimized hyperparameters, such



Loss as function of the epoch number



Metrics as function of the epoch number

Figure 3.8: Training figures for RDN after fine tuning network hyperparameters

an extensive fine tuning gave insight on the impact that each parameter can have. For example, it was noticed that with smaller batches and patches, the network tends to converge more quickly but also is more prone to overfitting.

As we see from figures 3.7 3.8, a set of good parameters could be found. Nonetheless, the difference in performence is not very significant and thus such an extensive fine tuning is only partially justified, given the amount of time and resource needed to implement it. A complete overview of the metrics proposed is supplied in paragraph 3.4.2. The main conclusion we could infer is that models and associated hyperparameters are robust even for specific applications as the one of this study. The research is very active in this field and mainly focused on the optimization of the models themselves. Therefore open source implementation of this kind of neural networks are nowadays trustworthy and one should put more effort in improving the quality of the used data.

3.4. Inference

3.4.1. Importance of the degradation model

To justify further the use of CSI and of a complex sensor model, a sample of some intermediary results it is reported in figure 3.9. A Pléiades image (LR) is upsampled by a factor 2 using by 5 means: bicubic upsampling, RDN/ESRGAN trained on a LR dataset obtained via bicubic downsampling, RDN/ESRGAN trained on a LR dataset obtained via the application of a realistic degradation model by means of the CSI. We can state the networks struggle in really super resolving the image when the bicubic downsample is adopted, while the SR images infered from a CSI sensor model trained network are visually satisfying. It is noteworthy that ESRGAN seems to be more accurate than RDN in the bicubic case, confirming the claims found in the literature. For the realistic degradation model sample, though, the two networks are comparable, supporting the choice of taking into account a non GAN during the project.

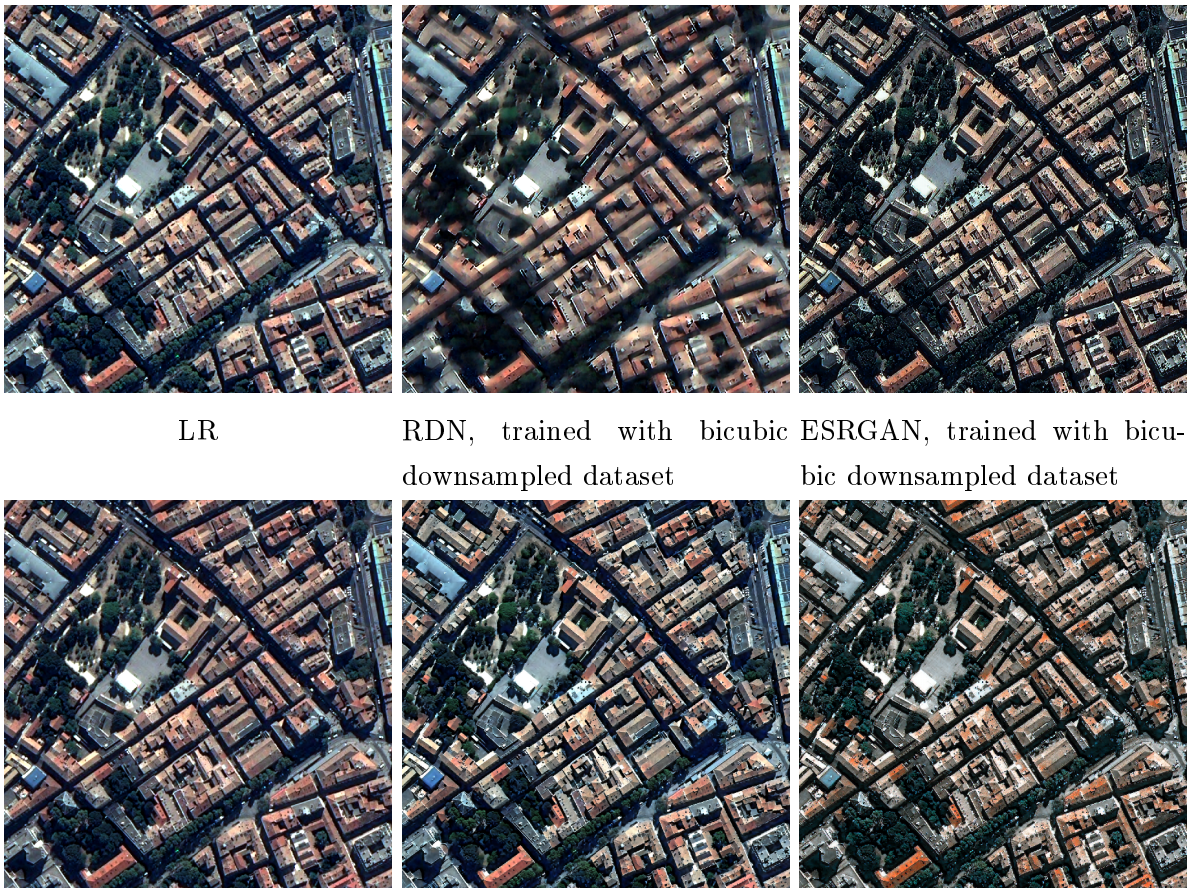
3.4.2. 2D metrics

We are always interested in having a quantitative measurement of the amount of degradation of an image to compare different methods. In super resolution literature, the Peak to Signal Noise Ration (PSNR) between prediction and ground truth is usually taken as primary metrics. Nevertheless, being based on the MSE and not on visual perception, it might not be always appropriate especially when we deal with synthetic images and deep neural networks. More sophisticated indicators exist and are easily found in reference literature. Here a brief overview of the metrics presented in this work, largely inspired by the work [19]. In the case of supe rresolution, metrics are defined as a measure of the difference between the tested, super resolved image I_2 and a reference, HR one I_1.

$$PSNR = 20 \cdot log_{10}(\frac{L^2}{RMSE}) \tag{3.4}$$

with $RMSE = \frac{1}{MN}\sum_{j=1}^{M}\sum_{i=1}^{N}(I_1(i,j) - I_2(i,j))^2$ and L the maximum value that can be attained by a pixel. It is the most commonly found metrics, yet it is not judged a reliable indicator anymore [19]. It's expressed in decibel and higher values indicate better quality.

$$SSIM = \frac{(2\mu_1\mu_2 + c_1)(2\sigma_1\sigma_2 + c_2)(cov_{12} + c_3)}{(\mu_1^2 + \mu_2^2 + c_1)(\sigma_1^2 + \sigma_2^2 + c_2)(\sigma_1\sigma_2 + c_3)} \tag{3.5}$$



LR | RDN, trained with bicubic downsampled dataset | ESRGAN, trained with bicubic downsampled dataset

Bicubic upsampling | RDN, trained with CSI downsampled dataset | ESRGAN, trained with CSI downsampled dataset

Figure 3.9: Pléiades image and SR sampled for different LR dataset generation mode

In (Eq. 3.5), μ, σ and cov are, respectively, the mean, standard deviation and covariance of two windows within the image. Multiple windows are utilized for the computation. $c_1 = (k_1 L)^2$, $c_1 = (k_1 L)^2$, $c_3 = \frac{c_2}{2}$, are three constants whose purpose is to stabilize the division if the denominator has a too low value. L is the dynamic range. SSIM values range from -1 to 1. The closer to one, the better the performance. This metrics has been elaborated to overcome PSNR, as it tells the local structure similarity (from here its name Structure SIMilarity index) between the two samples assuming that human perception is more sensitive to structures rather than the pixelwise differences.

$$ERGAS = 100\frac{dh}{dl}[\frac{1}{n}\sum_{i=l}^{n}\frac{RMSE^2}{\mu_1^2}]^{\frac{1}{2}} \tag{3.6}$$

The ERGAS (Eq. 3.6) index is also called Relative Dimensionless Global Error in Synthesis. $\frac{dh}{dl}$ is the ratio between pixel size of reference an input image, and n iterates on

the bands. A small ERGAS (close to 0) means good image quality.

$$SAM(V_1, V_2) = arccos(\frac{< V_1, V_2 >}{||V_1||_2 \cdot ||V_2||_2}) \tag{3.7}$$

The complete name of SAM (Eq. 3.7) is Spactral Angle Mapper. It computes the spectral angle between the pixel, vector of the reference image and prediction. It is worked out in either degrees or radians. It is performed on a pixel-by-pixel base. A value of SAM equal to zero denotes the absence of spectral distortion.

$$UQI = \frac{4\sigma_{21}(\mu_2 + \mu_1)}{(\sigma_2^2 + \sigma_2^2)(\mu_2^2 + \mu_2^2)} \tag{3.8}$$

UIQI (Eq. 3.8) is an acronym for Universal Quality Index. It accounts for the amount of transformation of relevant data from reference to tested image. As for SSIM, it ranges from -1 to 1, with 1 the best possible value.

$$SE = |(I_1 - I_2) \star sobel_x| + |(I_1 - I_2) \star sobel_y| \tag{3.9}$$

SE (Eq. 3.9) stands for Structural Error and it relies on the the convolution with Sobel x and y kernels, thus telling how the input image well represents contours.

3.4.3. Results

Finally, trainings were run with sufficient confidence on data, model and configurations (Tab. 3.1). Each training was run for an arbitrarily large number of epochs (more than 500), and only the best configurations are kept; more precisely, whenever the loss of the model on a validation test is smaller than the previous best loss, the new value became the reference and the weights of the running epoch are saved. The results shown are obtained using the weights of the epoch registering the best validation loss. Moreover, a test set of BD Merou acquisitions was kept out of the training in order test the network on new data. The metrics adopted are comprehensive of an image's quality and are described in section 3.4.2. Results will be shown for scale factors of 2 and 4, for test set from the BD Merou test set, and Pléiades high reslution images from the cities of Toulouse and Montpellier. As already underlined inparagraph 3.1, since the objective is to super resolve Pléiades images at 50 cm GSD, this is the value fixed for LR. The HR GSD, and so the SR's, depends then on the chosen scale factor: 25 cm for a zoom 2, 12.5 cm for a zoom 4.

Inference on BD Merou test set

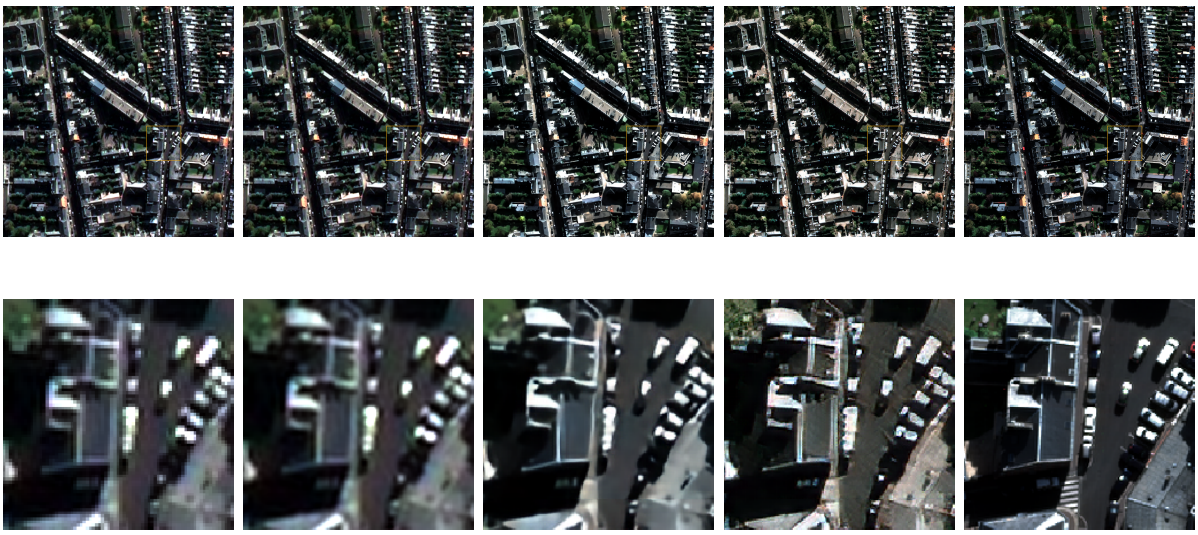


Figure 3.10: Test image from BD Merou, scale factor 2. A smaller crop taken in the orange square is zoomed.

Table 3.2: Scale factor 2

	PSNR [dB]	SSIM	ERGAS	SAM	UQI	Struc. err.
Bicubic	18.54	0.4705	0.6810	16.372	0.8471	80954
RDN	23.88	0.7753	0.3695	7.874	0.9562	45542
ESRGAN	20.97	0.6334	0.5151	11.946	0.9165	66457

Table 3.3: Reference metrics for BD Merou test set, scale factor 2

Tab. 3.3 and 3.4, quantitatively confirm that both trainings were succesfull as the networks are superior to bicubic in every measure. The only anomaly is represented by the SAM of the ESRGAN images, which is higher than the bicubic value. This means that a strong spectral distortion is intruduced by the GAN, something that is expected from this network, that is prone to introduce high frequencies even where there's no need for enhancing an image (Fig. 3.14) 5.4 are examples of ESRGAN's artifacts). We remark that comparing RDN to ESRGAN in terms of PSNR (dependant on the pixel error vis-à-vis the reference) is pretty unfair because the loss of the former relies solely on the pixel errors while for the latter other contributions are taken into account. That's why SSIM other metrics were considered. Nonetheless, RDN shows better performance also with respect to all the other metrics in both zoom 2 and zoom 4 cases. The SSIM would suggest that local structures are better reconstructed by RDN and this is pertinent with

the stronger coherence that RDN has with the ground truth that we can observe when we zoom. UIQI, on the other hand, acknowledges a good performance from both networks without much difference for a scale factor 4, whereas for the bicubic this value drops. The structural of ESRGAN is pretty high, much closer to bicubic with respect to the other metrics. This index tells how vertical and horizontal lines are well rendered, hence this is again coherent with the synthetic character of ESRGAN images, that struggles in producing straight lines in output (evident in the traffic lines of figure 3.13).

In zoom 2 we can say that ESRGAN metrics performance stands more or less in betwwen bicubic and RDN, being the two relative differences often similar. In scale factor 4 this is not anymore the case, as ESRGAN metrics are closer to bicubic than RDN ones. This fact is somehow unexpected since visually RDN seems to return the same image for zoom 2 and 4, while ESRGAN improves the level of detail between the two cases. This can be due to the ESRGAN's consistent artifact generation, that is more and more evident when increasing the zoom.

The radiometry is not conserved through the networks. Pixels composing an object may assume a different value in the SR image in each channel. This is also due to the fact that data seen by the networks during training are normalized, hence they work with normalized histograms and not with real ones. The radiometry shift is more evident for ESRGAN than RDN. This should be related to the loss used in the two cases. RDN was trained with a L1 loss ((Eq. 3.1)), whose purpose is to make the DNN output numerically as close as possible to the ground truth. On the other hand, ESRGAN has a loss more linked to the perception ((Eq. 3.2)) and thus it does not always prioritize the difference between pixels.

When looking at the full scale images of figure 3.10, we have the impression that super resolved images approach the target while having a different radiometry. But it is when we look closer that we can more appreciate the differences between the different version of the image. Well recognizable small scale objects such as cars, and straight features such as traffic lines and building edges, are often used as benchmark for evaluating the performance of a SR deep network. Still in Fig. 3.10 we see that the cars generated by RDN are closer to the ground truth ones, while the ESRGAN detail of the tree on the top left of the image closely resembles the HR view. We also have the feeling that homogenous regions, e.g. the left building's roof and the asphalt, are forced to be smooth in RDN case while unnecessarily textured in ESRGAN's.

