Enhancing the Spectral and Spatial Resolution of Remote Sensing Images

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

Charis Alexandros Lanaras
Diploma Engineer (Dipl.-Eng.), Aristotle University of Thessaloniki
born on 07 November 1990
citizen of Greece and Britain

accepted on the recommendation of

Prof. Dr. Konrad Schindler, ETH Zurich, Switzerland
Dr. Emmanuel Baltsavias, ETH Zurich, Switzerland
Prof. Dr. José M. Bioucas-Dias, University of Lisbon, Portugal
Prof. Dr. Devis Tuia, Wageningen University, Netherlands

2018
Abstract

Understanding the content of satellite and aerial images is one of the fundamental challenges of remote sensing research. The spectral resolution of the images is an important aid towards achieving a fine grained classification. In high spectral resolution (hyperspectral) images however, there is a trade-off between spatial and spectral information, even at the sensor level: in order to obtain a reasonable signal-to-noise ratio, sensors should have small pixels and be able to integrate over large spectral bands; alternatively, they can have fine spectral resolution, but with integration over large pixels. While increasing the spectral resolution, there is a limit on the minimum pixel size without suffering from noise. On the other hand, some multispectral sensors on satellites acquire multiple bands at different spatial resolutions. Hence, a few spectral bands suffer from limited spatial resolution, while observing the same scene. The reason for this is twofold: the limited bandwidth of the data transfer to the ground stations and the signal-to-noise ratio in narrow spectral bands. A prominent example of such a sensor is the European Space Agency’s Sentinel-2.

The objective of this thesis is to combine the spectral and spatial characteristics of the images to enhance their spatial resolution. This is done in a super-resolution fashion that uses the high frequency details from other wavelengths and fuses them to the bands with limited spatial resolution. This is, in most cases, an ill-posed problem, but special fundamental properties allow a simplification of the problem and the possibility of coming up with valuable solutions.

This thesis proposes a method to fuse a hyperspectral with a multispectral image by using spectral unmixing constraints (Chapter 3). The hyperspectral super-resolution is jointly performed with the unmixing of both images into pure reflectance spectra of the observed materials, along with the associated coefficients. As the observation model requires the relative sensor characteristics (spectral and spatial) these are also computed directly from the data.

Additionally, to deal with the multiresolution images, a new method is proposed to infer all the spectral bands of multiresolution sensors at the highest available spatial resolution (Chapter 4). To do so, a physical model of the acquisition process is introduced. The model is inverted, with appropriate regularisation, efficiently in the frequency domain. As this method uses a model it can easily be applied to any sensor with the multiresolution configuration.

Alternatively, the same problem (as in Chapter 4) is handled with a machine learning approach, focussing exclusively on Sentinel-2 data (Chapter 5). True to the statistical
learning paradigm, the super-resolution is learned solely from many training examples distributed globally, by using a convolutional neural network. The training samples are computed from Sentinel-2 images by downsampling, thus enabling access to a vast amount of data. The trained network can be efficiently applied to imagery from anywhere in a wide range of geographical regions.

Experiments on simulated and real data validate the presented methods quantitatively and qualitatively. Aerial and spaceborne images of the APEX and Hyperion sensors are used to validate the proposed hyperspectral super-resolution. In experiments with simulated and real data we observed superior results in terms of root mean squared error and spectral angle deviations compared to contemporary state-of-the-art. Moreover, Sentinel-2 is used as the main sensor to evaluate the multiresolution super-resolution. Simulations from hyperspectral data and real data are used in experiments. Quantitative results compared to state-of-the-art show a numerical improvement on the acquired results. Thus, the results of this dissertation are currently state-of-the-art and give the potential to many users of remote sensing data to improve the resolution of their final products.
Zusammenfassung


Um mit multispektralen Bildern umzugehen, wird eine neue Methode vorgeschlagen, welche alle spektralen Bänder von multispektralen Sensoren in der höchsten verfügbaren räumlichen Auflösung schätzt (Kapitel 4). Hierzu wird ein physikalisches Modell von der Bildaufnahme eingeführt. Das Modell kann mit passender Regular-
isierung effizient im Frequenzbereich invertiert werden. Weil diese Methode ein Modell benutzt, kann sie leicht für jeden Sensor mit der vorgenannten Multiauflösung Konfiguration angewendet werden.

Zusätzlich wird das gleiche Problem mit einem maschinellen Lernansatz angegangen, der sich exklusiv auf Sentinel-2 Daten fokussiert (Kapitel 5). Die Super-Auflösung wird ausschließlich von vielen global verteilten Trainingsbeispielen unter Anwendung eines „convolutional neuronalen Netzwerkes“ gelernt. Die Trainingsbeispiele sind von Sentinel-2 Bildern durch Heruntertaktung (downsampling) berechnet, was den Zugriff auf eine riesige Datenmenge ermöglicht. Das trainierte Netzwerk kann effizient auf Bilder aus fast allen Regionen der Welt angewendet werden.

Acknowledgements

My gratitude goes to all of those who contributed to the completion of this dissertation. While the list is long, it is unfortunately not possible to mention all of them here.

First and foremost, I would like to thank **Prof. Dr. Konrad Schindler** for giving me the opportunity to pursue my doctoral studies under his supervision at ETH Zurich. He has supported me during all the phases of my PhD and has shown a special interest and warm engagement in my work. His extraordinary reasoning skills, ability to simplify complex problems and diverse knowledge in computer vision are inspirational and make him an ideal “doctor’s father”. A very big thank you also goes to **Dr. Emmanuel Baltsavias**, my direct supervisor. Manos has been my mentor and we have had a good and close collaboration over the years. His critical remarks and constructive criticism have improved my thinking. He possesses a vast knowledge about remote sensing, relevant sensors and concepts. He recruited me within the RAPIDMAP project, which had us travel to Japan and supported my academic stay at the University of Tokyo. They both gave me the freedom to propose and implement my own ideas. The process of pursuing a PhD has helped me to learn a lot and to develop my research and teaching skills. It has been a pleasure working with you both!

My gratitude also goes to my PhD committee members. I want to give a big thank you to **Prof. Dr. Jose Bioucas-Dias** for taking the time to carefully review my thesis. He also hosted me in his laboratory at the Instituto Superior Tcnico in Lisbon and provided me with continuous support, both academic and administrative. This has led to a fruitful collaboration, which is an inseparable part of this thesis. A warm acknowledgement for reviewing and examining the thesis also goes to **Prof. Dr. Devis Tuia**. It is an honour to have both names on my thesis.

**Ms. Monique Berger Lande** also deserves also special recognition. She has proven extremely helpful in administrative issues and in instantly managing problems successfully. On the more social side, she has organized a numerous amount of group activities, including our favourite winter retreats!

Many of these group activities will remain unforgettable, due to the special members of the group of Photogrammetry and Remote Sensing at ETH Zurich. A special thank you goes to my friends and colleagues Katrin Lasinger, Dr. Silvano Galliani, Dr. Laura Leal-Taixe, Dr. Michal Havlena and Andres Rodriguez. Our many and interesting discussions until late in the evening, various events, long (but constructive) coffee breaks and travelling for conferences have made this experience great. Further, to each and every colleague at ETH Zurich who has also contributed: Maros, Nico, Manu, Nadine,
Mathias, Audrey, Mikhail, Jan, Javier, Wilfried, Corinne, Piotr, Pascal, Timo, Zeeshan, Sultan, David, Stefan, Jillian, Lisa and Fee. Thank you all for making the atmosphere friendly and joyful to work in. Of course, we will all remember Wilfried who did live to complete his PhD.

My acknowledgments are extended to all my friends and family for their support. My parents Tom and Cathy and my aunt Elizabeth were always there to assist me in every endeavour through the years. Importantly, a special thank you goes to my mum for carefully proofreading and correcting the final version of the manuscript. Moreover, my brother Philip has always been there to help and make me think about different aspects of my work. Finally, yet importantly, I want to thank Dimitria for her continuous support and care. She is constantly giving me strength and encourages me to follow my dreams.

Thank you all very much!
# Contents

1 **Introduction**  
1.1 Motivation ................................................................. 13  
1.2 Research Aims ............................................................. 14  
1.3 Methodology and Contributions ........................................ 15  
1.3.1 Methodology in Chapter 3 .......................................... 16  
1.3.2 Methodology in Chapter 4 .......................................... 16  
1.3.3 Methodology in Chapter 5 .......................................... 17  
1.4 Technical Relation of the Chapters ................................... 18  
1.5 Relevance for Science and Society .................................... 20  

2 **Background**  
2.1 Sensors ........................................................................... 23  
2.1.1 Hyperspectral Sensors ................................................. 25  
2.1.2 Multispectral Multiresolution Sensors ............................. 27  
2.2 Co-registration ............................................................... 30  
2.3 Hyperspectral Image Processing ........................................ 31  
2.3.1 Spectral Unmixing ....................................................... 32  
2.3.2 Endmember Extraction ............................................... 34  
2.3.3 Abundance Estimation ............................................... 36  
2.3.4 Non-negative Matrix Factorisation .................................. 38  
2.3.5 Sparse Unmixing ....................................................... 40  
2.4 Super-Resolution ............................................................ 42  
2.4.1 Multiframe Super-Resolution ......................................... 42  
2.4.2 Single Image Super-Resolution ...................................... 42  
2.4.3 Hyperspectral Single-Image Super-Resolution .................. 44  
2.4.4 Spectral Super-Resolution ............................................ 44  
2.4.5 Multi-sensor Super-Resolution ...................................... 45  
2.4.6 Pan-sharpening ........................................................... 46  
2.5 Convolutional Neural Networks ........................................ 46  
2.5.1 Forward Pass ............................................................. 47  
2.5.2 Backward Pass ........................................................... 48  
2.5.3 Applications and Architectures ...................................... 49  

3 **Hyperspectral Super-Resolution** ...................................... 51  
3.1 Abstract ........................................................................... 51  
3.2 Introduction ...................................................................... 52
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3 Related Work</td>
<td>54</td>
</tr>
<tr>
<td>3.4 Problem Formulation</td>
<td>55</td>
</tr>
<tr>
<td>3.4.1 Constraints</td>
<td>57</td>
</tr>
<tr>
<td>3.5 Proposed Solution</td>
<td>57</td>
</tr>
<tr>
<td>3.5.1 Super Resolution</td>
<td>57</td>
</tr>
<tr>
<td>3.5.2 Relative Spatial Response</td>
<td>60</td>
</tr>
<tr>
<td>3.5.3 Relative Spectral Response</td>
<td>62</td>
</tr>
<tr>
<td>3.6 Experiments</td>
<td>63</td>
</tr>
<tr>
<td>3.6.1 Datasets</td>
<td>63</td>
</tr>
<tr>
<td>3.6.2 Error Metrics and Baselines</td>
<td>64</td>
</tr>
<tr>
<td>3.6.3 Implementation Details</td>
<td>65</td>
</tr>
<tr>
<td>3.7 Experimental Results and Discussion</td>
<td>66</td>
</tr>
<tr>
<td>3.7.1 Relative Responses</td>
<td>66</td>
</tr>
<tr>
<td>3.7.2 Super-Resolution</td>
<td>69</td>
</tr>
<tr>
<td>3.7.3 Discussion</td>
<td>74</td>
</tr>
<tr>
<td>3.8 Conclusions</td>
<td>78</td>
</tr>
<tr>
<td>4 Super-Resolution of Multiresolution Images</td>
<td>81</td>
</tr>
<tr>
<td>4.1 Abstract</td>
<td>81</td>
</tr>
<tr>
<td>4.2 Introduction</td>
<td>82</td>
</tr>
<tr>
<td>4.3 Related Work</td>
<td>83</td>
</tr>
<tr>
<td>4.4 Problem Formulation</td>
<td>84</td>
</tr>
<tr>
<td>4.4.1 Subspace representation</td>
<td>85</td>
</tr>
<tr>
<td>4.5 Proposed Solution</td>
<td>86</td>
</tr>
<tr>
<td>4.5.1 Adapting the spatial regularisation</td>
<td>88</td>
</tr>
<tr>
<td>4.5.2 Implementation details</td>
<td>88</td>
</tr>
<tr>
<td>4.6 Experimental Results</td>
<td>89</td>
</tr>
<tr>
<td>4.6.1 Simulated data</td>
<td>90</td>
</tr>
<tr>
<td>4.6.2 Real Sentinel-2 data</td>
<td>92</td>
</tr>
<tr>
<td>4.7 Discussion</td>
<td>92</td>
</tr>
<tr>
<td>4.8 Conclusions</td>
<td>94</td>
</tr>
<tr>
<td>4.9 Supplementary Information</td>
<td>94</td>
</tr>
<tr>
<td>5 Super-Resolution of Sentinel-2 Images</td>
<td>97</td>
</tr>
<tr>
<td>5.1 Abstract</td>
<td>97</td>
</tr>
<tr>
<td>5.2 Introduction</td>
<td>98</td>
</tr>
<tr>
<td>5.3 Related work</td>
<td>101</td>
</tr>
<tr>
<td>5.4 Input data</td>
<td>103</td>
</tr>
<tr>
<td>5.5 Method</td>
<td>105</td>
</tr>
<tr>
<td>5.5.1 Simulation process</td>
<td>105</td>
</tr>
<tr>
<td>5.5.2 20m and 60m resolution networks</td>
<td>107</td>
</tr>
<tr>
<td>5.5.3 Basic architecture</td>
<td>107</td>
</tr>
<tr>
<td>5.5.4 Deep and very deep networks</td>
<td>110</td>
</tr>
<tr>
<td>5.5.5 Training details</td>
<td>111</td>
</tr>
<tr>
<td>5.6 Experimental results</td>
<td>112</td>
</tr>
<tr>
<td>5.6.1 Implementation details</td>
<td>112</td>
</tr>
</tbody>
</table>
## CONTENTS

5.6.2 Baselines and evaluation metrics ........................................... 112
5.6.3 Evaluation at lower scale ...................................................... 113
5.6.4 Evaluation at the original scale ............................................. 118
5.6.5 Suitability of pan-sharpening methods .................................... 121
5.7 Discussion ............................................................................. 122
  5.7.1 Different network configurations ........................................ 122
  5.7.2 Timing ............................................................................ 123
  5.7.3 Open-source publication of our models ................................. 124
5.8 Conclusions .......................................................................... 125
5.9 Acknowledgments .................................................................. 125

6 Conclusions and Outlook .............................................................. 127
  6.1 Summary .............................................................................. 127
  6.2 Discussion of Contributions .................................................. 128
    6.2.1 Hyperspectral Super-Resolution ........................................ 128
    6.2.2 Super-Resolution of Multispectral Multiresolution Images .... 128
    6.2.3 Super-Resolution of Sentinel-2 Images ............................... 129
  6.3 Discussion of Limitations ....................................................... 129
  6.4 Outlook ............................................................................... 131
    6.4.1 Future Directions .......................................................... 131

A Bibliography ........................................................................... 135

B List of Acronyms ...................................................................... 153

C List of Publications .................................................................. 157

D Curriculum Vitae ...................................................................... 159
Chapter 1

Introduction

1.1 Motivation

Most common cameras emulate the human vision in the sense that they record what a human would see, i.e., the visible spectrum. A lot of information however, that gives us a better understanding of the world, lies beyond the visible spectrum. As early as 1972, Landsat-1 the first major civil satellite earth observation mission was observing the near infrared part of the spectrum, that is especially useful for the monitoring of vegetation. Conventional multispectral sensors, usually covering only the visible and near infrared spectrum, record radiation reflected from Earth in a small number of broad spectral bands. For qualitative statements on subjects such as land coverage and the type of vegetation, these sensors deliver reliable data and information. However, quantitative information such as the provision of nutrients to crops, water quality of lakes or the identification of the mineralogy in rocks and soil demands higher spectral resolution data than that provided by conventional multispectral sensors. Conversely, hyperspectral sensors observe the sunlight reflected from Earth at many wavelengths across a wide range. High resolution measurements of the spectral distribution reveal characteristics of objects based on their chemical composition. This enables the identification of the spectral signatures of individual materials based on their molecular absorption and particle scattering.

In the past decades there has been an increased demand for higher resolution data. Depending on the application, it may be useful to increase the resolution in space or to obtain a finer quantisation of the visible or infrared spectrum. Consequently, researchers are eager to have a higher pixel density, that can offer more details that are critical in various applications, discussed in the next sections. Worldwide, there is a huge and rapidly growing demand for such geoinformation on a large scale, and hence creating a need for advanced (airborne and spaceborne) remote sensing.

The requirements of new applications are continuously increase and the currently available resolution will not be able to meet the future demands. While nowadays, there are solutions for very high resolution hyperspectral images on UAVs, large-scale coverage remains unrealistic with present technology, due to the high cost. On the
other hand, the expensive high precision optics and image sensors are an important concern in many applications. The technological advances of optics and sensors is in some aspects near the physical limits and we have reached this level, where any more improvements would mean having to sacrifice spectral, spatial or temporal resolution.

For imaging sensors actually, at the sensor level, there exists a trade-off between spatial, spectral and temporal resolution. For a given, small spectral bandwidth, one cannot drastically decrease pixel size (for better spatial resolution) as the light arriving at the sensor decreases and reduces the signal-to-noise ratio (SNR). The increased shot noise observed, degrades the image significantly. The effect is amplified for wavelengths that are affected more by the atmospheric absorption of the signal. On the other hand, for broader spectral bands the amount of light arriving is much larger, thus the pixel size can be reduced, but at the cost of lower spectral resolution. One could overcome the observed weak signal by integrating for a longer time, but this requires a static camera and scene, something that is not realistic for the moving aerial and spaceborne sensors. Thus, this physical constraint is difficult to overcome.

Further, an additional limitation is observed for some multispectral sensors on satellites with multiple (different) spatial resolutions. The main reason for the varying spatial resolutions is to reduce the amount of data that needs to be transmitted to the ground, as recognised by [Trautner (2011)]: “Data rates and data volumes produced by payloads continue to increase, while the available downlink bandwidth to ground stations is comparatively stable”. The bottleneck of slow data transmission to the ground is also difficult to overcome.

The two limitations mentioned above end up constituting very similar problems. Therefore a new approach towards increasing spatial and spectral resolution is required to overcome these drawbacks arising from the sensors and optics manufacturing technology.

1.2 Research Aims

With the limitations already mentioned in mind, this thesis investigates various aspects at the interface between enhancing the spectral and spatial properties of remote sensing images, by using software techniques and methodological solutions. In this context, this thesis explores the following research questions:

- The aim of the project is to combine high spatial and high spectral resolution images and analyse them in an integrated fashion, exploiting the complementary properties of the two data sources. Finding a way to combine high spatial and high spectral resolution is a fundamental objective of remote sensing research. It is an important unresolved question how to manoeuvre higher spectral resolution towards a main unresolved problem of computer vision, semantic understanding of images. The fundamental question is how to best combine the two data sources into an integrated product with the spectral detail of modern imaging spectroscopy and the geometric resolution of recent mapping cameras.
1.3. METHODOLOGY AND CONTRIBUTIONS

• A further question which arises is whether it is possible to apply a similar concept to multi-resolution multispectral images by exploiting the structure of the image. The aim is to reveal information that is not evident, by integrating information from all the bands using appropriate methods to compute all spectral bands at the highest spatial resolution of each sensor. Is there any advantage in introducing a new concept compared to existing techniques, e.g., pan-sharpening? The motivation for our research is that the increased spatial resolution brings practical advantages and fulfills the user demand by having a homogeneous resolution for both manual and automatic processing.

• Finally, an interesting question to ask is: Can modern deep learning utilise the abundance of remote sensing data available to learn a super-resolution only from training examples? In contrast to a physical model, the goal is to learn a super-resolution on a reduced scale from simulated examples and apply it directly to unseen data. The aim of a pure machine-learning driven approach is to generalise well on images globally without having a bias towards specific land-cover types.

By implementing the integrated image analysis we hope to improve the automated extraction of semantic information, leading to the estimation of more object classes and object attributes and with a higher accuracy. This is relevant for various applications, from land cover estimation to multi-temporal analysis and change detection, environmental monitoring, agricultural studies and estimation of vegetation properties, and the determination of essential climate variables from satellite data. The applications and benefits of geodata from imagery are diverse, e.g., for the sustainable development of limited land resources, land management and planning, environmental and climate monitoring, all need detailed information about the state and development of the human habitat.

Putting the objectives of the dissertation into context, it becomes evident that the research questions are not only relevant for airborne and spaceborne remote sensing, but also beyond. Sensors of increased spectral resolution can now even be placed on low cost devices e.g., UAVs or smartphones, where one cannot afford to install a high resolution sensor. It is worth investigating the potential of the integrated analysis of high spectral and high spatial resolution images also for this purpose.