As a matter of fact, for a scale factor 4, ESRGAN perceptually outperforms bicubic and RDN in terms of sharpness and detail of the learned structures, confirming the literature's

LR Bicubic x4 RDN x4 ESRGAN x4 HR

Figure 3.11: Test image from BD Merou dataset, scale factor 4

Scale factor 4

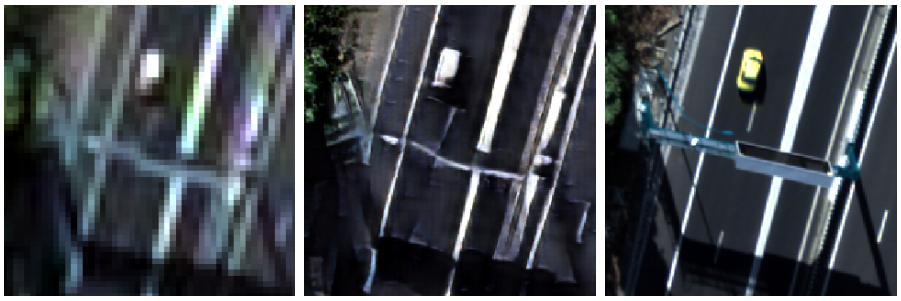
	PSNR [dB]	SSIM	ERGAS	SAM	UQI	Struc. err.
Bicubic	17.47	0.3191	0.5627	10.591	0.6186	91749
RDN	22.94	0.5898	0.2902	4.676	0.9025	58275
ESRGAN	19.31	0.4032	0.4421	13.452	0.8010	89130

Table 3.4: Reference metrics for BD Merou test dataset for bicubic, RDN and ESRGAN upscaling.

claim around GAN architecture. A non-GAN architecture, indeed, although able to return an image considerably more sharpened of a standard bicbic upsampling, seems not to be able to reach HR definition. But when looking closer, we see again how RDN is more attached to the truth whereas a GAN fills an image with artifacts. Flagrant are the details pictured in figures 3.12 3.13 3.14, where the ESRGAN generates fake details that could be real. This is relevant because it's a proof that this network learned to render cars, that are a real world common object. Less common objects, such as air-conditiong industrial plants (Fig. 3.12), or highway traffic signs (Fig. 3.13) cannot be learned because there are very few (or even absent) samples in the training set and therefore are upscaled as the object that more closely resembles their LR version, i.e. a car. In practice, better definition comes at the price of the so called hallucinations. This generative model shows here all its power in simulating the truth and, at the same time, all its limits because of synthetic characteristics of the outputs. 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 mismatch the area synthetically modified by the network. Furthermore, even if the hallucinations are coherent we risk big mistakes in the disparity map: for instance in Fig. 3.13 the traffic sign has i different altitude with respect to the ground, but being transformed into a line we totally lose such an information.

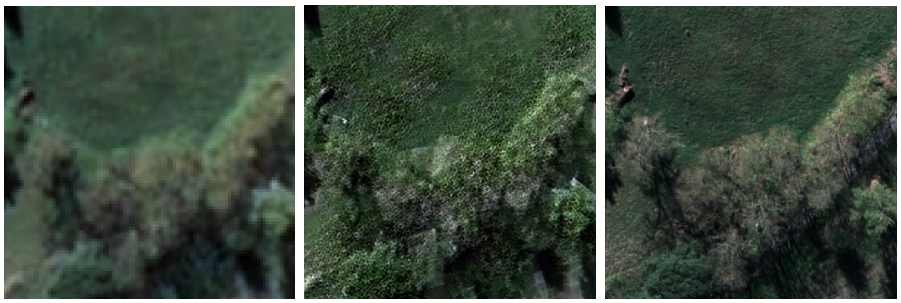
LR ESRGAN HR

Figure 3.12: Test image from BD Merou dataset, scale factor 4. First example of ESRGAN hallucination: an air-conditioning plant is transformed into a car



LR ESRGAN HR

Figure 3.13: Test image from BD Merou dataset, scale factor 4. Second example of ESRGAN hallucination: a traffic sign is super resolved as a traffic line



LR ESRGAN HR

Figure 3.14: Test image from BD Merou dataset, scale factor 4. Third example of ESRGAN hallucination: a tree texture is inappropriately extended to a meadow

Inference on satellite images

Once the ability of the trained network to well super resolve some test images was assessed, real satellite images were given as input to understand the performance on a real application case. As stated in section 3.1, the main objective is to improve the resolution on

Pléiades images. At this purpose, 2 Pléiades acquisitions on the french cities of Toulouse and Montpellier were used, their details are reported in paragraph 4.1. Although it is true that in the train set BD Merou some images of Toulouse were present, the same is not true for the city of Montpellier and the network performances on the two sites is totally comparable. We recall that quantitative measurements on these data are impossible as no reference is available. From a visual inspection, though, we can state that bicubic upsampling is outperformed by the networks. Although not measurable, the impression here is that network performance on real images is not as good as in the simulated validation test, yet the LR sample is improved in a perceptual sense. Edges are sharper and objects more detailed, even if not necessarily in a physical manner. The zoom factor 4 doesn't seem to add any relevant information with respect to the zoom 2 in bicubic and RDN cases. RDN at 12.5 cm might look sharper but the amount of detail at which the objects are rendered seems to be stagnant. On the other hand, ESRGAN has the ability to super resolve further the objects when increasing the zoom. However, the world returned by ESRGAN lens is far from the reality.

Conclusions

By extensively studying the samples it is possible to note down the following intuitions, that resume what illustrated in this section and validate the SR part of this work:

- Both networks have the ability of removing the blur and straighten the edges of any object: buildings, cars, traffic lines, etc. (Fig. 3.10).
- The radiometry is not conserved through the networks, while the geometry is fairly well rendered. Sometimes we can note some building edges and traffic lines that are curved in the SR images, while in reality they're straight, meaning that the networks only partially master the notion of geometric primitives.
- ESRGAN's results might be more pleasant when we look at the whole image, but it turns out to be conditioned by artifacts once we zoom (Fig. 3.15). When looking closer, we may prefer RDN because it seems to prioritize the conservation of contours of some objects like cars and buildings, yet they're still not realistic because their interior is artificially smoothened (Fig. 3.10, 3.15). On the other hand, ESRGAN presents an impressive rendering of some textured objects such as trees (Fig. 3.15), while it struggles to be reliable for human made objects (e.g. buildings, cars).
- We can appreciate two oppositely diverging behaviors in uniform zones for the two networks. Such regions of an image are characterized by similar radiometry, yet

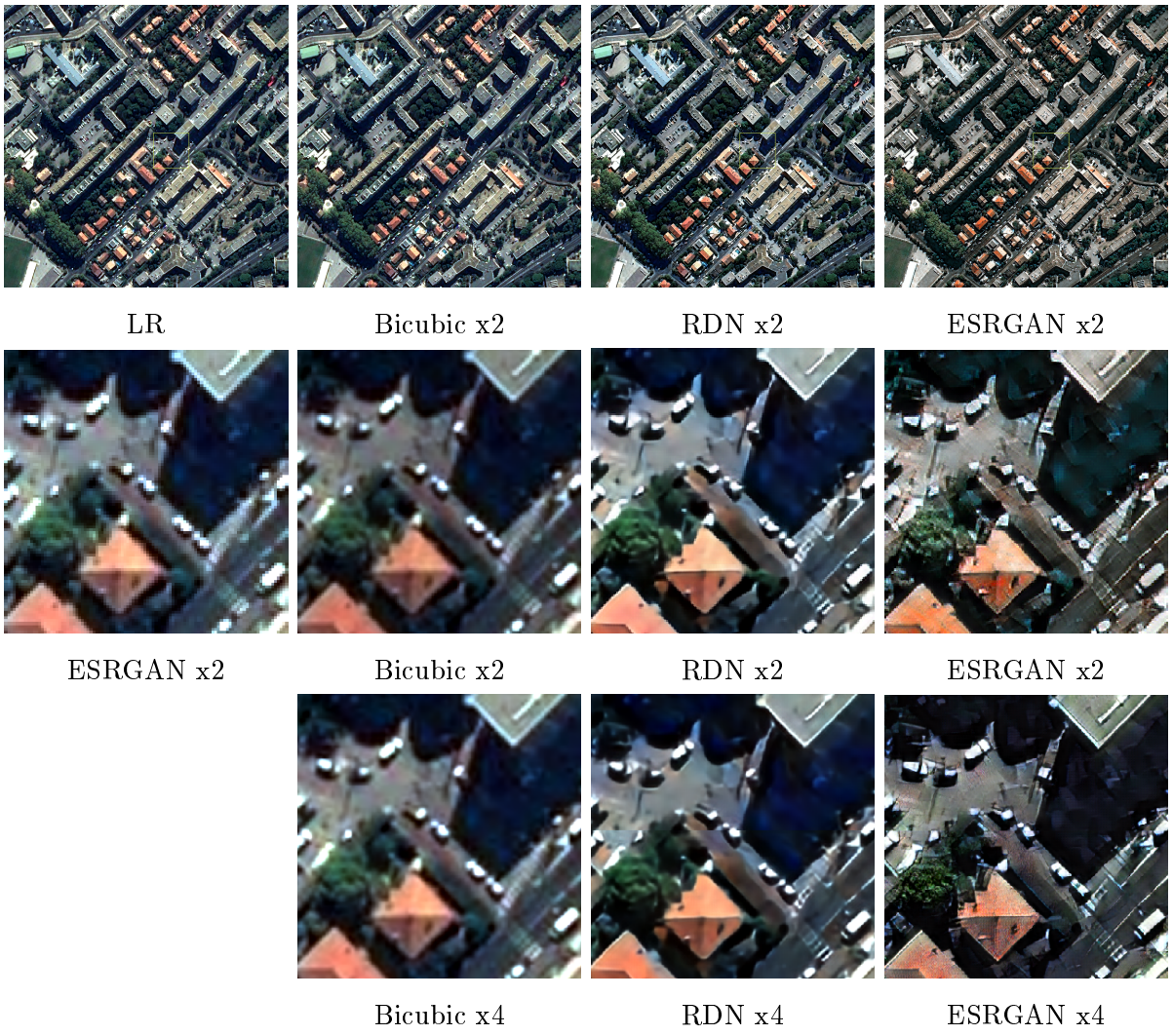


Figure 3.15: Test image from Montpellier dataset. A smaller crop taken in the green square is zoomed.

some details are still present when looking closer (Fig. 3.16). Such details seem to be flattened by RDN that therefore makes homogenous surfaces even smoother. On the other hand, ESRGAN amplifies them, adding non physical textures. In the context of a 3D pipeline, textures in uniform regions are supposed to help stereo matching algorithm, yet they have to be coherent between left and right images. These aspects will be treated further in sections 4.2 and chap. 5.

- What can guessed for a scale factor 2 emphasized in scale factor 4 Increasing the scale factor means pushing further the SR technology as the information available to reconstruct the signal is taken at a higher sampling distance. The distinction between theoretical resolution, i.e. ground extension of the acquisition divided by the number of pixel, and effective resolution, i.e. the size of the smaller detail

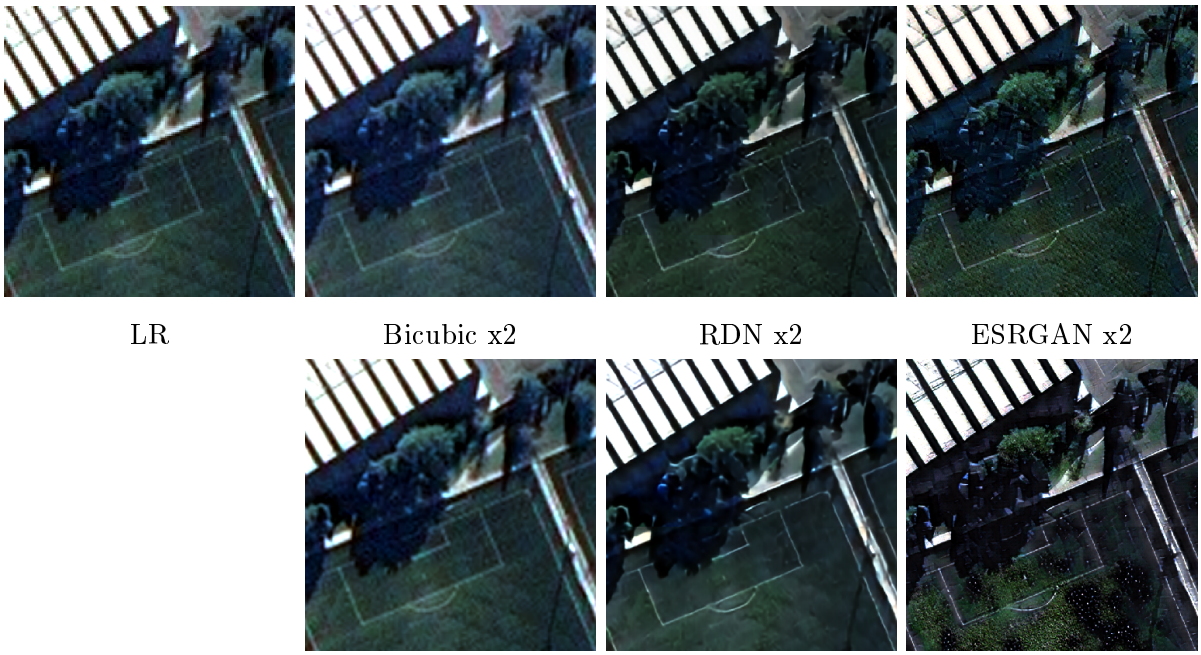


Figure 3.16: Test image from Toulouse dataset

that can be reconstruct through the upsampling technique, is more evident. For instance, RDN seems to have an asymptotical behavior with respect to scale factor: the amount of information that we can appreciate at 12.5 cm GSD is more or less the same that we observe at 25 cm. The ESRGAN, instead, manages to add further detail, very often in form of artifacts or hallucinations.

4 | DSM generation from pairs of super resolved images

So far we treated the single image super resolution, yet during this internship SR was not implemented for it's sake but for enhancing DSM generation via CARS-Pandora pipeline. Once SR proved to be robust enough for the specific application, the 3D aspects of this project could bey tackled. In this chapter we'll apply the trained network to the left and right images and study the impact at different levels of the above mentioned pipeline.

The considered scale factors between LR and SR images are 2 and 4. This means that, for LR images of Pléiades type, the GSD of 50 cm and SR ones are at 25 or 12.5 cm. The GSD of the final DSM can be set when running the CARS pipeline. For these results, a 50 cm resolution was chosen.

4.1. Methodology

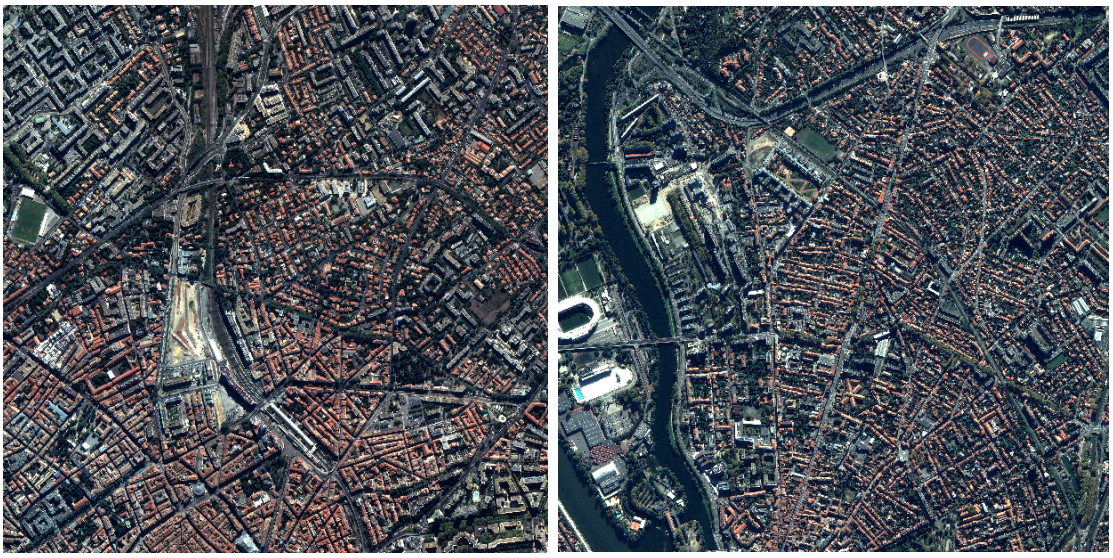
4.1.1. Definition of a dataset

First, it is necessary to define the test data. Since the the mission CO3D will supply VHR images at 50 cm GSD, the idea is to use use Pléïades couples in a site where possibly a good 2D and/or 3D reference was available. Fortunately, some accessible data correspond to these requirements. They will be referred as:

- Montpellier: sensed stereo data of the city of Montpellier, France, Fig. 4.1a. The left image image is 50726x62667 in panchromatic (PAN), the right one 48873x60854. The multispectral (MS) data (Red, Green, Blue and NIR) are four 4 times less resoluted. A lidar 3D model of 22680x22680 pixels rasterized at 50 cm resolution is also available, covering a considerable part of the optic acquisition. The characteristics of the instrument that took the acquisition are not known. Similar lidar scanners of the scale of a city are usually done, and can generate a point cloud of 30 pts/m^2

[39]. The confidence on the measure is typically in the order of tens of centimeters. From all these data, a smaller region of interest (ROI) of 3682x3622 pixels was selected, corresponding to about 3.33 Km^2. Indeed, given the local nature of the analysis performed there's no point in treating an enormous amount of data. The ROI was chosen in a densely populated area of the city that includes different types of buildings and urban configurations.

- Toulouse: same kind of data as Montpellier, with the definition of an analogous ROI, but for the city of Toulouse, France, 4.1b. The greatest difference lies in the lidar data which, unlike for Montpellier data, are not directly exploitable for visual purposes but would require additional process. This, together with the fact that some images of Toulouse were actually seen by the network during the training, makes Toulouse a little less ideal site for the study.



(a) Overview of Montpellier dataset's ROI (b) Overview of Toulouse dataset's ROI

4.1.2. Processing pipeline

In spite of the reduction of the ROI, the data were still too large in size to be passed to CARS without making the computational effort unsustainable. Hence, such a ROI has been divided in 9 overlapping tiles of approximately 1536x1536 pixels. Each tile undergoes the same treatment and the resulting DSMs are eventually merged to form the ROI's DSM.

The ROI are extracted from the raw data, in panchromatic band as well as in the corresponding multispectral raster, for both left and right acquisitions. Then, left and right

data go through a pansharpening operation; The *GDAL*[1] pansharpening function was used for its better handling of data that do not precisely have the same extent and because it is possible to set the weights used for the algorithm for each band. Such weights are parameters of the CSI, used by the pipeline to generate the panchromatic sample when the train dataset was generated (Sec.3.2).

From the pansharpened 4-bands rasters we extract the RGB image and this constitutes the LR sample, that will be further processed in 3 different ways to obtain the 3 high resolution versions: bicubic upsampling, super-resolved with RDN, super-resolved with ESRGAN. Inference by the networks is done straightforward with the process described in section 3.1. Bicubic upsampling is performed with *GDAL*[2] built in function.

Since in order to do stereo vision we need two gray scale images, only one band was extracted from RGB images. Blue band is discarded because it doesn't separate well the vegetation from the surroundings. Green and red bands are kept. Green band is what it' usually used for CO3D pipeline as the CO3D sensor will be array-like with a Bayer matrix, and hence green band contains more information than the others. This is actually not the case for Pléïades sensors and images. The red band was also selected because in historical european city centers like Toulouse and Montpellier it's common to find roofs in terracotta, material that can be associated to clay soils which typically reflect more in red wavelength.

Finally the four couples of images can be input into CARS-Pandora. No modifications to the CARS-Pandora code were necessary, but some adjustments at configuration / data stages had to be applied.

The most significant alteration regards the geometric model (RPC) of the super resolved / bicubically upsampled images. In practice, during the rectification step in CARS (see section 2.1.2), the epipolar images are resampled, from the input stereo pair, on a grid whose size is calculated irrespective of the pixel size specified in the image metadata. Therefore, the RPC has to be rescaled in order to be coherent with the image and have an unitary pixel size. Another little trick that may help with respect to CARS default usage is to set limits to the , (`elevation_delta_lower_bound` and `elevation_delta_upper_bound`) range of disparity calculated in the `prepare` step. When larger images are given as input, CARS tends to overestimate such a range (sometimes by one or two orders of magnitude) making it impossible for the correlator to estimate the good disparity.

[1] GDAL pansharpening documentation
[2] GDAL translate documentation

For what concerns Pandora, different configurations were tested since the correlation is the most critical step. In chapter 5 these aspects will be better illustrated. At this stage, it is useful to remark that the size of the window used for matching should changed between original image and SR in order to contain the same information. For example, a detail which is visible in a 5x5 window in a LR image, will be visible at the same scale in a 9x9 or 11x11 window in a SR image upsampled by a factor 2. Tables 4.1a and 4.1b resume the parameters used during the internship in the CARS-Pandora pipeline when different from default settings[3].

Epipolar err. up. bound	80
Epipolar err. low. bound	-80
Elevation delta up. bound	20

(a) CARS

Matching cost method	census, ZNCC
Window size	5,9
P1	8 (census), 2 (ZNCC)
P2	32 (census), 4 (ZNCC)

(b) Pandora

Table 4.1: CARS and Pandora configurations utilised if different from default

4.1.3. Metrics and means of analysis

Once generated, the different DSMs need to be evaluated in order to asses the SR contribution. 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 [16]. Moreover, the reference may suffer from the same biases as the tested DSMs in case a photogrammetry DSM is used. When relying in a lidar as reference, there might be some temporal differences (e.g. new buildings) in comparison to the utilised data. In other words, 3D (or 2.5D) metrics is not as developed and reliable as the 2D one. Therefore, limited attention will be given to the quantitative results because restricted to the measure of error difference, and not really capable of targeting some specific features, such as building shape, urban context rendering, which are in fact the main scope of the project. On the other hand, qualitative results such as DSM renderings will be the main focus of this chapter discussion.

Demcompare[4] is an open source tool developed by CNES that allows to compare two DEMs together: one taken as reference, the other is the tested one. The software computes a wide variety of standard metrics and allows one to classify the statistics with user defined

[3] CARS documentation, Pandora documentation
[4] Demcompare Github repository

criteria (default is slope classification). Furthermore, this tool is robust to different DEMs characteristics such as format, projection, etc. A coregistration step can be included in the pipeline, based on Nuth & Kääb universal coregistration method [37]. Some demcompare results will be proposed in order to supply quantitative measures, although it is likely less demonstrative than the reader's visual analysis.

In addition, with the purpose of circumscribing the real influence of SR to the photogrammetry pipeline, in chapter 5 we'll take a look into the disparity maps and cost volumes in the upper stages of the Pandora pipeline.

4.2. Results

Both datasets (Toulouse and Montpellier) were tested for the two selected bands (red and green) as well as for different configurations of the correlator (matching cost method and window size) and scale factors (2 and 4). In this way, remarkable properties of the different cases can be generalized, making the intuitions as less dependent from the peculiar configuration as possible. However, for each result shown, the parameters used in terms of disparity estimation, scaling and band used will be reported.

Figures 4.2 and 4.3 show input left image and output DSM for Montpellier and Toulouse datasets, respectively. In both cases the red band was extracted and a ZNCC (Eq. 2.4) cost measure was used. The window size is set to 5 for low resolution, while for high resolution it is approximately doubled for a scale factor 2 and quadrupled for a zoom 4. This is because the window must contain the same type of information in order to compare the two situations. If a car is present in the LR patch, we must see the very same car the SR frame, but better resolved.

From a quantitative point of view, tables 4.2 and 4.3 resume standard statistics for the four cases in Montpellier dataset for scale factors of 2 and 4, respectively. A computation of the error was only possible for the Montpellier dataset as in no reference was available for Toulouse one

If we consider mean and root mean square error (RMSE), it looks like the gain in upsampling the stereo pair is negligible or even that this step is detrimental to the results. As a matter of fact, standard statistics are not reliable indicators when we don't control perfectly the photogrammetry chain. Indeed, they are based on the assumption that the errors follow a Gaussian distribution and that no outliers exist.

But this is not the case as the correlator may results outside any distribution in certain

Stereo pair super resolution scale factor 2

	% valid points	Mean error	RMSE	Median error	NMAD
LR	95.53	-1.00	4.17	-0.55	1.27
Bicubic	94.29	-1.10	4.02	-0.49	0.92
RDN	94.99	-1.10	4.18	-0.47	0.84
ESRGAN	95.03	-1.01	4.11	-0.44	0.88

Table 4.2: Standard statistics for Montpellier benchmark calculated with Demcompare, red band, ZNCC matching cost method with window size 5 for LR and 9 for the other cases

Stereo pair super resolution scale factor 4

	% valid points	Mean error	RMSE	Median error	NMAD
LR	95.53	-1.00	4.17	-0.55	1.27
Bicubic	98.11	-1.24	4.43	-0.49	0.98
RDN	98.38	-1.28	4.65	-0.51	0.97
ESRGAN	98.18	-1.09	4.71	-0.49	1.13

Table 4.3: Standard statistics for Montpellier benchmark calculated with Demcompare, red band, ZNCC matching cost method with window size 5 for LR and 19 for the other cases

areas, for examples in façades and in uniform zones without texture. Moreover, the disparity range within which we search the real disparity is estimated by CARS in the `prepare` step (Fig. 2.2), unless known a priori. When underestimated, this leads to some glaring errors such as the missing reconstruction of the highest buildings in the scene (because they correspond to the highest/lowest values of disparity). This occurs especially for upsampled data where the values of such a range are doubled with respect to a standard GSD image. Therefore, error computation for bicubic and network cases may suffer from this problem. Whether or not this is linked to the networks behavior would require additional study. Figure 4.4 illustrates such an effect.

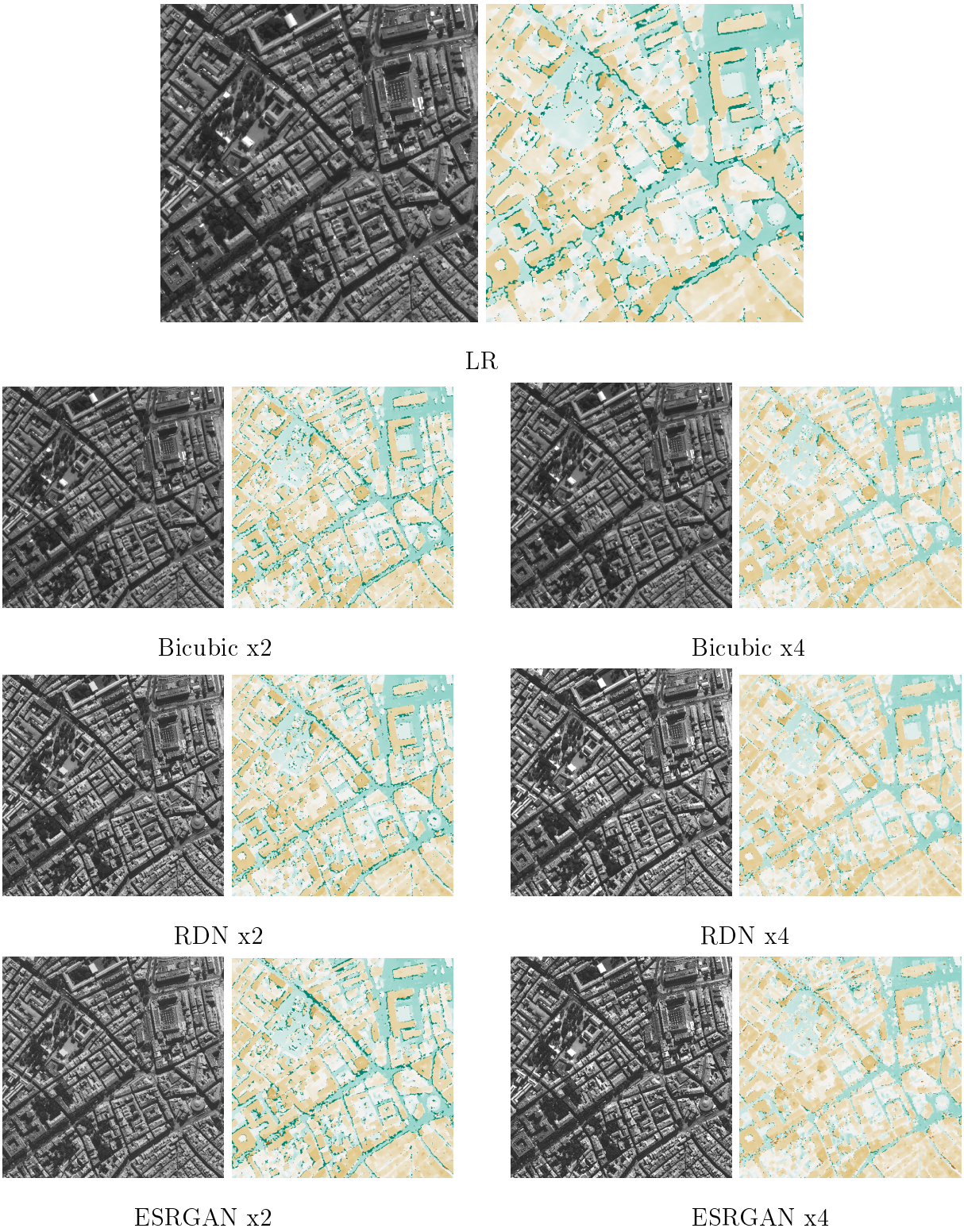


Figure 4.2: Crop from Montpellier dataset, left LR, SR and bicubic upsampled images for a scale factor 2 and resulting DSMs. Red band, ZNCC matching cost, window size 5 for LR, 9 for Bicubic, RDN and ESRGAN