1.3 Methodology and Contributions

This thesis is written as a cumulative dissertation, in accordance with the ETH Zurich, Department of Civil Environmental and Geomatics Engineering, Doctorate Ordinance of 2013. The technical “core” chapters (Chapters 3, 4, and 5) are included as they appear in the original publications, only with minor typographical corrections. They are complemented with an introduction (the current Chapter), some background information (Chapter 2) and finally conclusions and outlook (Chapter 6).

This dissertation builds upon the main idea of using a super-resolution technique to
increase the spatial resolution of the image, i.e., reduce the pixel size and increase the number of pixels per unit area, with the help of some auxiliary information. With a super-resolution approach, existing imaging systems and images can be utilised to achieve better results.

1.3.1 Methodology in Chapter 3

In Chapter 3 a method is proposed to fuse a hyperspectral image with a multispectral one to increase the spatial resolution of the hyperspectral image. This problem is formulated as a constrained coupled matrix factorisation, which is solved with alternating proximal gradient descent. The solution is facilitated by the fact that not many materials are present in each scene and moreover, that only a few of these materials can be present in each pixel. The image pixels are factorised into a (non-orthogonal) basis of endmembers and coefficients that can be interpreted as abundances. These concepts are discussed further in Section 2.3.1.

Contributions.

1. A joint super-resolution and spectral unmixing scheme of the hyperspectral and multispectral images is proposed. Besides obtaining a super-resolved image this also leads to a physically plausible decomposition in endmembers and fractional abundances.

2. The constraints used in the optimisation process introduce a physical meaning to the solution. Namely, the reflected intensity must be non-negative, the area covered by a material cannot be negative and that each pixel (individual observation) must be described by a mixture of a few materials. Importantly, the introduced constraints improve the solution.

3. While the spectral unmixing constraints impose sparsity on how many endmembers can be present in a pixel, further sparsity can be enforced, if desired. This is done with a sparsity promoting extension in the optimisation process.

4. The spectral and spatial response are required for the super-resolution, but they are not always perfectly known. Here, they are computed directly from the data, by taking into account constraints that make the results plausible.

5. Included in the results are experiments with data that were captured from different sensors onboard the same platform (Hyperion and ALI sensors, on the EO-1 satellite). The method can cope with difficulties that occur with real data, but are not present in simulated data.

1.3.2 Methodology in Chapter 4

The unsupervised method of Chapter 4 super-resolve multispectral images from a single sensor which has multiple resolutions (usually two or three groups of different spa-
1.3. METHODOLOGY AND CONTRIBUTIONS

The aim is to infer all the bands at the highest available resolution of the sensor. This is formulated as a convex problem and variational optimisation is used to solve it. Specifically, a variant of the alternating direction method of multipliers (ADMM) is used. Due to page limits in the published conference paper, some more details about the solution are given in Section 4.9.

Contributions.

1. A new physical model is introduced for the problem. It improves upon the earlier model by allowing for a different blur in each band and different resolutions of the input data. The formulation is complex, but still tractable.

2. Additionally, the data are projected to a subspace of lower dimensions. The vast majority of the signal is contained within fewer dimensions, without a significant loss of information. In such a way the number of unknown variables is reduced and the problem is less ill-posed.

3. The problem is further regularised with the spatial details observed in the high resolution bands. This novel spatial regularisation benefits from the knowledge of the location of the high frequency details and can be used for the deblurring of lower resolution bands.

4. Most of the operations to solve the problem are formulated as convolutions. Thus the solution and inversion of the model is efficiently computed in the frequency domain with big computational savings.

1.3.3 Methodology in Chapter 5

Chapter 5 deals with the same problem as Chapter 4. Here however, a pure machine learning approach is employed. A convolutional neural network (CNN) is trained to perform an end-to-end upsampling just from training images. The training images are simulated by downsampling in order to generate ground truth data for training. Due to the nature of supervised learning though, the network is tailored specifically to the specifications of Sentinel-2, i.e., the number and resolution of bands and their spectral properties. In principle, however, the method is generic, in the sense that it can be retrained for an arbitrary multiresolution multispectral sensor setup.

Contributions.

1. The mapping learned at lower resolution can also be applied to a different scale if the resolution difference is not big. We demonstrate that, even with this generalisation across scales, the method significantly outperforms previous state-of-the-art methods.

2. Simulating the input data to the network by downsampling enables the use of virtually an infinite amount of training data. This is especially important for the data hungry CNNs.
3. The network model is trained with samples selected globally over a wide range of geographical locations and multiple seasonal appearances. Hence, it generalises well across different climate zones and land-cover types, without the need to be retrained.

4. The present method is somehow similar to pan-sharpening. In a critical comparison to pan-sharpening methods however, all super-resolution methods outperform them by a big margin implying that pan-sharpening is not suitable for the super-resolution of Sentinel-2 images.

5. The pretrained network weights are made public to the community and can be refined for specific target regions.

6. The method is very efficient during inference. Modern CNNs have been optimised to run efficiently on powerful GPUs (Graphics Processing Units). In this way a Sentinel-2 tile of $110 \times 110$ km is processed in $\approx 3$ min, a speed-up of $>40 \times$ compared to existing methods using CPUs (Central Processing Units).

### 1.4 Technical Relation of the Chapters

This thesis is the result of a 4-year-long project. Even though the major technical contributions (Chapters 3, 4 and 5) were developed over this long period, the publications arising from this research were published shortly before (or after) the thesis itself was presented. Therefore, given the recent review of the literature included in each chapter, it was not considered necessary to include a separate literature review at this point. The reader is referred to the related work presented in the individual “core” chapters. In the following, the technical relations between the three “core” chapters are highlighted.

**Relation of Chapters 3 and 4.** Even though the two chapters might appear dissimilar at first sight, their inputs and objectives are very similar. In both chapters there are a set of high spatial resolution bands and a set of low spatial resolution bands, all with variable spectral properties (observed wavelengths). The goal in both, is to perform a joint analysis and increase the spatial resolution of the low resolution inputs, even if they observe different (either broader or finer) parts of the spectrum.

A physical observation model is employed to solve both. In both models, a subspace is exploited to express the data in a reduced dimensionality; the data are thus factorised in some basis and the respective coefficients. Due to the high spectral resolution of the data used in Chapter 3, a physical meaning is attached to the subspace and a non-orthogonal basis of “pure spectra” is used to express all the pixels, with spectral unmixing constraints. With the lower spectral resolution data in Chapter 4, an orthogonal basis is extracted (singular value decomposition), that guarantees to retain the great majority of the signal energy.

Additionally, both models require knowledge of the relative spatial blur (point spread function) of the bands/images with respect to the other inputs. With multiresolution
sensors (Chapter 4) the relative spatial blur of the bands is usually well known and in most cases it is reported by the provider of the data. On the contrary, in hyperspectral and multispectral data fusion, with variable inputs and different sensor combinations, the relative spatial blur has to be computed from the manufacturer’s specifications or ideally estimated from the input data.

There is however one fundamental assumption that differentiates the two models. The observation model in Chapter 3 assumes, that there must be a spectral overlap (at least partial) of the spectra observed by the two different input images. Moreover, the relative spectral response of the two sensors is required and can be estimated. This helps to express the two input images, with respect to the underlying unknown high spectral and high spatial resolution image. In Chapter 4 such an assumption is not made, even if there is a spectral overlap. In practice, an overlap is observed only very occasionally; two channels can observe the same part of the spectrum, however with different spectral and spatial resolutions, *e.g.*, bands B8 and B8A of Sentinel-2.

**Relation of Chapters 4 and 5.** These two chapters actually deal with the same objective, however a completely different technical approach is used. Building on the deep learning expertise in the Group of Photogrammetry and Remote Sensing and the rise in importance and performance of CNNs, the objective of Chapter 4 is reappraised in Chapter 5 using these innovative techniques.

In this extension (Chapter 5), the focus is set solely on Sentinel-2 data, meaning that the network model and the experiments are designed according to the specifications of Sentinel-2. This is done because the super-resolution is actually learned only from training examples and the network is expecting a predefined number of input bands. On the contrary, since Chapter 4 uses a physical model and involves no training, it is generic for all multiresolution sensors. This is especially important for very high resolution commercial sensors, for which the high cost of data makes it difficult to obtain sufficient training data.

In both chapters, some transformation of the feature space is performed. As discussed before, Chapter 4 uses a subspace that shrinks the feature dimension. This reduces the number of unknowns and stabilises the solution. On the other hand in Chapter 5 as deep neural networks are fairly robust against over-fitting, the feature space is expanded up to 128 dimensions and each of these dimensions can be interpreted as a different feature.

Finally, most of the operations in the physical model of Chapter 4 are equivalent to convolutions and are performed in the Fourier space for greater efficiency. These convolutions are based on physical principles (blurring and downsampling that happen within the sensor) and designed once by hand, using process knowledge that comes from empirical regularisation. The same sequence of these convolutions is applied until convergence. Fundamentally, this can be seen as a similar process to the learned convolutions within a CNN (Chapter 5). In the latter case, many more convolution operations are stacked on top of each other, but it is not an iterative process, *i.e.*, in total, the number of convolutions is comparable.
1.5 Relevance for Science and Society

This section includes a discussion of how the ideas introduced in this thesis can have an impact on science and society. In the greater context of the geosciences, remote sensing is becoming more and more important as a data source for earth systems monitoring and modelling, applications ranging from ecosystems and natural hazards through agriculture to spatial planning. Models and methods are being developed that need ever more data, with better spatial, spectral and temporal resolution and there is an increased interest in more details and finer granularity (spectral and spatial resolution). Unfortunately, this need for data is impeded by the limitations imposed by current technology and the economy to keep up with this demand. While previously the standard was to work with the available low resolution data, the prospect of having a software solution for increased resolution is now realistic.

Publicly available research. The methods presented in the current thesis for increased spectral and spatial resolution can be applied directly to real remote sensing images and are thus available for science as well as society. The results of the different approaches used are presented in the core chapters and the main contributions have already been discussed in Section 1.3. All the findings of the research carried out have been published and communicated to the public implementing an open access policy. Furthermore, the software code has been made available, thus ensuring that the work is reproducible and facilitating its use in future projects. The implementations of all the core chapters are public and can be found on the internet:


- Super-Resolution of Multispectral Multiresolution Images from a Single Sensor (Lanaras et al., 2017b). Code available: [https://github.com/lanha/SupReME](https://github.com/lanha/SupReME)


The code mentioned above can be modified and used in commercial and non-commercial software. In particular, the super-resolution of Sentinel-2 images (c.f. Chapter 5) could easily be integrated into the European Space Agency’s (ESA) Snap toolbox for processing Sentinel-2 data, by including the process into the full pipeline. Moreover, for commercial software (e.g., ENVI, Erdas Imagine, ArcGIS, QGIS) it could be a standard tool (as pan-sharpening) for Sentinel-2 data. It also appears straightforward to employ the method with other sensors, as long as enough data are available to train a neural network. This will hopefully enable further research that builds on the current findings.

Understanding of global dynamics. Multispectral images acquire important wavelengths that are crucial for monitoring specific environmental processes, e.g., the red
edge part of the electromagnetic spectrum, where there is a rapid change of reflectance associated with vegetation. Remote sensing helps to develop a scientific understanding of the Earth as an integrated system and how it responds to change. Moreover, it plays a vital role in predicting the variability and trends in climate, weather, and natural hazards and assists policy makers in taking decisions concerning the protection of the environment. An example is how remote sensing has been used to generate a complete land-cover map of the whole continent of Africa derived from Sentinel-2 data. In relation to the above, the higher spectral and spatial resolution developed in the current thesis could potentially enable the use of of multisresolution multispectral sensors in new, additional applications. Among some important applications are the following:

- processes occurring in the oceans and in the lower atmosphere, *e.g.*, clouds and aerosol properties (King et al., 2003)
- land cover/use (Stefanov and Netzband, 2005; Yüksel et al., 2008) and change detection (Lu et al., 2004)
- monitoring of snow and ice movements (Stroeve et al., 2006)
- tracking of active fires (Giglio et al., 2006; Roy et al., 2008)
- investigation of the hydrologic cycle and water resource management (Malenovský et al., 2012; Bastiaanssen et al., 2005)
- study of the Earth’s albedo and reflectance (Naegeli et al., 2017; Vuolo et al., 2016)
- evaluation of the Earth’s surface vegetation (Frampton et al., 2013; Hill, 2013)
- high temporal numerical weather prediction (Warner, 2010)

**Hyperspectral observations.** On the other hand, hyperspectral images can resolve spectral reflectance into hundreds of spectral bands versus the few multispectral bands. The disadvantage however is that global coverage is a far goal, because as yet there are insufficient satellites/sensors. The key importance of hyperspectral imaging is to extract geochemical, biochemical and biophysical parameters, that can then provide information on the status and evolution of various terrestrial and aquatic ecosystems.

The improvements discussed in the dissertation with respect to higher resolution of hyperspectral data can provide benefits in the subsequent processing, particularly where advanced tools for the classification of data exist (Camps-Valls et al., 2014; Tuia et al., 2015). The availability of many spectral bands, complex land ecosystems, land assets and land use can be accurately classified, by utilising the molecular absorption and particle scattering signatures that are identified. More specifically, forestry and agriculture benefit from more detailed classification of, species (Dalponte et al., 2012), the health of the vegetation (Zarco-Tejada et al., 2002) and the growth seasons (Liu et al., 2004). Further applications include mining and geology (van der Meer et al., 2017).
2012), with more accurate remote mineral exploration (Resmini et al., 1997). Better predictions of crop yield and assessments (Zarco-Tejada et al., 2005, Clevers and Gitelson, 2013), containment mapping (Pickles and Ebrom, 2017) are possible and also the monitoring of the deterioration of roofing sheets on buildings (Bassani et al., 2007). Hyperspectral data can also be used to provide information about the damage caused by pollution to plants and soil (Choe et al., 2008).

Furthermore, hyperspectral imaging has been used in the domain of computer vision. The extra information included in hyperspectral images enables applications such as tracking (Van Nguyen et al., 2010), segmentation (Tarabalka et al., 2010) and face recognition (Pan et al., 2003), document analysis (Kim et al., 2011, Padoan et al., 2008) and analysis of paintings (Elias and Cotte, 2008). It is also being used in practice, e.g., medical applications (Lu and Fei, 2014), food inspection (Wu and Sun, 2013), inspection of medical drugs (Franch-Lage et al., 2011) and forensics. More details about some important representative hyperspectral sensors can be found in Section 2.1.

Chapter 2

Background

In this chapter, relevant background information is presented, to set the stage for the subsequent technical chapters. First, an overview of sensors which are relevant to the research goals is presented in summary. Second, the prerequisite co-registration is reviewed. Third, the concept of super-resolution is presented, showing how the term is also used to refer to several different, related forms of image enhancement. Forth, basic hyperspectral image processing is highlighted, coupled to fundamental linear algebra operations. Finally, a brief overview of convolutional neural networks is given and additionally, some high-performing network architectures are discussed. Even though whole text books exist which cover the above topics in great detail, most relevant aspects in the context of the thesis are briefly presented to make the document more self-explanatory.

2.1 Sensors

This dissertation explores passive, optical remote sensing images. Thus, one is able to extract information from objects or scenes on the earth’s surface, based on their radiance, without any physical contact.

Optical spaceborne and aerial sensors capture the earth’s radiance (reflected sunlight) in various regions of the electromagnetic (EM) spectrum, see Figure 2.1. The EM spectrum can be split into multiple spectral regions. The visible (VIS) range and near-infrared (NIR) are sometimes grouped together as visible and near infrared (VNIR). A panchromatic (PAN) image records “all the colours” (from Greek) meaning typically the visible spectrum. However, sometimes part of the NIR spectrum is included. The short-wave infrared (SWIR) part of the spectrum covers the range from approximately 1.4 to 2.5 µm. This is followed as the wavelength increases (up to 15 µm) by the thermal infrared (TIR) spectrum (also called Long-wave infrared), where the observed energy is no longer reflected, but emitted from the earth. When multiple spectra of this type are obtained simultaneously from the same sensor, the images formed at different wavelengths are stacked together and called bands. The data used in this thesis are restricted to the VNIR and SWIR range of the spectrum. TIR bands, which are emissive,
Infrared Microwave Visible Ultraviolet X-ray

Multispectral

Hyperspectral

Figure 2.1: The electromagnetic spectrum as sampled by multispectral and hyperspectral sensors. Atmospheric absorption does not allow a full coverage of the EM spectrum.

are excluded because these bands depict greatly different information with different edges and texture, that would hinder their combination with the reflective features in the VNIR and SWIR bands.

Hyperspectral sensors, also called imaging spectrometers typically capture the VNIR and SWIR spectrum. Their typical characteristic is that they acquire narrow and contiguous spectral bands, up to a few hundreds. Their spectral width, measured as the Full-Width-Half-Maximum (FWHM) is usually below 10 nm. On the other hand multispectral sensors capture much broader spectral bands. A schematic comparison between the two can be seen in Figure 2.1. In the following, a few hyperspectral and multispectral multiresolution sensors are presented. A more complete overview of satellite sensors is given in the Earth Observation portal website at [https://eoportal.org/web/eoportal/satellite-missions/](https://eoportal.org/web/eoportal/satellite-missions/).

Whisk-broom and push-broom sensors. There are two typical categories of remote sensing sensors on moving platforms: A whisk-broom sensor uses a mirror to reflect light onto multiple detectors, each collecting one pixel at a time, and to sweep in a direction perpendicular to the flight path, while, a push-broom scanner, uses a line of detectors that is arranged perpendicular to the flight direction. As the carrier flies forward an image is collected one line at a time, where all the pixels in each line read simultaneously. Compared to a whisk-broom sensor the push-broom scanner receives a stronger signal because it can observe an area for a longer time. A drawback however, is that the individual detectors of a push-broom sensor can have variable sensitivity that causes stripes in the data. Most airborne and spaceborne sensors use a pushbroom design.
Table 2.1: Basic characteristics of spaceborne hyperspectral sensors. Decommissioned, current and future missions.

<table>
<thead>
<tr>
<th>Sensor</th>
<th>Platform</th>
<th>Year Launched</th>
<th>Number of bands</th>
<th>Spectral Range [nm]</th>
<th>Band width [nm]</th>
<th>Resolution [m]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decommissioned</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HYPERION</td>
<td>EO-1</td>
<td>2000</td>
<td>220</td>
<td>350-2550</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>Current</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHRIS</td>
<td>PROBA-1</td>
<td>2001</td>
<td>62</td>
<td>410-1050</td>
<td>5-12</td>
<td>18-36</td>
</tr>
<tr>
<td>Future</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRISMA</td>
<td>ISS</td>
<td>2018</td>
<td>237</td>
<td>400-2500</td>
<td>12</td>
<td>30</td>
</tr>
<tr>
<td>DESIS</td>
<td>ISS</td>
<td>2018</td>
<td>235</td>
<td>400-1000</td>
<td>2.5</td>
<td>30</td>
</tr>
<tr>
<td>EnMAP</td>
<td>ISS</td>
<td>2019</td>
<td>244</td>
<td>420-2450</td>
<td>6.5-10</td>
<td>30</td>
</tr>
<tr>
<td>HISUI</td>
<td>ISS</td>
<td>2019</td>
<td>185</td>
<td>400-2500</td>
<td>10-12.5</td>
<td>30</td>
</tr>
<tr>
<td>HyspIRI</td>
<td></td>
<td>2022</td>
<td>214</td>
<td>380-2500</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.1.1 Hyperspectral Sensors

The development of the hyperspectral imagers started initially for research purposes, but the technology has expanded to cover commercial applications. The following paragraph introduces a few, representative hyperspectral sensors that can be grouped as spaceborne and airborne. While the list is ever-growing a few important sensors for the envisaged spectral and spatial enhancement are discussed.

Satellite Sensors. Table 2.1 contains an overview of decommissioned, current and future satellite missions. Most of the sensors have a Ground Sampling Distance (GSD) of 30 m aiming at global coverage. Interestingly among these, DESIS and HISUI will be mounted on the International Space Station. At the time of the writing this thesis the only active hyperspectral sensor in space is CHRIS.

Hyperion was one of the most well-known spaceborne hyperspectral sensors. It was launched in 2000 by NASA onboard the EO-1 (Earth Observation-1) satellite and was decommissioned in 2017. The instrument operated in a push-broom fashion, providing a spatial resolution of 30 m for all bands. It had a focal plane array that provided separate SWIR and VNIR detectors. Moreover, there was another imager onboard EO-1, the Advanced Land Imager (ALI), which obtained ten spectral bands also operating in a push-broom fashion. The spatial resolution of the nine multispectral bands is 30 m and of the panchromatic band 10 m (Pearlman et al., 2001). The joint acquisition of images from ALI and Hyperion had the advantage of (almost) simultaneous capture.