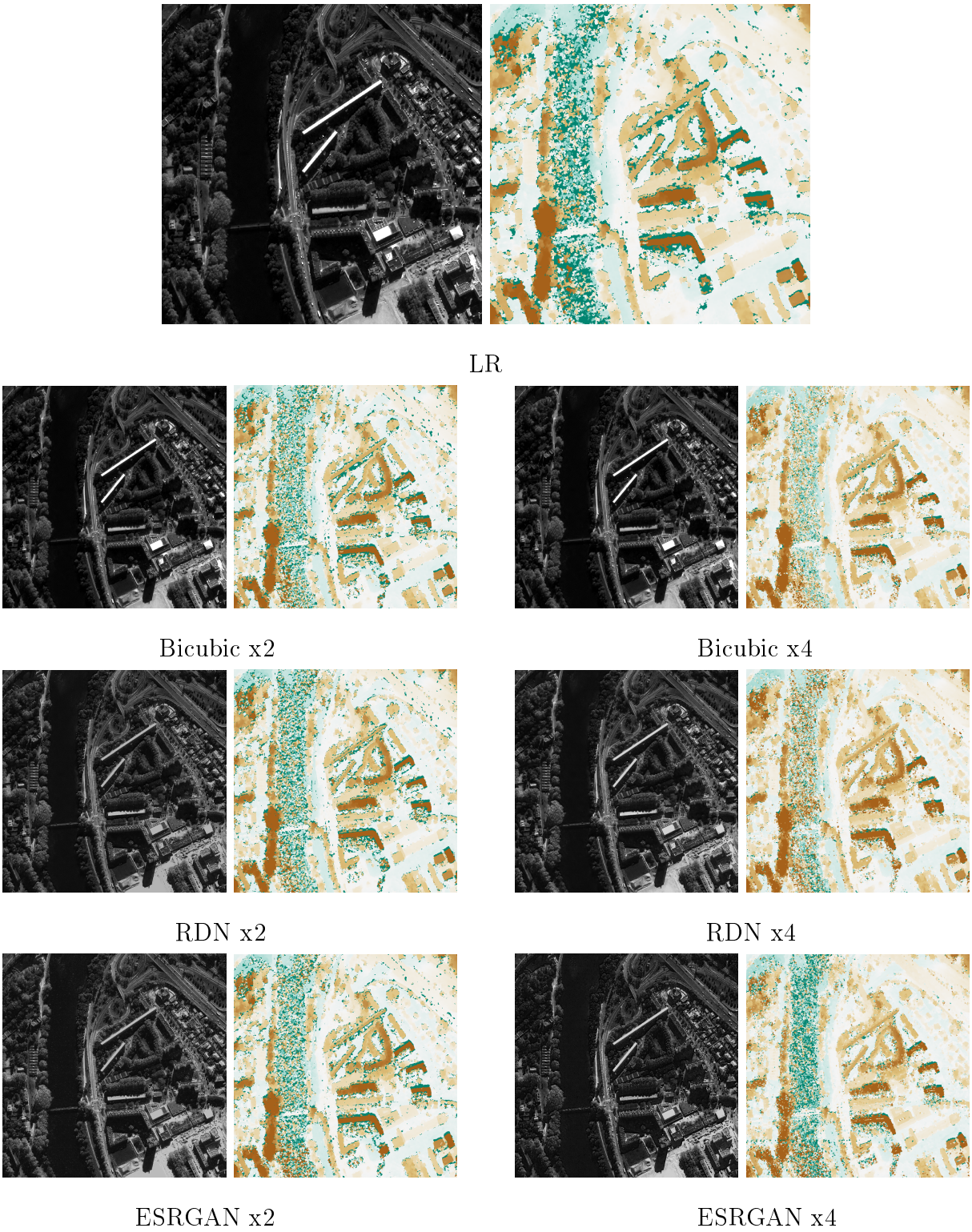
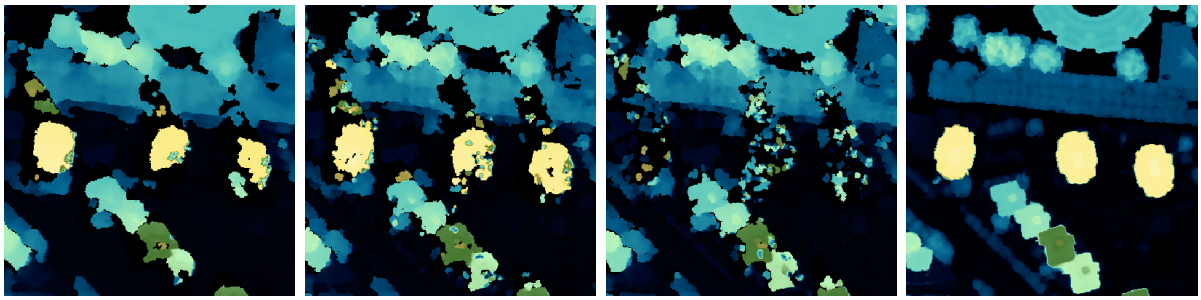


Figure 4.3: Crop from Toulouse dataset, left LR, SR and bicubic upsampled images for a scale factor 4 and resulting DSMs. Red band, ZNCC matching cost, window size 5 for LR, 19 for Bicubic, RDN and ESRGAN



LR, disparity range estimated by CARS [-32, 10] pixels

Bicubic, disparity range estimated by CARS [-68 18] pixels

RDN, disparity range estimated by CARS [-38 17] pixels

Lidar, disparity range estimated by CARS [-36 10] pixels

Figure 4.4: "Disappearing building" phenomenon: detail from Montpellier DSM. Disaprity values are multiplied by the scale factor when using upsampled images which is 2 in this case

Since this is a common issue when dealing with DEMs, *demcompare* proposes the NMAD index (Eq. 4.2) [16]. It's a sort of estimate for the standard deviation more resilient to outliers in the dataset.

$$NMAD = 1.4826 \cdot median_j(|\Delta h_j - m_{\Delta h}|) \tag{4.1}$$

Δh_j are the individual errors and $m_{\Delta h}$ is the median of the errors. We see that, according to this indicator, there is a remarkable improvement for DSMs generated by upsampled images, up to about 34% reduction for RDN recall that for Toulouse dataset it was not possible to generate these statistics as no reference was available.

If we look at table 4.3, we can surprisingly remark that statistics are worse for the scale four case. However, this also comes with an increase in valid points percentage of more or less 3%. This means that less points are marked as invalid by CARS. It is therefore unfair to compare directly the two tables, since they do not compute the statistics on the very same points. We can rather observe the differences between the upscaling methods. In this sense, bicubic upscaling seems more beneficial to DSM generation: the output raster has the lowest RMSE and a NMAD almost equal to the RDN one. ESRGAN super resolution, unlike for the zoom 2 case, leads to a larger error in terms of both RMSE and NMAD. This should be due to the highest amount of noise generated with ESRGAN inputs (observable for example in Fig. 4.3, ESRGAN x4 case) which in turn is caused by the artifacts introduced with such a network that lead to mismatches.

No data values are often in correspondence to building façades, that are details notoriously difficult to reconstruct in a DSM: when seen from above from two even slightly different angles, their appearance may change a lot. Additionally, because of the perspective, often they are represented by very few pixels. These two factors together make very challenging to perform stereo matching on façades.

In general, an artificial increase in stereo pair resolution leads to more noise. Hence, even if 3D objects might be slightly better rendered, the larger noise amount causes larger global error as it can be seen in Tab. 4.3. This is confirmed by the standard deviation values of DSMs, reported in table 4.4. No data were not taken into account for this computation.

	LR	Lidar
σ [m]	4.580	4.576

	Bicubic x2	RDN x2	ESRGAN x2
σ [m]	4.634	4.621	4.643

	Bicubic x4	RDN x4	ESRGAN x4
σ [m]	4.661	4.705	4.944

Table 4.4: Standard deviation of analyzes DSMs, excluding no data values. MTP dataset, red band, ZNCC matching cost

We can appreciate how upscaling produces an increase in standard deviation, which is very similar to the reference (Lidar) for the LR case, has a slight increase for the zoom two tests and it's much larger for a zoom 4 of the stereoscopic couple. Moreover the standard deviation remains approximately the same for bicubic between the two factors, while its value rises considerably for the ESRGAN x4 case. This confirms the visual impression when looking at figures 4.3 where the noisiest DEM is the one labeled ESRGAN x4, and it is again coherent with the scarce reliability that ESRGAN outputs have with respect to the reality: artifacts may lead very easily to wrong matching hence to introduce bad values that are able to pass through the noise filters of the photgrammetry pipeline.

Additionally, it is possible to estimate the confidence on the measure by computing statistics on small ROIs that are approximately flat. Some examples are football pitches, roofs of industrial complexes, rivers, squares. By selecting a sufficiently large area or by repeating this measure on many different zones we can estimate the confidence on the altitude as the range of oscillations of the interested pixels with respect to their mean. An example is reported in Tab. 4.5 and Fig. 4.5. Many other samples were taken into consideration but here we only report this one as it is the most effective. Indeed, the particular roof pattern helps stereo matching because of the univocal information it vehicles. Football pitches are, for instance, less adapted as they often do not have details that can be used

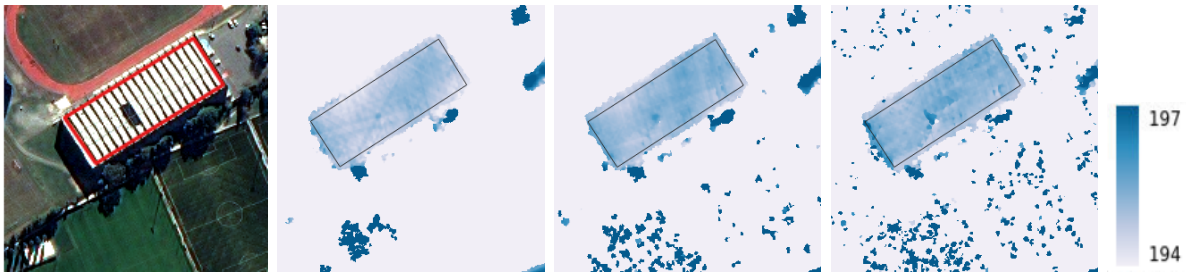
during matching. This sample comes from Toulouse dataset, where no reference is available. The accuracy of the lidar used in Montpellier dataset was estimated in the very same way, since the characteristics of the instrument were not known. The lidar estimated confidence is around 30 cm.

	LR
μ[m]	194.91
h_{max}[m]	195.63
h_{min}[m]	194.11

	Bicubic x2	RDN x2	ESRGAN x2
μ[m]	195.22	195.13	195.16
h_{max}[m]	195.91	196.21	196.08
h_{min}[m]	194.32	194.48	193.98

	Bicubic x4	RDN x4	ESRGAN x4
μ[m]	195.31	195.25	195.25
h_{max}[m]	196.22	196.32	196.70
h_{min}[m]	194.52	190.40	191.57

Table 4.5: Mean, maximum and minimum values statistics computed on the zone of Fig. 4.5. Value range increasing with the scale factor.



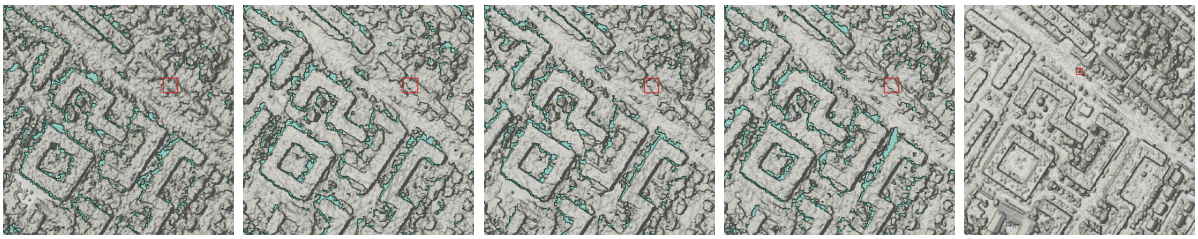
RGB LR image LR ESRGAN x2 ESRGAN x4

Figure 4.5: Roof used for statistics calculation in Tab. 4.5, RGB image and detail from the DSM produced via LR, ESRGAN scale factor 2, ESRGAN scale factor 4 input stereo pairs. Note the increasing amount of noise with the scale factor as well the amplitude of magnitude of the oscillations around the mean value of the roof's height

It is evident from Tab. 4.5 and Fig. 4.5 that DSMs generated via photogrammetry have less accuracy than lidar ones, but this is an expected result and the two technologies cannot be directly comparable. However, being lidar confidence on the altitude measure empirically estimated to be about one order of magnitude higher than stereo-reconstruction (20-30 cm vs 2-3 m), it is correct to assume laser scanning measures as the ground truth.

Another way to see the resolution of the two type of DSMs is to evaluate the smallest scale of the recognizable objects. In figure 4.6 we can observe that the lidar DSM well separates small size buildings, cars and trees. In the red box we can distinguish a car, approximately 3 meters long. The DSMs generated via CARS do not reach the same

resolution in terms of semantics. We can well see big buildings, while small size ones are barely recognizable. Cars and trees are totally lost. In the red square, a recognizable building edge measuring about 9 meters. When upscaling the input images, we gain in object definition and no real differences can be observed between the different upsampling cases, whereas when we keep original image resolution ("LR" case), the impression is that even less detail is available, as we struggle to recognize any object which is not a large building.



LR Bicubic x2 RDN x2 ESRGAN x2 Lidar

Figure 4.6: Detail from Montpellier dataset DSMs generated via LR and the 3 upsampling mode generated image inputs. Lidar DSM is also reported. In the red box the smaller scale recognizable feature

Quantitative results may suggest that a slight gain in DSM accuracy can be achieved by upscaling the input stereo satellite pair. On the other hand, we can observe an increase in noise (Tab. 4.5). A higher scale factor leads to more DSM completeness but not necessarily to a lower error but instead to even more noise, especially when using neural network super resolution techniques that do not preserve radiometric consistency. A visual analysis was simultaneously carried out in order to understand what might be the areas where the proposed methods perform well and where instead they have their weak points. Qualitatively we can distinguish two different trends: in regions characterized by high contrast (building edges, contours, streets) SR may in fact provide a better suited input pair, since it enhances further the contrast making it easier for the stereo algorithm to find the good match. On the other hand we have uniform or textured areas, where neural networks may add inconsistent textures leading the stereo matching to find the wrong match and this introduces noise in the stereo pipeline.

High contrast areas

When looking at the center building of figure 4.7, its edges are not well defined in the LR 3D model, in the sense that they are far from being straight. The situation is improved when passing to a GSD of 25 cm for the stereo couple considered. In particular, in RDN case the longest wall is relatively little distorted. This might be due to the strongest

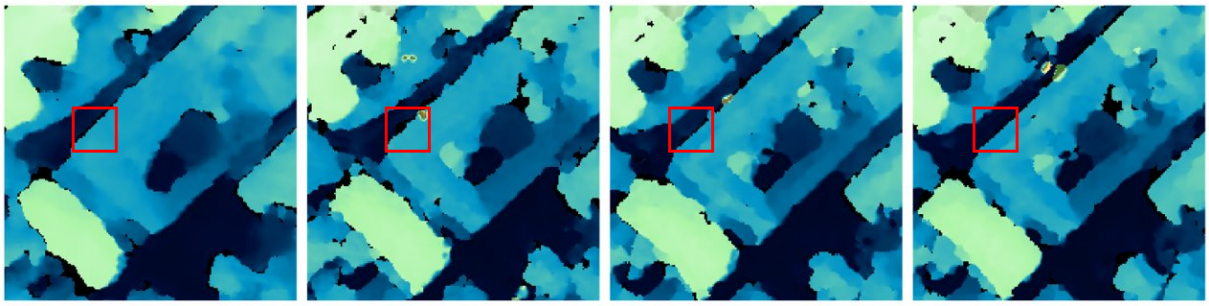


Figure 4.7: Detail from Montpellier dataset, red band ZNCC matching cost, scale factor 2. Building edges are better defined when upscaling the input stereo pair, and further straightened when using SR networks

capacity of RDN of sharpening objects without adding texture, yet one example can not be taken as evidence. However, this example will be reconsidered in chapter 5 where it shown how SR networks can enhance stereo matching for such a case. Indeed, they amplify the radiometric difference between street and building making it clearer the distinction between these two objects. In the right conditions, this property can be reflected in the 3D model at the end of the pipeline.