CHRIS (which stands for Compact High Resolution Imaging Spectrometer) is a hyperspectral instrument onboard the Belgian satellite PROBA-1 (Van Mol and Ruddick, 2004). The satellite is agile and able to point at distinct angles with respect to the target in a 55 degree cone (centred at the zenith). Furthermore, it can acquire up to five images of each target during the acquisition process that enables the collection
of BRDF (Bidirectional Reflectance Distribution Function) data. The sensor is capable of acquiring 62 spectral bands at 34 m GSD. However, it can be programmed to sacrifice spectral resolution for better spatial resolution and to acquire 18 spectral bands at 17 m GSD. Finally, it can acquire 37 spectral bands at 17 m GSD with a reduced swath width. In practice the sensor is configured to use one of five imaging modes. Currently, CHRIS is the only active hyperspectral sensor in space.

The Italian Space Agency ASI (Agenzia Spaziale Italiana) is also planning PRISMA (PRecursore IperSpettrale della Missione Applicativa), a medium-resolution hyperspectral imaging mission (Galeazzi et al., 2008). EnMAP (Environmental Mapping and Analysis Programme) is a future German imaging spectrometer mission (van der Linden et al., 2015). It is planned to be launched into space in 2019 and its life expectancy is 5 years. Due to the tilting mode of the satellite it will be possible to acquire the same point on earth every four days. DESIS is an Earth Sensing Imaging Spectrometer (Eckardt et al., 2015) of the German Aerospace Centre DLR (Deutsches Zentrum für Luft- und Raumfahrt). The sensor is planned to be bundled on the ISS (International Space Station). Specifically, it will be a sensor on the MUSES (Multi-User System for Earth Sensing), which is a commercial Earth-sensing precision pointing platform.

The Japanese Ministry of Economy, Trade, and Industry is developing HISUI, a spaceborne hyperspectral Earth imaging system (Matsunaga et al., 2016). It is also a future mission that will be deployed on the ISS for three years, planned to start in 2019.

Hyperspectral InfraRed Imager (HyspIRI) is a future mission of NASA (Lee et al., 2015) currently in the trial stage. The mission is comprised of two main instruments: a visible short-wavelength infrared (VSWIR) imaging spectrometer with 214 bands and a mid to thermal infrared (TIR) multispectral imager with 8 bands collecting data in the spectral range of 3 to 12 µm. The planned launch year is currently 2022 and since the mission is still in the trial stage the characteristics included in Table 2.1 might change.

Airborne Sensors. The number of airborne sensors is significantly greater than the spaceborne sensors and it is out of the scope of the thesis to list all of them. In the following, a few representative sensors are introduced, most of which have been used extensively for research and from which publicly available images exist. Table 2.2 gives an overview of the presented sensors and their basic characteristics.

Unfortunately, only limited details can be found about most of the sensors. Maybe the most revolutionary sensor, that promoted a big advancement in the field of imaging spectroscopy is AVIRIS (Airborne Visible / Infrared Imaging Spectrometer), designed by NASA. It is still one of the highest quality data sources for the scientific community. Other sensors focused on research include ROSIS (Reflective Optics System Imaging Spectrometer) of the German Aerospace Centre (DLR), APEX (Airborne Prism EXperiment) a Swiss-Belgian consortium on behalf of the European Space Agency (ESA) and HYDICE (HYperspectral Digital Imagery Collection Experiment) of the U.S. Navy. Commercial solutions are also available, e.g., the CASI (Compact Airborne Spectro-
Table 2.2: Basic characteristics of selected airborne hyperspectral sensors.

<table>
<thead>
<tr>
<th>Name</th>
<th>Year</th>
<th>Number of Bands</th>
<th>Spectral Range [nm]</th>
<th>Band width [nm]</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVIRIS</td>
<td>1987</td>
<td>224</td>
<td>400-2450</td>
<td>9.4-16</td>
</tr>
<tr>
<td>ROSIS</td>
<td>1993</td>
<td>128</td>
<td>440-850</td>
<td>5</td>
</tr>
<tr>
<td>APEX</td>
<td>2005</td>
<td>300</td>
<td>380-2500</td>
<td>10</td>
</tr>
<tr>
<td>HYDICE</td>
<td>1995</td>
<td>210</td>
<td>400-2500</td>
<td>7.6-14.9</td>
</tr>
<tr>
<td>CASI</td>
<td>1989/2002</td>
<td>288</td>
<td>400-1050</td>
<td>2.2</td>
</tr>
<tr>
<td>HYMAP</td>
<td>1996</td>
<td>126</td>
<td>450-2500</td>
<td>15-20</td>
</tr>
<tr>
<td>DESIS</td>
<td>Future</td>
<td>240</td>
<td>450-950</td>
<td>2.3</td>
</tr>
</tbody>
</table>

graphic Spectrometer) designed by the company ITRES (http://www.itres.com/). The newest CASI-1500 is a visible near infrared (VNIR) sensor which offers 1500 pixels across its field of view, achieving spatial resolutions as high as 25 cm. Most of the well-known research datasets come from these airborne sensors, i.e., Indian Pines, Cuprite, Moffett Field (AVIRIS), University of Pavia (ROSIS), Washington DC (HYDICE) and University of Houston (CASI).

**Sensors on UAVs.** Even more in number than airborne sensors, sensors on UAVs (unmanned aerial vehicles) are even harder to fully list. There are more than a dozen sensors that are light weight and suitable for mounting on a UAV. Interestingly, they have more spectral bands compared to conventional airborne sensors and still have very narrow spectral channels (typically below 10 nm). The sensors cover variable wavelengths from 400 up to 2500 nm with up to 900 spectral bands and weigh as little as 270 g (PhotonFocus, BaySpec). The majority however, weigh up to 1-2 kg. For a complete overview please refer to Adão et al. (2017). Moreover, companies such as GAMAYA (https://gamaya.com/) use hyperspectral images from UAVs for precision agriculture, e.g., early detection of weeds and better prediction of crop yields.

### 2.1.2 Multispectral Multiresolution Sensors

As mentioned in the introduction, some widely-used satellites have instruments (in some cases with multiple sensors or detectors) with multiple spectral bands, however with different spatial resolution (usually two or three groups of different spatial resolutions exist, see Table 2.3). The very positive aspect in this case is that these spectral bands are imaging the same area quasi-simultaneously, which means that very similar illumination and atmospheric conditions are encountered and other multi-temporal change problems are avoided. It also means that normally the view directions of the bands are very similar, while there is a good to excellent geometric co-registration between the spectral bands of each sensor. Examples of such multi-spectral sensors with variable spatial resolution include: MODIS, VIIRS, ASTER, WorldView-3 and Sentinel-2. As can be seen in Table 2.3, most of these sensors do not have a panchromatic band (which could possibly be used for super-resolution), the major spectrum ranges
Table 2.3: List of the major spaceborne multispectral sensors with variable spatial resolutions.

<table>
<thead>
<tr>
<th>Satellite</th>
<th>Sensor</th>
<th>No. of bands</th>
<th>Wavelengths covered</th>
<th>Spatial resolution of spectral bands [m]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terra</td>
<td>ASTER</td>
<td>14</td>
<td>VNIR, SWIR, TIR</td>
<td>15, 30, 90</td>
</tr>
<tr>
<td>Terra, Aqua</td>
<td>MODIS</td>
<td>36</td>
<td>VNIR, SWIR, TIR</td>
<td>250, 500, 1000</td>
</tr>
<tr>
<td>Suomi-NPP, NOAA-20</td>
<td>VIIRS</td>
<td>22</td>
<td>VNIR, SWIR, TIR</td>
<td>375, 750</td>
</tr>
<tr>
<td>WorldView-3</td>
<td>WV110,  CAVIS</td>
<td>28</td>
<td>VNIR, SWIR</td>
<td>0.31, 1.24, 3.7, 30</td>
</tr>
<tr>
<td>Sentinel-2</td>
<td>MS Imager</td>
<td>13</td>
<td>VNIR, SWIR</td>
<td>10, 20, 60</td>
</tr>
</tbody>
</table>

covered are VNIR and SWIR and the resolution difference between the spectral bands is a factor of 2 to 6. With the exception of WorldView-3 and Spot-6/7 data, the data from the other sensors can be downloaded and used free of charge. Details of the multiresolution sensors are discussed below.

The Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) is an imaging instrument, resulting from a cooperative effort between the USA and Japan. The sensor was launched in 1999 onboard Terra. The ASTER sensor consists of 3 separate instrument subsystems with 3 different spatial resolutions. In particular, it acquires 4 bands in the VNIR with 15 m GSD, 6 bands in the SWIR with 30 m GSD and 5 bands in the TIR with 90 m GSD. Unfortunately, in 2008 the SWIR bands were declared non-operational, because of a malfunctioning of the detectors. Thus, since then the SWIR data are marked as unusable. ASTER can also serve as a higher-resolution complement to the other Terra instruments, e.g., MODIS.

The Moderate Resolution Imaging Spectroradiometer (MODIS) is an imaging instrument (whisk-broom sensor) of NASA aboard the Terra and Aqua satellites which were launched in 1999 and 2002, respectively. Terra MODIS and Aqua MODIS jointly cover the entire earth’s surface every 1-2 days and capture data in 36 spectral bands in 3 different spatial resolutions. The sensor acquires 2 bands with 250 m GSD, 5 bands with 500 m GSD and 29 bands with 1000 m GSD, ranging in wavelength from 0.4 to 14.4 µm. From the MODIS data multiple level products are derived, which help to understand the global dynamics and processes occurring on the land, in the oceans, and in the lower atmosphere. MODIS has been succeeded by the VIIRS instrument.

VIIRS (Visible Infrared Imaging Radiometer Suite) is a sensor onboard the Suomi National Polar-orbiting Partnership (Suomi NPP) and the National Oceanic and Atmospheric Administration’s-20 (NOAA-20) weather satellites. The VIIRS sensor was first launched in 2011 to expand upon the data collected by the MODIS and AVHRR sensors. VIIRS is a whisk-broom radiometer acquiring 22 spectral bands from wavelengths of 0.4 to 12.5 µm with 2 different spatial resolutions. Five of the VIIRS bands are captured at 375 m GSD, designated as “l-bands” and sixteen are captured at 750 m GSD, also referred to as “M-bands”. One of the M-bands sensitive to the VNIR spectrum can observe nighttime lights on earth and is called Day/Night band, or DNB.

WorldView-3 (WV-3) is a commercial Earth observation satellite that was launched in 2014 and is owned by DigitalGlobe, which has an average revisit time of 1 day. The
2.1. SENSORS

resolution of its panchromatic band is 0.31 cm GSD and it is listed among the very high resolution imaging satellites, such as Ikonos (0.8 cm), QuickBird (0.61 cm), GeoEye-1 (0.41 cm), WorldView-2 (0.5 cm) and Pléiades (0.5 cm). Very high resolution satellites have in general one panchromatic band and 4 multispectral bands at lower resolution covering the visible blue, green, red and NIR parts of the EM spectrum. As discussed in Section 2.4.6 pan-sharpening is used to enhance the spatial resolution of the multispectral bands. The WV-3 Imager (WV110) however, apart from the panchromatic band of 0.31 cm GSD, acquires 8 VNIR bands with 1.24 m GSD and 8 SWIR bands with 3.7 m GSD. Commercially, the SWIR bands are only delivered at 7.5 m GSD. Moreover, onboard the same satellite additional bands are acquired, 12 bands at 30 m GSD covering the VNIR and SWIR spectrum, called CAVIS bands. The CAVIS sensor monitors the atmosphere and is used mainly to improve WorldView-3’s imagery in case of haze, soot, dust or other obscurants. Overall WV-3 acquires images with 4 different resolutions.

Sentinel-2 (S2) is an earth observation mission, part of the European Space Agency’s (ESA) Copernicus programme, which is fully operational as of 2017. It consists of two identical satellites, Sentinel-2A and 2B, which use the same sensor (push-broom style) and fly on the same orbit with a phase difference of 180 degrees. The revisit period of one satellite is 10 days and collectively 2A and 2B achieve a temporal resolution of 5 days. The maximum resolution of S2 is 10 m GSD and the swath width is 290 km. These figures compare favourably to the “competing” and complementary Landsat-8 where S2 has better spatial, spectral and temporal resolution. A comparison of the spectral bands acquired between Landsat-8 and Sentinel-2 is presented in Figure 2.2. Because of S2’s good characteristics and recency, this thesis focuses specifically on super-resolution of Sentinel-2 images. More details about the Sentinel-2 mission are given in Chapters 4 and 5.
2.2 Co-registration

In order to be able to combine different image data, *i.e.*, from different sensors, the images must be geometrically co-registered (aligned). This is an important prerequisite and a very important aspect of super-resolution, whenever data from additional sensors are required. Registration errors, even at subpixel level, can considerably disturb the subsequent integrated analysis. In common applications, the registration must be more accurate than approximately 0.3 pixels to ensure correct assignment between pixels of different sources. Assuming an equal error of 0.3 pixels in both directions, this leads to a $\approx 50\%$ overlap of the two image pixels, which we consider to be the minimum for a combined processing of the images. Moreover, especially for images with multiple bands, it is also crucial that the band-to-band co-registration is accurate. Band-to-band mis-registration occurs frequently in hyperspectral images that are acquired with different detectors (Figov et al., 2007), *e.g.*, for VNIR and SWIR. Additionally, band-to-band mis-registration can be caused by the spectral smile effect present in push-broom type sensors (Richter et al., 2011). Data providers often claim that all the spectral bands of a sensor are well co-registered, however experience shows that various mis-registration effects are not always fully compensated.

In order to compute the correct co-registration, different measures of similarity can be used; a few examples are the sum of squared differences (SSD), normalised cross-correlation (NCC), phase correlation, and mutual information. Besides that, most software tools compute a global transformation for the whole image, *e.g.*, simple translation, similarity or affine. This is sufficient for typical scenarios when data are simultaneously captured from the same platform. Nevertheless, the challenging case of spatial co-registration occurs when images are acquired from different viewpoints, as due to the terrain a global transformation would be very complex. Consequently, co-registration is often performed after the generation of orthoimages, if a Digital Elevation Model (DEM) is available. The co-registration is even more challenging, when the inputs are of different spatial resolutions and different spectral coverages, as is the case for most super-resolution problems.

In the present thesis, where necessary, we have used co-registration methods previously developed in the Group of Photogrammetry and Remote Sensing of ETH Zurich. The work builds on existing image matching methods, by introducing matching error detections and reaches accuracies better than 0.2 pixels. Lanaras et al. (2014) compared different methods to co-register images based on effective phase correlation and mutual information. It is noted for completeness that the co-registration methods were originally developed for images with large differences in appearance (including different modalities such as Synthetic Aperture Radar – SAR). The main focus is a co-registration refinement, as with most geodata the approximate co-registration can be easily acquired or estimated. These methods rely on small patches extracted from the whole image. A global transformation is then fitted to the set of corresponding points. This decreases the complexity and also allows the focus to be on a complex global transformation from many individual points. The data used in this thesis do not exploit serious differences for standard co-registration, however implementations and
practices from Lanaras et al. (2014) are used.

2.3 Hyperspectral Image Processing

Working with hyperspectral data requires organizing and analyzing data in high-dimensional spaces. This greater dimensionality compared to multispectral images significantly increases the volume of data to be stored, transmitted, and processed; and introduces new challenges to conventional image processing techniques. Moreover, the large quantities of available data in the near future (e.g., the estimated 4.5 TB of daily data from HyspIRI) further indicates the need for efficient hyperspectral image processing.

Hyperspectral images exhibit a high degree of correlation between bands and hence lie on a lower-dimensional manifold within the high-dimensional data space. This leads to problems, e.g., when training a classifier it unnecessarily increases the required number of training samples. Working in high-dimensional data spaces leads to a number of technical difficulties, because the space will inevitably be very sparsely populated and distance measurements become meaningless. This phenomenon is termed the curse of dimensionality.

One fundamental observation however, that helps alleviate the high dimensionality of hyperspectral images is the fact that they are spectrally and piece-wise spatially smooth: the values of neighbouring wavelengths and spatial locations are highly correlated. Thus, low-rank modelling and sparse reconstruction can be used to exploit this fact. I.e., most information in hyperspectral images can be represented by a few dimensions. In practice, not all hyperspectral bands carry significant information, but the significant bands change depending on the application.

Dimensionality reduction (DR) is the process of specifically identifying and eliminating redundancies of hyperspectral data, with the main goal of retaining as much spectral information as possible. DR can be especially useful for storage, transmission, classification, and visualisation of hyperspectral remote sensing data. There are many ways to perform DR and they can be grouped in supervised and unsupervised methods. The supervised methods include manual band selection, according to the application or rely on labelled samples to retain important features. On the other hand, unsupervised methods are only based on the statistics of the images. One important example is the principal component analysis (PCA) that captures the maximum variance in the data and decorrelates the bands. The minimum noise-fraction (MNF) transformation (Green et al., 1988) optimises the SNR to obtain the reduced features. These representations, however, do not have any physical interpretation, see Figure 2.3. For some applications however, DR can eliminate important information, e.g., in target detection. In reconnaissance and surveillance applications hyperspectral images are used to detect and identify very specific targets. These targets are usually represented in only a few pixels and have small power compared to other signals in the image, and thus would be omitted as noise.
CHAPTER 2. BACKGROUND

Figure 2.3: Dimensionality reduction of a hyperspectral image of APEX with PCA to three bands. Left: A true colour representation of the scene. Middle: A colour composite of the first three principal components. Right: A colour composite of principal components 4 to 6. Notice, how in some components details appear that are not visible in the true colour image. E.g., different patterns in the trees in the upper left corner.

Among the various challenges in hyperspectral image processing, the focus of this subsection is on the concepts used in the following chapters of the thesis. For a complete overview of recent developments and advances in hyperspectral image and signal processing the reader is referred to Bioucas-Dias et al. (2013) and Ghamisi et al. (2017).

2.3.1 Spectral Unmixing

In hyperspectral image processing pure pixels are those pixels that represent one single class/object type/material, as opposed to mixed pixels with spectral values that are a combination of two or more materials. Mixed pixels occur very frequently and are especially evident at the boundaries of objects. Mixed pixels occur for three main reasons: the low spatial resolution of the sensor, multiple reflections of the signal or intimate mixtures. In the first and second category mixing happens in a macroscopic fashion, meaning that only a few macroscopic components are involved, while in the third case mixing happens in a microscopic fashion, see Figure 2.4. The presence of mixed pixels in the data might compromise the analysis for coarse spatial resolution. When mixing occurs it is not possible to directly determine the materials present in the scene, from the measured spectral vectors. However, the rich spectral resolution can be used to unmix hyperspectral pixels and recover the information. Solving spectral unmixing requires the inversion of a mathematical model of radiative transfer, that describes the transfer of energy, as photons interact with the materials in the scene.

Linear mixing. Linear mixing is a special case of spectral mixing, that occurs within the sensor due to its limited spatial resolution (Keshava and Mustard, 2002). This implies, that the signal coming from each individual material is separable up to the moment it is recorded by the sensor (see Figure 2.4, left). The task of linear spectral unmixing is to define a set of endmembers (pure spectral signatures of macroscopic components) and the corresponding abundances that denote the fractional coverage
2.3. HYPERSPECTRAL IMAGE PROCESSING

Figure 2.4: Representation of three different types of spectral mixing. Left: Linear mixing occurring in the sensor with two or more distinct endmembers, Middle: Non-linear mixing: A combination of linear and bilinear mixing (single and double scattering) in a multilayer scenario, Right: Non-linear mixing in an intimate medium.

of each endmember in one pixel. Thus, endmembers and abundances can be seen as a non-orthogonal basis and the corresponding coefficients, respectively. As opposed to DR, one of the ways to reduce the dimensionality of a hyperspectral image is to attempt to see the image as a linear mixture of natural materials, which attaches a physical meaning to the reduced dimensionality representation. The physical meaning comes with some constraints. I.e., since endmembers are reflectance spectra, they have to be non-negative and accordingly the abundances must satisfy the abundance non-negativity constraint and the abundance sum to one constraint, since they represent fractions of a pixel. This fact is exploited extensively in Chapter 3. While the linear mixing model is an approximation of the light scattering in real scenarios, it has proven to work quite robustly in practice (reflectances are mostly low, so with repeated reflections the radiance quickly goes to zero). While significantly different from hyperspectral image classification, linear spectral unmixing can be seen as a soft classification, where each pixel can belong to more than one class. The processes of estimating the number of endmembers in a scene, endmember extraction and abundance estimation are all related by the linear mixing model. These are presented in Subsections 2.3.2 and 2.3.3.