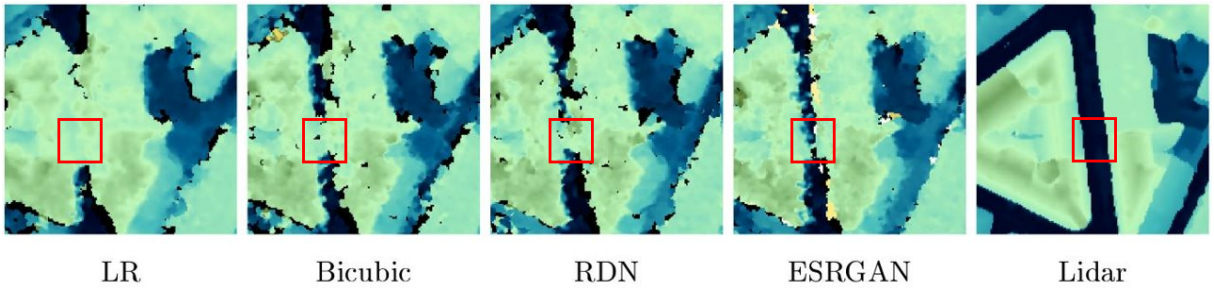


Figure 4.8: Detail of a narrow street from Montpellier dataset, green band census matching cost, scale factor 2.

Between the two buildings of figure 4.8 there is a narrow street, that is almost totally canceled in the LR case in favor of the buildings, while in the DSM generated with upscaled images we better guess the extent of the street. In this case, ESRGAN behaves better than the other methods as the way is continuously visible. Even if in the 3D models generated by a LR stereo pair very often we can observe worse rendering of the buildings, i.e. less recognizable shape and junction of different constructions, it is not as common to individuate differences between the DSMs generated via the 4 upscaled versions of the original image. This is relevant as one of the main applications of DSMs is to precisely map a city and it is therefore indispensable to well separate the buildings composing a

city. Indeed, it is not always achieved as it can be seen in figures 4.3 and 4.6. However, in the case of Fig. 4.8 ESRGAN shows to be beneficial to the photogrammetry products. In chap. 5 we'll dig further into this example.

Uniform areas

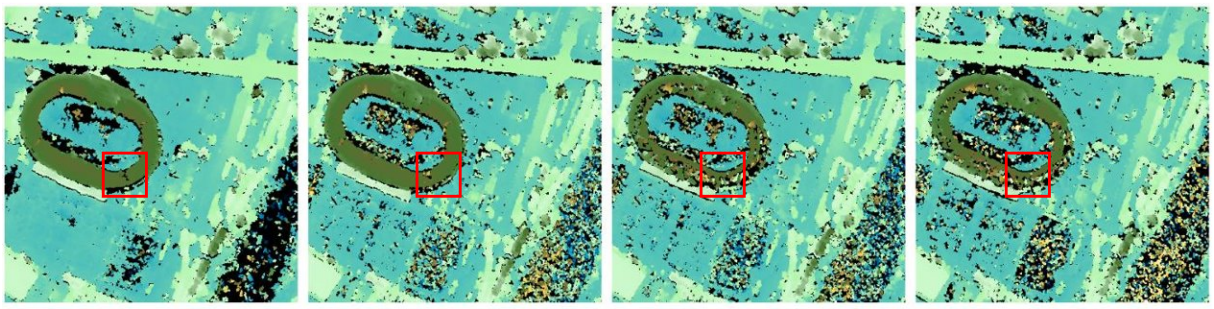


Figure 4.9: Detail of the rugby stadium and Garonne river from Toulouse DSM. Green band ZNCC matching cost, scale factor 2.

In figure 4.9 we can immediately notice how the stadium rendering is completely deteriorated when utilising CNNs. This is because some details characterizing the upper side of the stadium are curiously lost during super resolution step (Fig. 4.10). This is unexpected and might be linked to the fact that a stadium is a peculiar building and nothing similar was "seen" by the networks during the training. The results is a radiometric flattening of stadium roof that in turn leads to a more complicates matching and strong errors in disparity estimation.

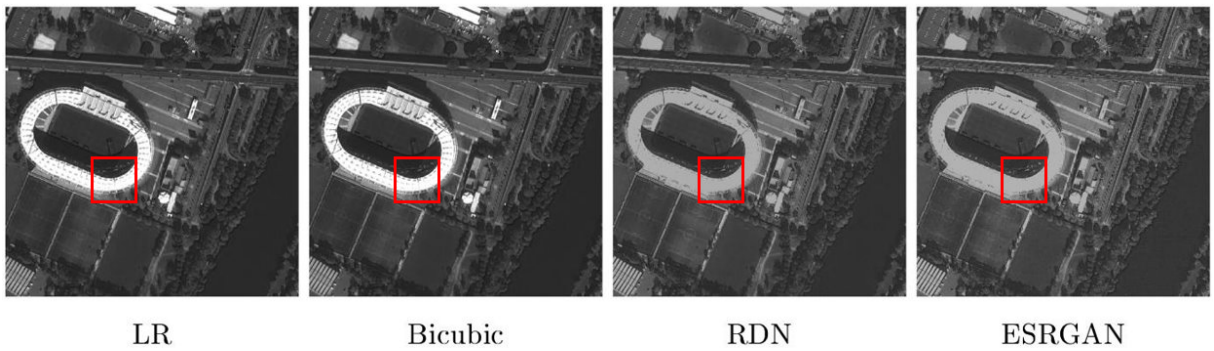
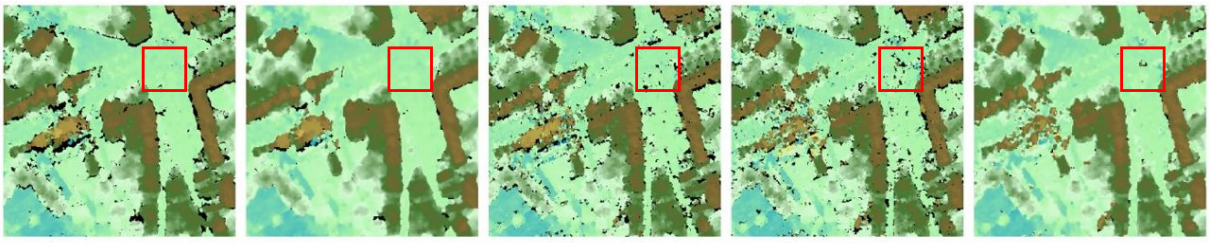


Figure 4.10: Detail of the rugby stadium and Garonne river from Toulouse dataset, left image. Green band, scale factor between LR and Bicubic, RDN, ESRGAN is 2.

Moreover, in this capture there are two really uniform areas, the football pitch on the bottom center, and the river Garonne, bottom right. Both suffer from the important noise, especially in the ESRGAN case. This should be linked to the ESRGAN's texturing

property (section 3.4), and it is likely due to the fact that left and right images are not textured in the same way, introducing confusion at the correlation step. Indeed, ESRGAN patterns always stem from real local changes in pixel values, which are amplified through the network. In other words, high frequencies are put where they should not be.

One can argue that two other football pitches are present in the picture, yet their elevation is still rendered for all methods. However, by a closer look we notice that these two pitches, unlike the third one, have visible lines so the correlator is able to often find good references and thus to make the good estimate.



LR at 0.5 m DSM resolution | LR at 1 m DSM resolution | Bicubic at 0.5 m DSM resolution | RDN at 0.5 m DSM resolution | RDN at 1 m DSM resolution

Figure 4.11: Detail from Toulouse dataset, DSM. Green band census matching cost, scale factor between LR and bicubic, RDN input pairs is set to 2. Increased noise from LR to bicubic/SR

While for large uniform areas the increase in noise is evident, this property can be seen wherever there is a local homogeneous texture as well as in façades. From figure 4.11 it is possible to appreciate an increase of the noise when passing in the image 25 cm GSD domain, whether via a bicubic interpolation or a super resolution network. It's a rule of thumb that, to have a satisfactory DSM, its resolution should be 3/4 times higher then the ground sampling distance of the images in order to filter out the noise from the point cloud. We totally find in when increasing at 1 m the rasterization interval for RDN, with an input image sampled at 25 cm (Fig. 4.11).

This may have origin in the rasterization step, when an interpolation of the point cloud generated at the end of triangulation step (CARS illustration sec. 2.1) is applied. We expect a denser point cloud to be interpolated with more accuracy hence letting less noise through. But actually it's the opposite. This may mean that the noise introduced in the disparity maps from wrong matches caused by artifacts or inconsistent information between left and right image is not totally filtered out in the successive steps of the stereo pipeline.

4.3. Wrap-up

Globally, we can appreciate in most cases a slight quantitative (Tab. 4.2, 4.3) and qualitative improvement when upscaling the input pairs. Nonetheless, side effects may attenuate such enhancements or even cancel out the benefits. It is possible to note the following points:

- From a quantitative viewpoint, it is possible to highlight an improvement in NMAD statistics when using upsampled images as input of the stereo pipeline. Percentage of valid points may also benefit when a greater scale factor is employed. On the other hand, other statistics seem to contradict this enhancement.
- Visual inspection and global and local standard deviation computations certify an increase in noise that appears when upscaling stereo pairs, especially when using neural networks.
- An increase in scale factor of the stereo images does not reflect in an improvement of DSM quality. It does help to reconstruct more points, especially in correspondence of façades, but it also lead to an even noisier 3D model.
- Matching in correspondence of high contrast patches can benefit in some cases from SR networks, thanks to the stronger sharpening that halps stereo matching in recognizing the right features.
- Uniform zones seem to suffer even more than usual when SR is applied. This is again probably due to inconsistent texturing by neural networks that leads Pandor to fail more often
- No real global improvement could be highlighted with respect to bicubic upscaling.

This last point brings us to a necessary digression. One may expect to transfer the differences seen in the images in the 3D model. In practice, it is not as evident to distinguish this and similar features. This would mean that the use of SR networks for such a task is not always justified and that a bicubic upsampling would be enough. However, by a closer look we can sometimes recognize some behavior specific to the upsampling method: it is the case of the straight edges produced by RDN inputs in figure 4.7 (networks improve contrast) and the fail in stadium rendering by network generated pair in figure 4.9 (networks may never have seen a building as large as a stadium).

Furthermore, when comparing different configurations of CARS-Pandora, we can see a noticeable impact on the final DSM for all the cases analysed, perhaps more important

than the preprocessing adopted for the stereo couple. This can be seen by looking at figures 4.7 (ZNCC matching cost, (Eq. 2.4)) and 4.8 (census matching cost (Eq. 2.3)). They belong to the same dataset but the rendering of building shapes and the amount of noise differ considerably.

All these clues lead to the conclusion that the used photogrammetry pipeline is 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. This justifies a further analysis illustrated in chapter 5.

5 | Cost profile analysis

The outcomes of section 4.2 left multiple open questions: is super-resolution even beneficial to photogrammetry or should we limit ourselves to standard interpolation methods? Why a superior 2D metrics doesn't propagate in a 3D context? If SR spectra are more filled up, why such a high frequency information seems to be ineffective to reconstruct a high frequency signal such as the difference in altitude between buildings? Is it because of the consistency of the added information, or does it have to do with some process in the CARS-Pandora pipeline?

To try to answer these and other questions a further analysis was performed, trying to isolate the SR images contribution. The critical step in this sense is the matching cost computation [42] where a similarity measure is performed in order to associate patches of the left and right images along an epipolar line. Different images (i.e. different patch contents) lead inevitably to different measures. By looking at such measures for each possible disparity, it is attainable to understand the role played by a signal of diverse nature . For this examination, the Pandora was stopped before sgm optimization [15], right after the cost computation.

Figures 5.1,5.2, 5.3, 5.4,5.5, 5.6 report some significative examples, showing a comprehensive analysis of what happens for a given pixel when it comes to disparity computation. They are composed of:

- A plot of the costs associated to the pixel for each disparity of the considered range (that hereby we define as *cost profile*). There are four lines, one for the LR disparity, the other three for bicubic, RDN and ESRGAN upsampling profiles. The ground truth value is also present in the plot. The unit of measurement for the disparities is the pixel, so that LR disparity values have been mulitplied by the scale factor in order to be visually comparable to the higher resolution samples. The matching cost method considered for these plots is ZNCC (Eq. 2.4), with a window size of 5 for the LR case, 9 and 19 for the scale factors 2 and 4, respectively. Patch siwes have to be adapted to the scale factor in order to contain the same, more resolved information as in the LR image. Thie matching cost method is an arbitrary choice, due to the

fact that this measure is less affected by noise and returns more interpretable cost profiles. We recall that for ZNCC measure, the higher the value, the more similar two patches, hence we will look for the maximum costs.

- A zoom on the left image around the chosen pixel. A red square positions the window used by Pandora. Such a window is in turn zoomed and its spectrum is also shown. Looking at the frequency content of a window may allow us to understand if and how high frequencies interact with disparity estimation.
- A zoom on the right image around the matching pixel computed by Pandora. Zoom on the window and window spectrum are shown as well.
- A zoom on the right image around the ground truth matching pixel computed by Pandora. Zoom on the window and window spectrum are shown as well.

Scale factor 2

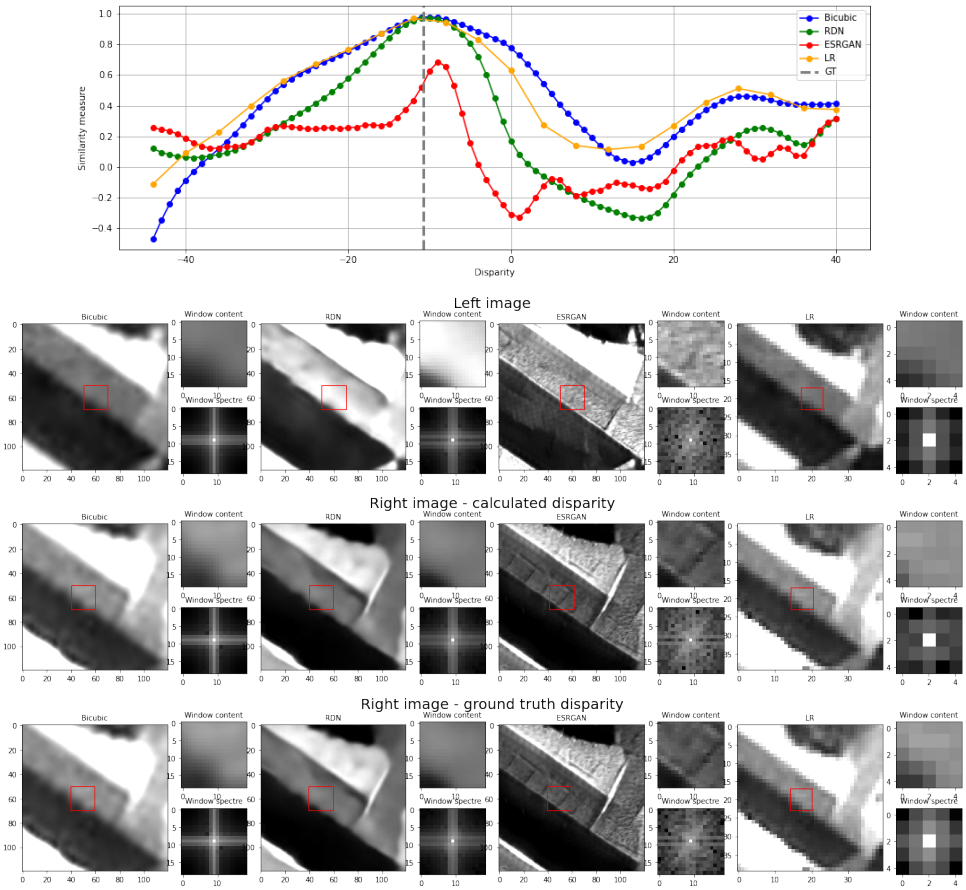


Figure 5.1: Building edge, detail from Montpellier dataset, red band ZNCC matching cost

Only the Montpellier dataset could be used for this chapter, since is the one with n

available ground truth, i.e. the associate lidar. On can arge that the lider is a ground truth for what concerns altitude measure but now we're focusing on disparity maps, that are a different measure. In order to generate the ground truth disparity, the library Beefröst was used. It's a tool developed in a collaboration between CNES and CS Group and presented in paragraph 5.1 (and more comprehensively in [5]). It produces stereo-rectified images and ground truth disparity maps, from satellite imagery and lidar.

5.1. Beefröst

The development of such a tool arise from the interest around the generation of ground truth stereo datasets, from satellite imagery and lidar, in order to train deep learning based solution for stereo matching [5]. In order to do so, Beefröst needs as input two satellite images and the associated lidar altitude data. The main outputs of our pipeline are stereo-rectified images pairs, and their corresponding disparity maps. The rectification step is based on the CARS pipeline [35] (Sec. 2.1), that returns the epipolar grids. It follows an optional coregistration step: the DSM is computed (again via CARS) and an affine transformation allows the it to be superimposed to the lidar. This procedure is often useful because the DSM or lidar georeferencing might not be perfect. The disparity maps computation consists of multiple steps. First, each pixel of the stereo-rectified image is mapped into the original sensor image coordinates using the rectifying grid. Then, the original sensor image corresponding point is localized onto the aligned lidar and the height is stored into the left disparity map. Next, a pixel-wise ratio and bias is applied to such heights (Eq. 5.1).

$$disparity(i,j) = \frac{height(i,j) - bias(i,j)}{ratio(i,j)} \tag{5.1}$$

The height corresponds to the altitude difference for a 1 pixel disparity and the ratio could be approximated by the resolution divided by the stereoscopic angle between the two views. The two informations can be retrieved from the RPC of the image. All the operations are performed on the left and the right image.

5.2. Case studies

Similarly to the DSM chapter, when studying disparity maps and cost profiles we can distinguish two main trends can be distinguished: a potential improvement of matching for windows characterized by high contrast, and a detrimental effect where the match is performed on an uniform region.

High contrast areas

In 2D, building edges are well straightened when using SR networks. This reflects in many cases in a better 3D rendering of such a feature. When it comes to disparity computation, this can be seen in the cost profiles of figure 5.1. This case study belongs to the same area shown in figure 4.7. The peak, corresponding to the pixel disparity, is better defined for the case of super resolution thanks to a change in the concavities of the functions. In this case, bicubic upsampling doesn't really add much with respect to the original, LR information. On the contrary, both neural network samples present a further, alarming, maximum around a disparity value of -20, not pointed out in LR-bicubic trends. Looking at the spectra, we see the edge orientation well reflected in the Fourier transforms for RDN and ESRGAN generated patches, while very little of this geometry information can be seen in LR and bicubic frequency representation. Such a spectral information might have been useful in better defining the actual disparity value with respect to other candidates. For this particular point, the correct disparity is identified in all cases, so the more rounded shape of LR and bicubic cases does not propagate into a wrong altitude. Nevertheless, we can imagine that along this edge the more punctual guess in the RDN, ESRGAN leads to a more defined delineation of roof and ground disparities, as well as to a more, justifying what we see in figure 4.7.

If we look at the mid-line of the roof in figure 5.2, we can see how the contrast between the two roof sides present in the LR image is strongly enhanced by the networks, that even create here a black line. This can be seen as a high contrast feature, something that allows network based methods to improve the cost profile. The frequencies added by the networks fill a wider area of the spectra, leading to superior definition for the high contrast detail shown. Although a common bias is present with respect to ground truth, the maximum corresponding to the disparity much more discriminated by RDN, ESRGAN, whereas the LR, bicubic profiles are similar and very flat. Again, all methods manage to have their maximum at the right disparity, but we would be much more confident in the measure obtained with super resolved images. Moreover, in semi-global matching uncertainties propagate and a well defined peak leads more often to a better result [42].

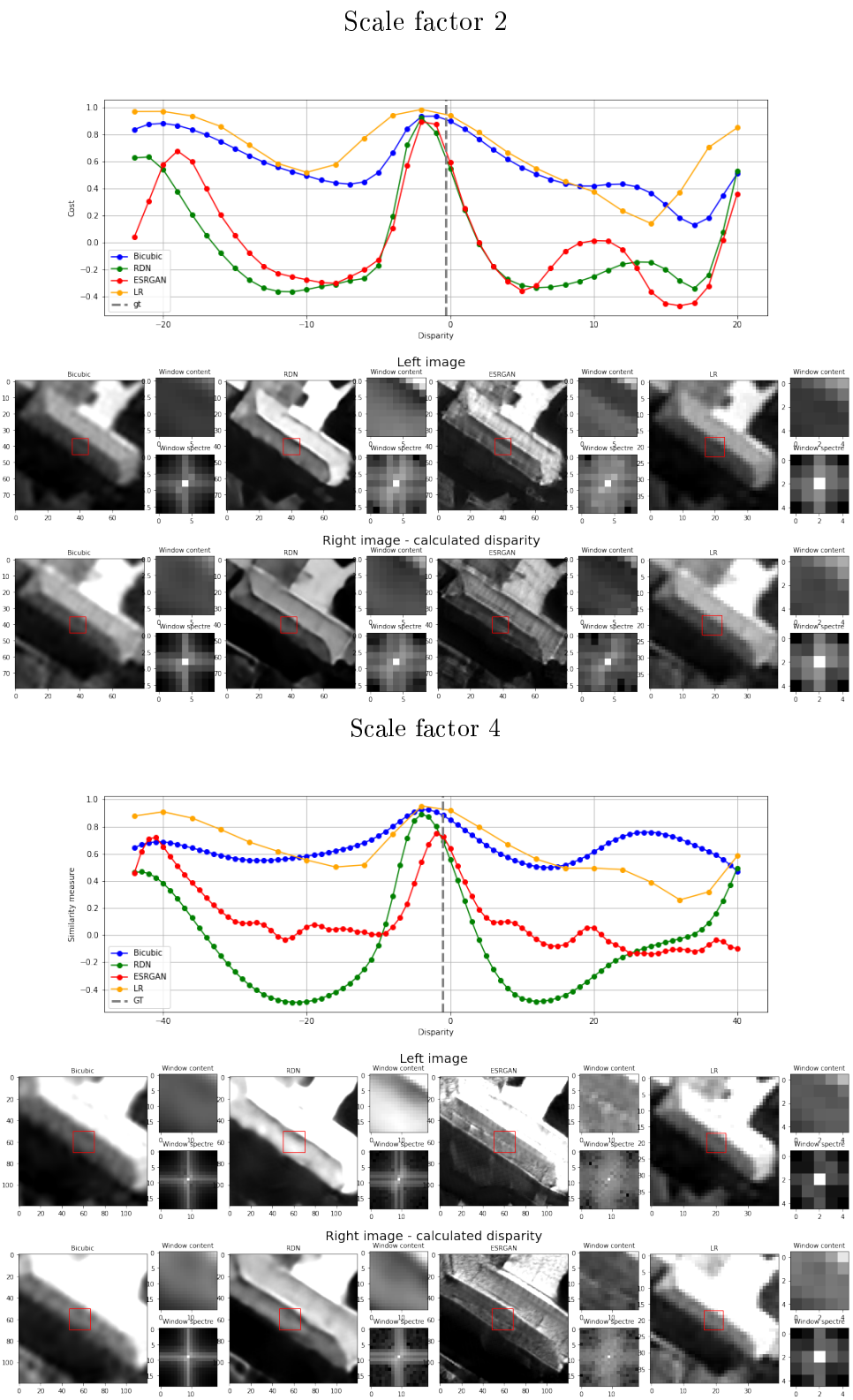
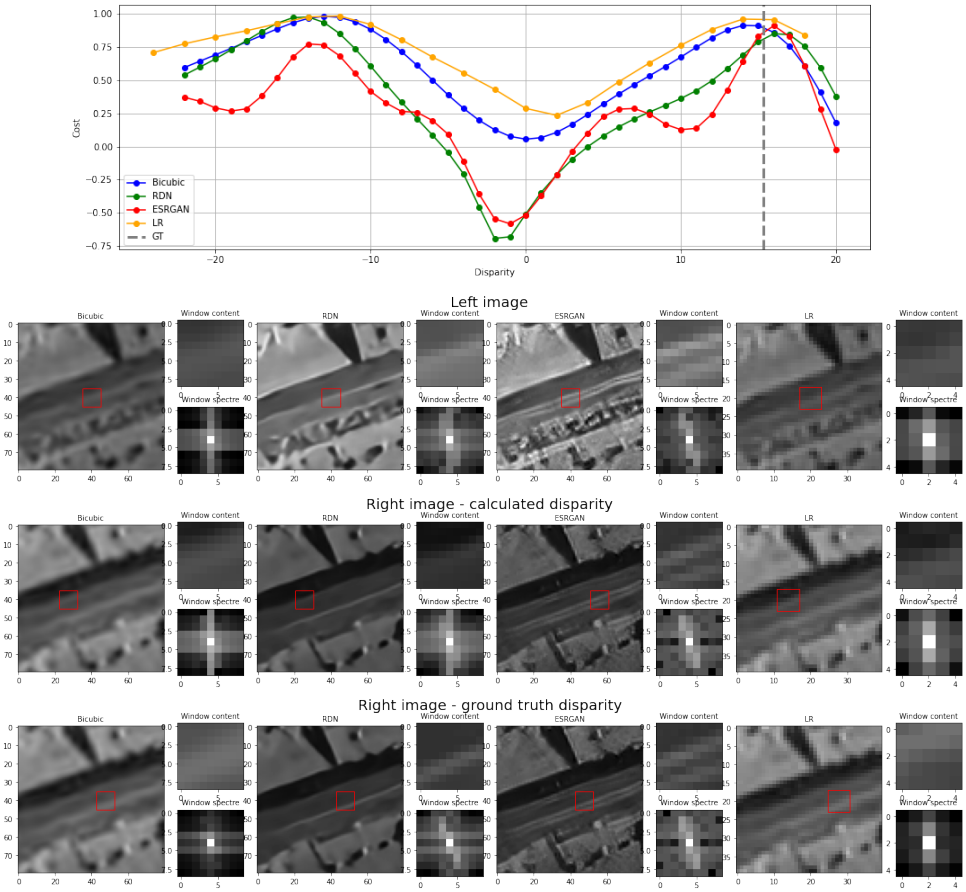


Figure 5.2: Roof sides' edge. Detail from Montpellier dataset, red band ZNCC matching cost.

When passing to scale 4, things become even more interesting. The bicubic cost profile maintains its very shape as expected. Indeed, as already recalled, bicubic interpolation can only introduce more blur and hence lead to the very same matching, just at a lower sampling distance. Less obvious is the fact that RDN zoom 4 profile completely recalls the zoom 2 one. However, this behavior is actually coherent with the conclusions highlighted for chapter 3: the trained RDN was not capable of add anything significant when increasing the scale factor, acting much like an interpolation. This can be seen in both spectre of the patch and cost profile. Finally, ESRGAN profile is relatively changed between the two zooms. It is remarkable that, even if there's a higher difference between maximum and minimum value for scale factor 2, we may prefer the zoom 4 profile because it strongly flattens almost all the other measures leaving less room for concurrent maxima. This fact means that matching relevant information was added by enhancing the scale factor, confirming the impression that ESRGAN left in chapter 3, where we observed how additional information (realistic or synthetic) is introduced by this architecture for a stronger scale factor.

Another interesting example that supports these hypothesis id shown in figure figure 5.3 and it covers the very same area shown in 4.8. We remarked 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. One could expect that this is due to less blurred building walls on the street sides, but there's likely another reason. This street is characterized by a traffic line which acts as an important clue for the matching algorithm. The SR networks have the capability to sharpen this line which becomes a more recognizable feature, as the spectra confirm. This translates into a better definition of the two maxima of the cost profile with respect to LR and bicubic curves. Moreover, it's interesting to notice that ESRGAN outperforms even RDN in this case, since its correct maximum (the one corresponding to the ground truth) is well higher than the wrong one, whereas for the other methods the attribution falls to the wrong maximum, taking the change of contrast due to shadow light transition instead of the one marked by the line on the asphalt. This superior performance can be associated to the fact that the line is not only sharpened but also doubled, adding additional information. As there's no ground truth in HR for these images, we don't know completely whether this is an artifact or not. In any case, this tendency of ESRGAN of forcing high frequency details even where there aren't there, might sometimes be beneficial for a matching algorithm.

Scale factor 2



Scale factor 4

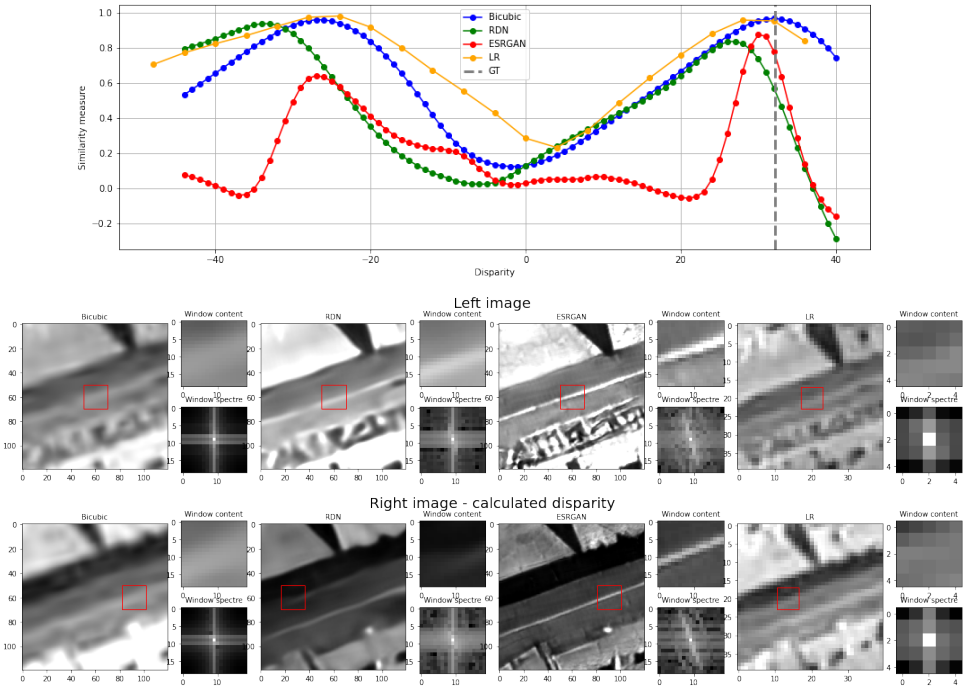


Figure 5.3: Traffic line in a narrow street. Detail from Montpellier dataset, green band ZNCC matching cost.

The double line actually disappears when passing at scale factor 4, but again this line is key for understanding the observed enhancement in the DSM. Its impressive definition in the ESRGAN image fairly leads to a cost profile in which we are very confident as its shape is almost ideal. These cost profiles confirm what seen a little bit earlier in the paragraph for figure 5.2, i.e. that for 1D features the ESRGAN contribute is significant (and in can be increased with the scale factor), while bicubic and also RDN convey more or less the same information which in these cases cannot guarantee a confident match. This example also supports the hypothesis that stronger spectral structures information can be helpful for DSM production in urban areas, as a the ESRGAN is rich in 1D information.