Non-linear Mixing. When the EM radiation interacts with more than one reflective material before reaching the sensor, this process is called non-linear mixing (see Heylen et al., 2014, for a review). This concept can be classified into two main concepts: multilayer reflection and intimate mixtures. Multilayer mixing occurs when the reflected signal has been scattered from multiple macroscopic surfaces/objects. This multiple reflection complicates the unmixing procedure as the finally recorded signal is an element-wise product of the individual endmembers. The bilinear mixing model is a special case of multilayer mixing, assuming that only double reflection exists and that the elementwise products of the endmembers form new endmembers (see Figure 2.4, middle). After this point the linear mixing model can be applied with the fundamental assumption, that the endmembers (and their inner products – that form the new endmembers) are known. If the signal is reflected more than two times the estima-
chapter becomes computationally inefficient, as the problem grows exponentially with the number of bounces (respectively, composite endmembers). However, we can assume that the amplitude of the signal decreases significantly after 3 or more reflections and can be neglected.

Non-linear mixing that occurs at a microscopic level is more challenging, as the materials are in close proximity and it depends on their composition (Figure 2.4, right). Hapke’s model (Hapke, 1981) provides a, data driven but physically inspired, approximate analytic solution to the radiative transfer equation and is still the most widely used approach. In non-linear unmixing the vast majority of methods require prior knowledge of the endmembers and fully unsupervised non-linear unmixing research is very limited (Heylen et al., 2011b, Heylen and Scheunders, 2012).

2.3.2 Endmember Extraction

A plethora of algorithms exist, that can identify the endmembers of a scene. However, most of them require the number of endmembers present in the image as an input. The most important methods to estimate their number are Virtual Dimensionality (Chang and Du, 2004) and HySime (Bioucas-Dias and Nascimento, 2008). The endmember extraction can be classified in three main categories:

- pure pixel assumption
- minimum volume algorithms
- statistical methods

The first two can be described as geometrical approaches, which exploit the fact that in a linear mixture the observed spectral responses form a simplex. The vertices of this simplex are the sought endmembers.

The pure pixel assumption states that at least one pure pixel of each endmember exists in the scene. Pure pixel algorithms identify a set of “pure” pixels in the data and assume that these are the vertices of the simplex, see Figure 2.5 (left). Even if the assumption of having pure pixels of all endmembers does not always hold, in practice they work surprisingly well. The first well-known example is the pixel purity index (PPI) proposed by Boardman et al. (1995). PPI projects the spectral data onto many random unit vectors (skewers). Pixels with the largest count of extremes in all projections are associated with pixel purity and are noted. A second well-known method is Vertex Component Analysis (VCA) proposed by Nascimento and Bioucas Dias (2005). VCA iteratively projects the spectral vectors onto a direction orthogonal to the subspace spanned by the endmembers that have already been identified. The new endmember signature is determined as the extreme of the projection. The idea of these two methods is to find the extreme values of the projection onto some subspace, which has also been used in Successive Volume Maximisation (SVMAX) by Chan et al. (2011). Furthermore, another way to exploit the pure pixel assumption is to maximise the volume of a simplex with a variable set as vertices (endmembers). N-FINDR reported by Winter (1999) was the first such approach. Similarly, another method maximising the
2.3. HYPERSPECTRAL IMAGE PROCESSING

Figure 2.5: Illustration of the concept of a simplex in linear mixing. The vertices are the sought endmembers (e₁, e₂, e₃). **Left:** Pure pixel assumption. The endmembers (pure pixels) are part of the data points (purple circles). **Middle:** Minimum volume principle. The pure pixels (green circles) are not part of the data points, but some points lie on the facets. **Right:** Highly mixed scenario. There are no points close to the facets.

The category above that utilise the geometric properties of the points in the high-dimensional space fail in highly mixed scenarios, when there are no data points near the facets, see Figure 2.5 (right). On the contrary, **statistical methods** make assumptions about the statistics of the observed spectra and can deal better with this case, by formulation of a statistical inference problem. DECA (Nascimento and Bioucas-Dias, 2012) a dependent component analysis algorithm models the abundances as mixtures of Dirichlet densities, but quickly reaches its limits when the number of endmembers grows. A basic disadvantage of statistical methods is that they have higher computational complexity than the geometrical methods and do not seem to be necessary in typical remote sensing settings.
2.3.3 Abundance Estimation

The estimation of endmembers is often interrelated with estimating the abundance vector, a process sometimes referred to as the inversion step. An important aspect of the inversion step in the linear mixing model is the inclusion of two constraints coming from elementary physics: full additivity and non-negativity. It is assumed, that the endmembers present in the scene are known and given. The main idea behind the inversion is to minimise a distance metric, which is done with direct least squares estimation, gradient descent approaches, or basis pursuit, including variants based on variational calculus.

To keep the notation consistent throughout the thesis, the same formulations as in Chapter 3 are used in the following. Let \( z \in \mathbb{R}^B \) be an observed pixel of an image with \( B \) spectral bands. The linear mixing model is:

\[
    z = E a + n ,
\]

where \( E \equiv [e_1, e_2, \ldots, e_p] \in \mathbb{R}^{B \times p} \) is a matrix containing \( p \) pure spectral vectors (endmembers) as columns, \( a \in \mathbb{R}^p \) is the fractional abundance vector and \( n \in \mathbb{R}^B \) is a vector of additive noise assumed to be independent and identically distributed. This mixing model can be easily extended to multiple pixels \( N_m \) of an image in the following way:

\[
    Z = EA + N ,
\]

where \( Z \in \mathbb{R}^{B \times N_m} \) is a given image with the pixels ordered in lexicographical order concatenated in the second dimension, \( A \in \mathbb{R}^{p \times N_m} \) a matrix with the per-pixel abundances ordered in the same way as \( Z \) and \( N \in \mathbb{R}^{B \times N_m} \) a matrix of additive noise with the same ordering. Moreover, the abundance fraction matrix is \( A \equiv [a_1, a_2, \ldots, a_{N_m}] \in \mathbb{R}^{p \times N_m} \), where \( a_i \in \mathbb{R}^p \) represents the fractional abundance vector of the \( i \)th pixel. For the mixing model to hold, the number of endmembers \( p \) in the scene must be smaller than the number of spectral bands \( B \), i.e., \( p < B \). If this inequality does not hold, it means that at least one endmember can be expressed as a linear combination of the remaining ones and thus should not be considered as a separate endmember. This could happen when the number of spectral bands is limited and two similar endmembers are confused by the limited spectral resolution.

If one assumes that there is no noise in the image, the inversion could be done by using unconstrained least squares and perfect unmixing results would be obtained by minimising the following expression:

\[
    \min_a \frac{1}{2} \| z - E a \|_2^2 ,
\]

where \( \| \cdot \|_2 \) is the vector L2 norm and the solution of eq. (2.3) is

\[
    a = (E^T E)^{-1} E z ,
\]

given that \( E \) is of full column rank. A similar formulation can also be written for an
2.3. HYPER SPECTRAL IMAGE PROCESSING

image of form $Z$, by minimising the Frobenious norm of the matrix as

$$\min_A \|Z - EA\|^2_F,$$

which is equivalent to solving the L2 norm equation for each pixel individually. The Frobenious norm of a matrix $X$ is the root of the squared sum of all individual elements as:

$$\|X\|_F = \sqrt{\sum \sum |x_{ij}|^2} = \sqrt{\text{trace}(X^\top X)}.$$  

The solution in this case is $A = (E^\top E)^{-1}EZ$. In the following text the two expressions (single pixel / image) will be used interchangeably.

Nonetheless, as discussed before, hyperspectral images are inevitably contaminated with some level of noise (e.g., from the atmosphere), therefore the unconstrained inversion does not give satisfactory results. To constrain the problem and get reasonable results in the presence of noise, **full additivity** (Abundance Sum Constraint – ASC) can be considered, i.e., $\sum a_j = \|a\|_1 = 1$, respectively $1^\top A = 1^\top$ to get the per column sum of $A$ to equal 1. Note, that the dimensions of $1$, a vector of ones, has to be consistent with the dimensions of $A$. This restricts the least squares solution to lie on a hyperplane. The solution can be computed in closed-from using Lagrange multipliers,

$$a = (E^\top E)^{-1}(E^\top z - 1^\top (1(E^\top E)^{-1}1)^{-1}(1(E^\top E)^{-1}1^\top z - 1).$$

The complementary constraint, **non-negativity** (Abundance Non-negativity Constraint – ANC), does not allow a closed-form solution and has to be solved iteratively. It is however convex, meaning that it has a unique solution. The solution to eq. (2.3), requires all elements of $a$ to be non-negative, i.e., $a_j \geq 0, \forall j \in [1, \cdots, p]$. A well-known algorithm for this problem is the non-negative least squares (NNLS) algorithm (Lawson and Hanson, 1995). In practice it is known to be rather slow, and several variants exist to accelerate its solution. Accordingly, the problem falls into the domain of quadratic programming with linear inequalities as constraints. It can be transformed as:

$$\min_{a \geq 0} \left( \frac{1}{2} a^\top Qa + c^\top a \right),$$

where $Q = E^\top E$ and $c = E^\top z$. As $Q$ is positive semidefinite and the non-negativity constraints form a convex feasible set, this problem is convex.

A small experiment to show how well the two constraints fit together is depicted in Figure 2.6. A set of 30 endmembers are extracted from a hyperspectral image with VCA. Then non-negative least squares are used to compute the 30 fractional abundances for each pixel. Summing up all the abundances per pixel gives an estimate of how much the ASC is violated when not enforcing it. Despite constraining the abundance values to be non-negative, they rarely sum up to one.

**Soft constraint.** The ASC can also be treated as a soft constraint, by including it as an extra observation in the regression/least squares problem. Eq. (2.3) can thus be
Figure 2.6: Left: A true colour composite of a hyperspectral image. Right: Per-pixel sum of abundances obtained with non-negative least squares. The values differ significantly from 1 without the Abundance Sum Constraint (ASC) (see the colour scale). A total of 30 endmembers were extracted with Vertex Component Analysis (VCA).

transformed as:

\[ z = \begin{bmatrix} z \\ \delta \end{bmatrix}, \quad E = \begin{bmatrix} E \\ \delta 1^\top \end{bmatrix}, \quad a = \begin{bmatrix} a \\ \delta \end{bmatrix}. \quad (2.8) \]

The magnitude of \( \delta \) controls the strength of the ASC. If it is set to a big number it enforces the constraint in a stronger way, while for smaller \( \delta \) the constraint is only softly imposed and variations are observed. The estimation might become numerically unstable with a very high \( \delta \) when \( E^\top E \) is inverted. This formulation moreover, does not conflict with the ANC and can be used in combination.