Uniform and textured areas

In figure 5.3 a probable artifact allows ESRGAN to better resolve the matching. But when artifacts are not coherent between left and right image, this leads more easily to matching troubles. Figure 5.4 illustrates this phenomenon. The pixel and the surrounding window corresponds to a uniform portion of a roof. Here, some not interpretable detail in the low left image, are turned by the networks into a texture, very evident in the ESRGAN image. High frequencies are forced and this results in a random pattern. In the right SR images such an artifact is not present and the roof surface is pretty smooth. Hence, Pandora cannot find the matching window, it simply doesn't exist. This results in a globally lower similarity measure and more uncertain cost profiles. In particular, ESRGAN profile has a limited range and the correct maximum is not really distinguishable. The correlator then looks for windows characterized but some sort of high frequency details, such as the contrast between the illuminated and shodowed sides of the roof, failing to find the right disparity, unlike the other methods that are closer to the ground truth. When passing at scale factor 4, this issue persists as it is again present in the ESRGAN window. The cost profile is really confused and the disparity taken is wrong as it corresponds to the roof edge, where high frequencies in fact are present. Again we can state that the bicubic cost is basically unvaried across the scale factors while in this case, RDN shows the better performance when it comes to matching as even if the profile doesn't exclude other maxima, still it highlights the right disparity. This means that, unlike what seen in the other examples, also RDN has the capability of introducing new high frequency information for a lower GSD, that in this case was beneficial for stereo reconstruction.

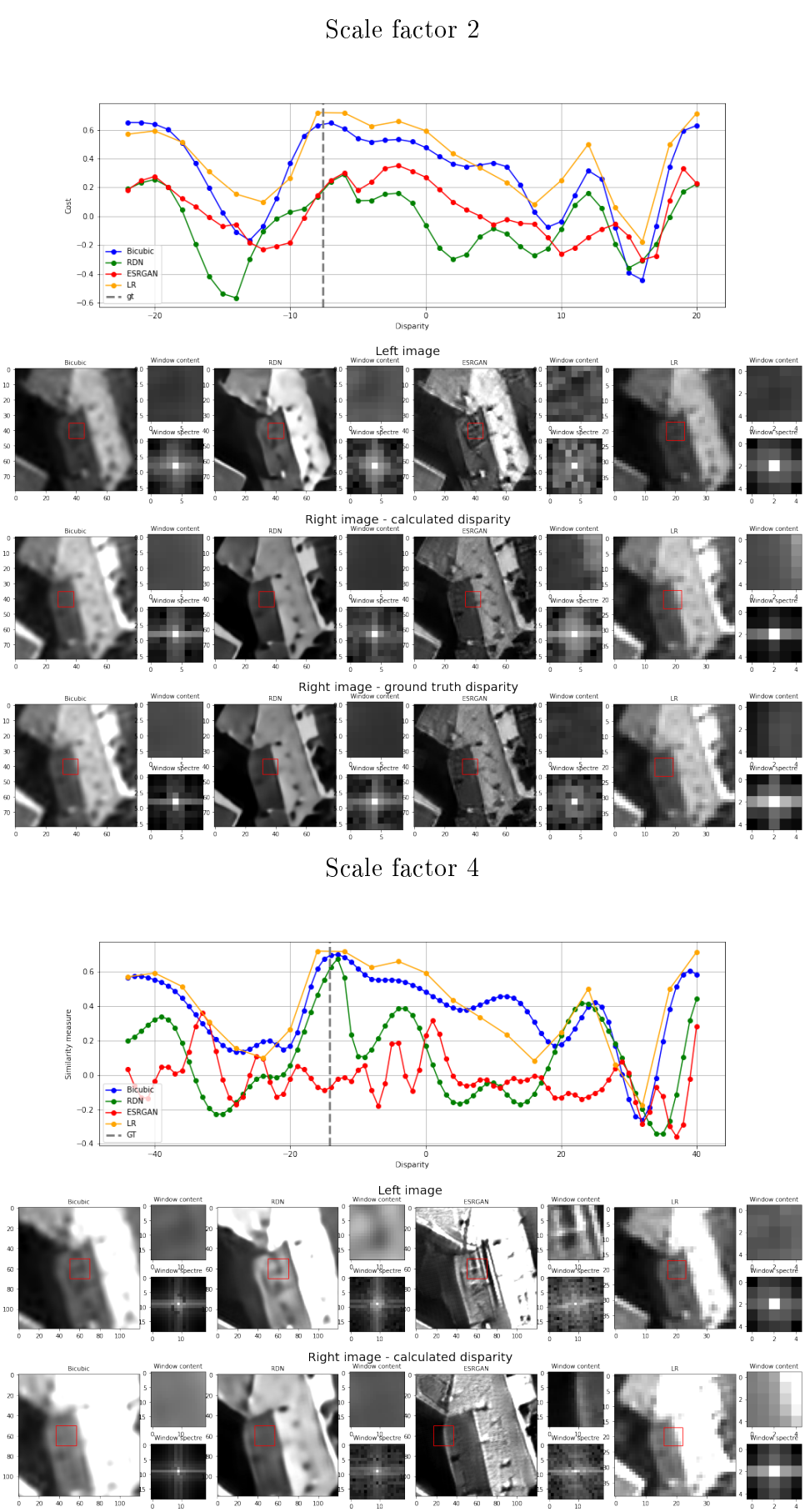


Figure 5.4: Generation of artifacts on roofs. Detail from Montpellier dataset, red band ZNCC matching cost.

Scale factor 4

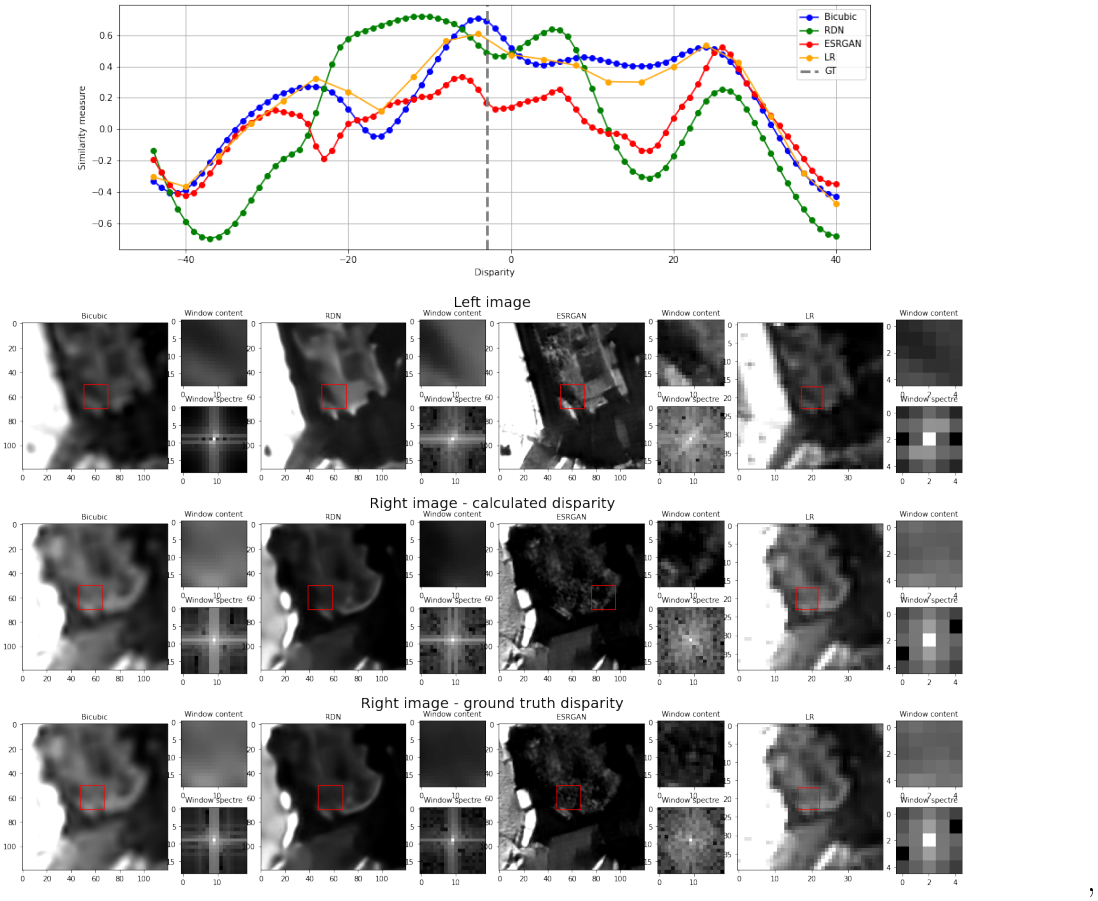


Figure 5.5: Stereo matching example on a tree. Detail from Montpellier dataset, red band ZNCC matching cost.

Fig. 5.5 shows a clear example of how an untrustworthy artificial intelligence can be detrimental when insert in a complex process and it reconnects to the ESRGAN hallucination phenomenon (Fig. 3.12, 3.13, 3.14). The analyzed pixel belongs to a tree, which, because of its uncommon contrast in the left image, is mistaken for a building by the networks. This is evident especially in the ESRGAN case where on the left image we can figure the roof and the façades of a building, while on right we clearly see a tree. As one can expect, ESRGAN similarity measure prediction is totally out of the target as its maximum lies elsewhere. RDN also shows inconstency between left and right patch and a consequent wrong estimation. LR and bicubic image generated cost profiles, although not really discriminatory, assess fairly well the position of the real optimum.

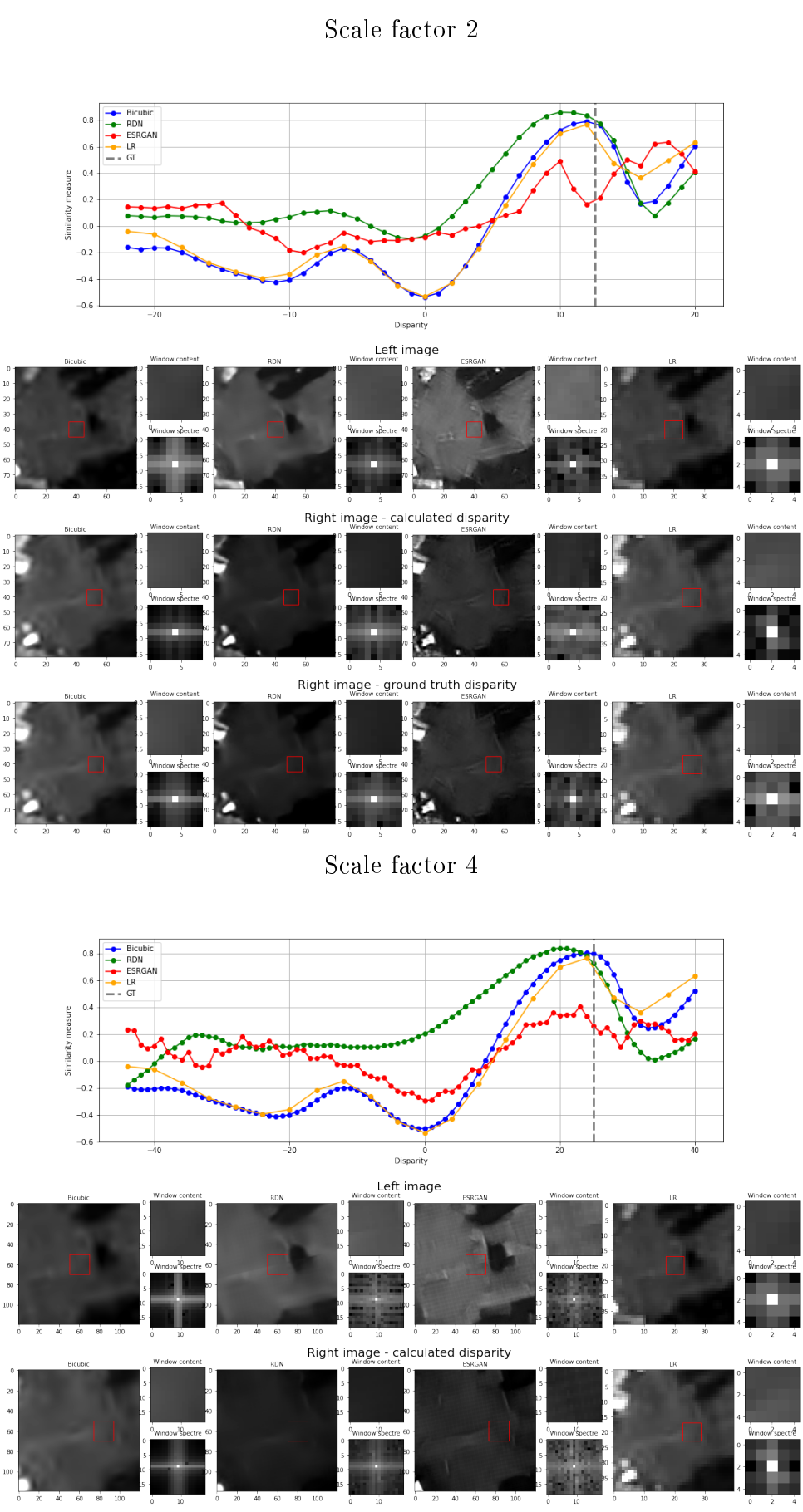


Figure 5.6: Stereo matching example on a square. Detail from Montpellier dataset, red band ZNCC matching cost.

As already pointed out when illustrating figure 4.10 4.11, in uniform zones the networks cause noisier DSMs. This is because 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 the pixel shown in figure 5.6 isn't well matched by all methods, although enough information is present originally to guess the actual optimum in LR and bicubic interpolation cases. The RDN labeled function resembles LR and bicubic profiles but it is biased, perhaps because radiometry is too much modified. ESRGAN matching presents a very flat profile that even shows a local minimum in correspondence of the real disparity. This induced by the presence of artificial details in the looked area and it shows how in this case high frequencies are introduced incoherently between left and right images. Moreover, it is interesting how ESRGAN stereo couple produces the only profile that changes considerably when zooming to a factor 4, suggesting again that RDN has less capability to enforce further information for stronger scale factors.

5.3. Wrap-up

The following main conclusion can be drawn by this cost profile analysis :

- Contrasts are enhanced by the SR networks, which leads to more discriminative cost profiles for pixels whose windows include discontinuities such as light-shadow thresholds (Fig. 5.2), edges (Fig. 5.1) or distinct lines (Fig. 5.3). This is coherent with the hypothesis made in 1.3, i.e. a more filled spectrum should improve the photogrammetry algorithm in correspondance to discontinuities. This doesn't always translate to a better DSM but we can observe it distincly in some cost profiles.
- High frequencies reconstruction by SR networks is not necessarily consistent. This leads to artifacts (especially for ESRGAN) which can be beneficial, when coherent between left and right image (Fig. 5.3), or detrimental when not present in both images (Fig. 5.4).
- Super resolution doesn't seem to improve matching of uniform zones (Fig. 5.6). On the other hand, it might add further noise to DSMs because such patches are further homogenized (RDN) or inconsistely textured (ESRGAN).
- Cost profile analysis finally allow to find the 2D results in the stereo pipeline, meaning that the characteristics of the successfully super resolved images presented in chapter 3 could be seen in the shape of these curves.
  - Bicubic interpolation doesn't add much relevant information with respect to

LR as the two profiles are often similar, but it increases the sampling which can be benificial

- RDN can be seen as a good interpolator that is powerful in well defining the disparities and it remains fairly coherent with the reality
- ESRGAN, thanks to its peculiar architecture and loss function, force high frequency information but in an uncontrolled way and this can lead to outperform in matching (Fig. 5.3 5.2) as well as to wrong estimations due to the inconsistency of the information (Fig. 5.4). In a DSM, we see mostly this second contribution as the noise is consistently increased as shown in Chap. 4.

6 | Conclusions

The thesis is placed in the general context improvement of an existing photogrammetry 3D process based on stereo matching. In particular, the objective is to explore some tracks for enhancing CS product Pandora hence CNES/CS CARS pipeline. This is relevant for the more general framework of the CO3D mission, whose processed data will depend from the reliability of these softwares under development, as well as for the upcoming products of AI4GEO consortium.

A literature basis has been delineated on the super resolution topic in the beginning of the internship (chap. 2). SR has a long story in imaging and not only DNN exist for its resolution. In either case, deep learning methods represent the current state-of-the-art and might be even easier to implement, given the great amount of open source material, therefore there's no specific reason for using a traditional technique. Nowadays, the focus in SR is on GANs and hybrid methods. In the field of remote sensing, the tendency is to extend results of computer vision research but dealing with some additional limitations. Remote sensing SR is characterized by more a minor amount of available data (especially in HR) with respect to computer vision. This issue seems to have been overcome by the creation of larger datasets (borrowed by object detection and scene classification related works, sec. 2.2.4) and a systematical employment of transfer learning from computer vision. Instead, the approach proposed during this project relies on one hand, in limiting the target of a SR method to a well specific kind of data (VHR Pléiades images). On the other, on focusing on the generation of a realistic high quality dataset, made possible by the CSI tool and implementation of a Pléïades sensor model. The results support the applied methodology, as real Pléïades products are satisfactorily super resolved (chap. 3), and thus the implemented network could be inserted into the 3D pipeline. Yet a strong artifact production could be observed for the GAN, making it less suitable for stereo matching.

There are not many references in literature that assess the influence of single image SR for DSM generation. They confirm that denser and more detailed clouds can be generated. More importantly, they suggest some gain in quality can be achieved, so the problem of

DSM quality improvement might be converted into an easier image SR problem, where models and high quality data are widely available. A large scale experiment was conducted at this aim. Results shown chapter 4 seem to agree with literature. Moreover, in an urban area context, that is the main focus of the study, buildings and streets are slightly better reconstructed. SR network stereo pair enhanced DSMs mey be slightly more accurate but tend to deteriorate the reliability of the measure. The gain with respect to bicubic upsampling is not always flagrant and, in addition, uniform and textured zones are prone to 3D failures when the upsampling is performed through a network. For what concerns the comparison with other standard upsampling techniques (bicubic upsampling was the reference used), the underlying idea is that the high frequency offered contribution SR networks can help in reconstructing a high frequency signal such as differences in altitude in a city. Looking at DSMs generated via CARS-Pandora, though, this is not always the case. In order to account for this result, a further analysis was performed at the stereo matching step, and it is presented in chap. 5. Here, we can highlight interesting differences in the way the algorithm attributes disparities with respect to the type of input image (low resolution, interpolated or super resolved). The main outcome of this analysis is that super resolved images lead to less stereo mismatches and more defined cost profiles when matching is performed on a neighborhood characterized by high contrast. This is relevant also because more discriminator cost profiles can lead to a more accurate / reliable disparity map and thus DSMs [42]. Yet, this is counteracted by noise injection in uniform or textured areas due to artifacts and inconsistent information, that propagate throughout the stereo pipeline eventually canceling out the benefits we may have in correspondence to edges and lines. As a matter of fact, the image of a city is the composition of uniform patches separated by high frequency features, so that it might not be worth to be more precise in stereo matching on edges and lines while introducing uncertainty elsewhere.

DNNs are powerful instruments and they confirmed their potential in this study, yet they might not reliable enough for an application where accuracy is strongly needed. Indeed, SR networks modify radiometry and might generate artifacts that move away their outputs with respect to the real information. This is a factor that discourages their use in the DSM pipeline. With more mature networks or trainings, that do not introduce fake information, or by finding a way to make coherent the upsampling between left and right image, we could better exploit SR networks potential in the CARS-Pandora pipeline.

Further extension of this study can take different directions. For example, if the concern is not only the improvement on DSM quality but also some other tasks, multi task approach might be envisaged to improve simultaneously image, DSM and related application quality or performances. In [2] an unique framework for DSM refinement and roof type

classification is proposed. Moreover, considering that CO3D will take two or more views of the same scenes, one can exploit the additional information that can be encoded in the second image. This would be rather belong to multi image super resolution which is a completely different set of techniques and would require another study. Yet, some approaches aiming at exploiting multiview information to boost single image SR have been proposed [21] [28].

Moreover, during this project, a limited amount of data was used for training. Perhaps, increasing dataset size and/or usage of even more specific data (i.e. stereo acquisitions), can lead to more generalize and performing SR for the application concerned. Furthermore, we would like from a 2D image to know whether it will be beneficial for DSM before running the pipeline. Standard 2D metrics were not very useful in this case as a superior 2D metrics for the networks with respect to bicubic is to found again in 3D. Thus, it would be interesting to find or define some metrics that are directly connected to the quality of a stereo matching. Finally, enforcing coherency between left and right images could potentially limit the mismatches caused by uncontrolled artifact generation.

Bibliography

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

[2] K. Bittner, M. Körner, F. Fraundorfer, and P. Reinartz. Multi-task cgan for simultaneous spaceborne dsm refinement and roof-type classification. *Remote Sensing*, 11 (11):1262, 2019.

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

[4] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3086–3095, 2019.

[5] M. Cournet, E. Sarrazin, L. Dumas, J. Michel, J. Guinet, D. Youssefi, V. Defonte, and Q. Fardet. Ground truth generation and disparity estimation for optical satellite imagery. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:127–134, 2020.

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

[7] J. Delon and B. Rougé. Small baseline stereovision. *Journal of Mathematical Imaging and Vision*, 28(3):209–223, 2007.

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

[9] L. Dumas. La mise en correspondance d images. *CS Group*, 2018.

[10] R. Fernandez-Beltran, P. Latorre-Carmona, and F. Pla. Single-frame super-resolution

in remote sensing: A practical overview. *International journal of remote sensing*, 38 (1):314–354, 2017.

[11] D. Guo, Y. Xia, L. Xu, W. Li, and X. Luo. Remote sensing image super-resolution using cascade generative adversarial nets. *Neurocomputing*, 443:117–130, 2021.

[12] J. M. Haut, R. Fernandez-Beltran, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla. A new deep generative network for unsupervised remote sensing single-image super-resolution. *IEEE Transactions on Geoscience and Remote sensing*, 56(11):6792–6810, 2018.

[13] J. M. Haut, M. E. Paoletti, R. Fernández-Beltran, J. Plaza, A. Plaza, and J. Li. Remote sensing single-image superresolution based on a deep compendium model. *IEEE Geoscience and Remote Sensing Letters*, 16(9):1432–1436, 2019.

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

[15] H. Hirschmüller. Semi-global matching-motivation, developments and applications. *Photogrammetric Week 11*, pages 173–184, 2011.

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

[17] B. Hou, K. Zhou, and L. Jiao. Adaptive super-resolution for remote sensing images based on sparse representation with global joint dictionary model. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2312–2327, 2017.

[18] M. Irani and S. Peleg. Improving resolution by image registration. *CVGIP: Graphical models and image processing*, 53(3):231–239, 1991.

[19] P. Jagalingam and A. V. Hegde. A review of quality metrics for fused image. *Aquatic Procedia*, 4:133–142, 2015.

[20] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang. Edge-enhanced gan for remote sensing image superresolution. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5799–5812, 2019.

[21] D. Jin, M. Ji, L. Xu, G. Wu, L. Wang, and L. Fang. Boosting single image super-resolution learnt from implicit multi-image prior. *IEEE Transactions on Image Processing*, 30:3240–3251, 2021.

[22] R. Keys. Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, and signal processing*, 29(6):1153–1160, 1981.

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

[24] C. Latry and B. Rougé. Super resolution: quincunx sampling and fusion processing. In *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No. 03CH37477)*, volume 1, pages 315–317. IEEE, 2003.

[25] L. Lebègue, E. Cazala-Hourcade, F. Languille, S. Artigues, and O. Melet. Co3d, a worldwide one one-meter accuracy dem for 2025. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:299–304, 2020.

[26] M. Lebrun, A. Buades, and J.-M. Morel. " implementation. *Image Processing On Line*, 2013:1–42, 2013.

[27] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[28] L. Li, W. Wang, H. Luo, and S. Ying. Super-resolution reconstruction of high-resolution satellite zy-3 tlc images. *Sensors*, 17(5):1062, 2017.

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

[30] P. Lier, C. Valorge, and X. Briottet. Imagerie spatiale: Des principes d'acquisition au traitement des images optiques pour l'observation de la terre. *Editions Cépadues*, 2008.

[31] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[32] T. Lu, J. Wang, Y. Zhang, Z. Wang, and J. Jiang. Satellite image super-resolution via multi-scale residual deep neural network. *Remote Sensing*, 11(13):1588, 2019.

[33] W. Ma, Z. Pan, J. Guo, and B. Lei. Super-resolution of remote sensing images

based on transferred generative adversarial network. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 1148–1151. IEEE, 2018.

[34] W. Ma, Z. Pan, J. Guo, and B. Lei. Achieving super-resolution remote sensing images via the wavelet transform combined with the recursive res-net. *IEEE Transactions on Geoscience and Remote Sensing*, 57(6):3512–3527, 2019.

[35] J. Michel, E. Sarrazin, D. Youssefi, M. Cournet, F. Buffe, J. Delvit, A. Emilien, J. Bosman, O. Melet, and C. L'Helguen. 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, 2020.

[36] N. L. Nguyen, J. Anger, A. Davy, P. Arias, and G. Facciolo. Proba-v-ref: Repurposing the proba-v challenge for reference-aware super resolution. *arXiv preprint arXiv:2101.10200*, 2021.

[37] C. Nuth and A. Kääb. Co-registration and bias corrections of satellite elevation data sets for quantifying glacier thickness change. *The Cryosphere*, 5(1):271–290, 2011.

[38] M. Pashaei, M. J. Starek, H. Kamangir, and J. Berryhill. Deep learning-based single image super-resolution: An investigation for dense scene reconstruction with uas photogrammetry. *Remote Sensing*, 12(11):1757, 2020.

[39] N. Riviere, A. Amditis, A. Amiez, G. Athanasiou, J. Berggren, A. Boulch, N. Bozabalian, D. Duarte, P.-E. Dupouy, P. Escalas, et al. 3d laser imaging techniques to improve usar operations for wide-area surveillance and monitoring of collapsed buildings.

[40] G. Rohith and L. S. Kumar. Paradigm shifts in super-resolution techniques for remote sensing applications. *The Visual Computer*, pages 1–44, 2020.

[41] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. Image super-resolution via iterative refinement. *arXiv preprint arXiv:2104.07636*, 2021.

[42] E. Sarrazin, M. Cournet, L. Dumas, V. Defonte, Q. Fardet, Y. Steux, N. Jimenez Diaz, E. Dubois, D. Youssefi, and F. Buffe. Ambiguity concept in stereo matching pipeline. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:383–390, 2021.

[43] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1):7–42, 2002.

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

[45] R. Tsai. Multiframe image restoration and registration. *Advance Computer Visual and Image Processing*, 1:317–339, 1984.

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

[47] N. Weir, D. Lindenbaum, A. Bastidas, A. V. Etten, S. McPherson, J. Shermeyer, V. Kumar, and H. Tang. Spacenet mvoi: a multi-view overhead imagery dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 992–1001, 2019.

[48] J. White, A. Codoreanu, I. Zuleta, C. Lynch, G. Marchisio, S. Petrie, and A. R. Duffy. Super-resolving beyond satellite hardware using realistically degraded images. *arXiv preprint arXiv:2103.06270*, 2021.

[49] D. Yang, Z. Li, Y. Xia, and Z. Chen. Remote sensing image super-resolution: Challenges and approaches. In *2015 IEEE international conference on digital signal processing (DSP)*, pages 196–200. IEEE, 2015.

[50] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.

[51] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21 (12):3106–3121, 2019.

[52] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In *European conference on computer vision*, pages 151–158. Springer, 1994.

[53] J. Zbontar, Y. LeCun, et al. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.*, 17(1):2287–2318, 2016.

[54] D. Zhang, J. Shao, X. Li, and H. T. Shen. Remote sensing image super-resolution via mixed high-order attention network. *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

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

[56] Y. Zhang, Z. Zheng, Y. Luo, Y. Zhang, J. Wu, and Z. Peng. A cnn-based subpixel level dsm generation approach via single image super-resolution. *Photogrammetric Engineering & Remote Sensing*, 85(10):765–775, 2019.

List of Figures

List of Tables

List of Acronymes

Acronyme	Description	Page
AI	Artificial Intelligence	5
CARS	Chaçine Automatique de Restitution Stéréoscopique	10
CCD	Coupled Charge Devie	21
CMOS	Complementary Metal-Oxide-Semiconductor	18
CNES	CentreNational d'Études Spatiales	1
CO3D	Constellation Optique 3D	1
CSI	Chaîne Simulation Image	36
DDN	Deep Neural Networks	4
DEM	Digital Elevation Model	1
DSM	Digital Surface Model	1
DSI	Disparity Space Image	17
DTM	Digital Terrain Model	1
ESA	European Space Agancy	5
ERGAS	Erreur Relatif Global Adimensionnel en Synthèse	53
ESRGAN	Enhanced Super Resolution Generative Adversarial Network	40
GAN	Generative Adversarial Dense Network	29
GSD	Ground Sampling Distance	1
IBP	Iterative Back Projection	24
MISR	Multi Image Super Resolution	22
MTF	Modulation Transfer Function	46
HR	High Resolution	3
LR	Low Resolution	3

Acronyme	Description	Page
PSNR	Peak Signal to Noise Ration	52
RDN	Residual Dense Network	40
RMSE	Root Mean Square Error	52
ROI	Region of Interest	10
RPC	Rational Polynomial Coefficients	10
SAM	Spectral Angle Mapper	54
SC	Sparse Coding	25
SGM	Semi Global Matching	17
SE	Structural Error	54
SISR	Single Image SUper Resolution	22
SR	Super Resolution	18
SSIM	Structure SIMilarity index	52
UQI	Universal QUality Index	53
VHR	Very High Resolution	20
ZNCC	Zero Normalized Cross Correlation	15

Acknowledgements

I would like to thank the company, CS Group, which hosted me for this six months internship, as well as the Centre national d'études spatiales (CNES). The CNES allowed me to use their assets both in term of computational power, i.e. the access to the cluster, and proprietary software such as the CSI for the generation of the dataset, as well as the data themselves (BD Merou PELICAN database). Without these elements this work would be definitively harder and less pleasant and very likely most of the shown results would have been barely possible

I acknowledge my academic supervisor at Politecnico di Milano, prof. Marco Gianinetto, whose vaste knowledge in remote sensing gave me insight into my work. He granted a lot of his time during our calls until late evening, although this internship is physically and thematically distant from his current job.

A special mention to my tutor, Loïc Dumas, who selected me among a large array of candidates, giving me a very important opportunity for my future, as it is my first working experience in the industrial sector. Then, during the internship, he lead through many difficulties and taught me a lot not only from a technical point of view, but also from a human one.

I would like to remember my intern colleagues, Alice De Bardonnèche, Mathis Roux and Alexandre Fiche, for sharing with me their experience and who are always there for any possible difficulty but also for making working days time spent with friends.

I want to thank all the 3D team in CS Group: Mickaël Savinaud, the responsible of the skill center, who confirmed me within the team by proposing a job offer; Fahd Benatia and Jasmin Siefert, with whom I shared good part of this super resolution experience. Marina Bertolino, who supported me for the installation and utilization of the CSI; Jonathan Guinet, Yohann Steux, Aurelie Emilien, for their availability and kindness whenever I encountered some issue in photogrammetry, as well as for some tips in machine mearning matter; Quentin Fardet who provided me the basis for the cost profile analysis.

Additionally, I have to be grateful to my institutions Politecnico di Milano and ISAE-

SUPAERO for having given me the opportunity of carrying out this double degree, and for the outstanding formation I received in these 3 years.

Finally, how not to spend a word for all my friends and colleagues that I met during this unforgettable period, each one leaving me some of their values and hopefully making me a better person.