**Fully Constrained Least Squares.** The combination of both full additivity and non-negativity constraints defines the fully constrained least squares as:

\[
\min_{a} \frac{1}{2} \| z - E a \|_2^2 \\
\text{subject to } a_j \geq 0 \quad \forall \ j \\
\sum_{j} a_j = 1.
\quad (2.9)
\]

Multiple references exist for the fully constrained inversion (Heinz and Chang, 2001, Heylen et al., 2011a). It can also be considered as a quadratic programming problem with linear equality and inequality constraints and it is also convex. An efficient solution using an Alternating Direction Method of Multipliers (ADMM) variant, termed SUnSAL (Bioucas-Dias and Figueiredo, 2010), has been employed to solve this problem within the thesis.

### 2.3.4 Non-negative Matrix Factorisation

A very different approach to the inversion problem is the use of non-negative matrix factorisation (NMF). Here, unmixing is done simultaneously for the whole image and
multiple pixels are considered at the same time. In these methods the endmembers are usually allowed to change from their initial values, so matrix factorisation methods can be considered as variable endmember methods. This offers the possibility to fine-tuning the endmember spectra to the scene. The main concept is to iteratively update endmembers and abundances, by only updating one of the two in each step, while enforcing non-negativity. A review of matrix factorisation methods can be found in Berry et al. (2007). The problem is defined as minimising the function \( f(E, A) = \frac{1}{2} \| Z - EA \|_F \).

\[
\min_{E,A} f(E, A) = \min_{E,A} \frac{1}{2} \| Z - EA \|^2 \\
\text{subject to } E \geq 0, A \geq 0
\] (2.10)

where the inequalities in \( E \) and \( A \) must hold in an element-wise fashion. The solution to this problem is convex in either \( E \) or \( A \), but not in both at the same time. More specifically, the solution is not unique as for an arbitrary non-negative matrix \( D \), whose inverse \( D^{-1} \) is also non-negative, \( EDD^{-1}A \) is also a valid solution, giving an infinite number of combinations. Due to the fact that the problem is not convex, initial non-negative solutions/approximations of the matrices \( E \) and \( A \) have to be defined in advance.

Given the initial solutions, gradient descent can be used to solve eq. (2.10), in which each iteration uses the following update rules:

\[
A \leftarrow A - \epsilon_A \frac{\partial f}{\partial A} \\
E \leftarrow E - \epsilon_E \frac{\partial f}{\partial E}.
\] (2.11)

Depending on the step size \( \epsilon_A \) and \( \epsilon_E \) the result of \( A \) and \( E \), might become negative. This is obviously undesirable and it is solved with a non-negativity projection step. The projection simply sets to zero all negative values after each update, thus at the same time promoting sparsity. Due to its similarity with the problem presented in Chapter 3, this projected gradient descent concept is employed later on within that chapter.

Moreover, a convenient multiplicative update is used to solve the matrix factorisation in eq. (2.10), according to this rule:

\[
A \leftarrow A \cdot (E^\top Z)/(A^\top AE + \epsilon) \\
E \leftarrow E \cdot (ZA^\top)/(EEA^\top + \epsilon)
\] (2.12)

As all the variables used within the updates are non-negative, they do not create any negative values. Some care must be taken to avoid divisions by zero; this is done by adding \( \epsilon \), a small positive value, to the denominator. The main disadvantage of the multiplicative rule update is that, if an element of either \( E \) or \( A \) is set to zero, then it cannot become positive through the updates again. The formulation of non-negative matrix factorisation is utilised by Yokoya et al. (2012) in their work on hyperspectral and multispectral image fusion.
Variants of the pure NMF can also be defined using some constraints. A predominant example is to impose sparsity, limiting the number of non-zero elements in $A$ (Hoyer, 2004), by computing a sparsity measure with a formula that relates the L1 and L2 norms. Similar constraints have been imposed in the remote sensing community by Jia and Qian (2009) and Qian et al. (2011). Furthermore, not only sparseness constraints, but also other linear constraints can easily be added. E.g., here the ASC is considered:

$$\min \frac{1}{2} \| Z - EA \|^2 + \lambda J_1(E) + \mu J_2(A)$$

subject to $E \geq 0, A \geq 0$ (2.13)

where $J_1(E)$ and $J_2(A)$ are penalty terms introduced to enforce certain application-dependent constraints and $\lambda$ and $\mu$ are the respective weights that control the trade-off between the approximation error and enforcing constraints. The constraint could be for instance in the form of $J_1(E) = \| E \|^2_F$ to penalise solutions of large Frobenious norm. Another constraint could be the ASC, i.e., the additivity of the columns in $A$ to sum to one:

$$J_2(A) = \| 1^\top A - 1^\top \|^2_2,$$

(2.14)

where $1$ is consistent with the dimensions of $A$. Then the multiplicative rules can be extended as

$$A_{ij} \leftarrow A_{ij} \frac{(E^\top Z)_{ij}}{(A^\top AE)_{ij} + \lambda(\partial J_1(E)/\partial e_{ij})}$$

$$E_{ij} \leftarrow E_{ij} \frac{(ZA^\top)_{ij}}{(EEA^\top)_{ij} + \mu(\partial J_2(A)/\partial a_{ij})}$$

(2.15)

Finally, since each individual subproblem of NMF is convex, one could use the so-called alternating least squares (ALS) with a unique solution for $E$ and $A$ at each step. In the multiplicative and gradient descent updates, if the solution starts heading downhill towards a fixed point, it must remain on this path and it cannot escape. Theoretically, ALS can prevent the solution from going down a poor path. The changes at each individual iteration though are too abrupt and do not benefit from the good initial solutions supplied. Therefore, in practice ALS does not yield good convergence and it is generally avoided.

### 2.3.5 Sparse Unmixing

Sparse linear unmixing is used to determine an optimal, small subset of spectral signatures of materials contained in each pixel and to estimate their fractional abundances, assuming that the pure spectral signatures are given in advance. With this assumption, the observations are mixtures of a subset of given signatures available in a potentially very large “spectral library”. This library must be able to model each (mixed) pixel in the scene. Sparse unmixing has close links with compressed sensing, basis pursuit, and matching pursuit. Its main advantage is that it alleviates the need to estimate the number and signatures of the endmembers in advance and relies on the existence of
2.3. HYPER SPECTRAL IMAGE PROCESSING

high quality endmembers from a spectral library. However, even though these spectral libraries contain many entries, usually a pixel does not have more than 10 endmembers, which makes the problem very sparse and easier to solve.

In this case, \( E \) is considered to be a spectral library that has more endmembers \( p \), than spectral bands \( B \), i.e., it corresponds to an underdetermined system of equations. Moreover, as the hyperspectral image is corrupted by noise, the equation \( z = Ea \) does not hold precisely. Subsequently, the L2 norm of the difference \( \| z - Ea \|_2 \) is allowed to have a maximum deviation with value \( \eta \), a tolerance against noise and modelling errors. The sparse unmixing problem is concerned with finding the minimum number of endmembers that can explain the spectrum of \( z \) with a maximum deviation of \( \eta \) and is formulated accordingly as follows:

\[
\min_a \| a \|_0 \quad \text{subject to } \| z - Ea \|_2 < \eta, \ a \geq 0 ,
\]

where \( \| \cdot \|_0 \) corresponds to the L0 norm (the number of non-zero entries). L0-minimisation is NP-hard, but greedy algorithms such as orthogonal matching pursuit (OMP) provide a polynomial time approximation to the L0 norm. Surprisingly though, the solution of minimising the L1 norm is very close to that of minimising the L0 norm (Tibshirani, 1996), with the advantage of being convex. The L1 solution is known as basis pursuit and can be solved with linear programming. This fact makes the problem tractable:

\[
\min_a \| a \|_1 \quad \text{subject to } \| z - Ea \|_2 < \eta, \ a \geq 0 .
\]

The solution of which can be found by minimising the respective Lagrangian to the transformed formulation:

\[
\min_a \| Z - Ea \|_2^2 + \lambda \| a \|_1 \quad \text{subject to } \ a \geq 0 ,
\]

that is very similar to the well-known least absolute shrinkage and selection operator (LASSO), but without the extra non-negativity constraint (ANC). Iordache et al. (2011) present multiple variants of basis pursuit and OMP solutions for sparse unmixing of hyperspectral data combining different constraints. Along the same lines, Iordache et al. (2012) use total variation to incorporate spatial information, exploiting the fact that neighbouring pixels are likely to have the same endmembers with slightly different mixtures. Moreover, recently Esmaeili Salehani et al. (2016) start from an L1 norm optimisation and gradually convert it to a L0 norm via an \( \arctan \) smoothing function that approximates the norm between L0 and L1.

So far, the sparse regression problems were only constrained using the ANC. Akhtar et al. (2015b) argue that with appropriate normalisation of the spectral library the ANC implicitly enforces the ASC, implying that there in no need to specifically enforce ASC.

Even though the idea of sparse unmixing is very promising, the heterogeneity in the available spectral libraries constitutes a significant hurdle. The three main disadvantages are: (i) The limited number of samples. It would be very challenging to capture the spectra of all materials. (ii) The conditions at the time of data capture. The observed spectra are rarely acquired under the same atmospheric conditions and the
atmospheric influences can in practice only be corrected approximately. \( \text{iii} \) The spectra are by nature very correlated, due to physical processes, such as water absorption.

An important step towards the general use of spectral libraries for unmixing purposes is the fact that publicly accessible spectral libraries already exist, \textit{e.g.}, U.S. Geological Survey (USGS) spectral library \cite{Kokaly2017}, the ASTER spectral library \cite{Baldrige2009} and the open source library SPECCHIO \cite{Hueni2009}.

# 2.4 Super-Resolution

Super-Resolution refers to techniques that increase the resolution of images. The result is an image with an increased number of pixels. As the term can be used in many situations, this section gives an overview of the most common usages. Moreover, in Figure \ref{fig:super-resolution} there is a schematic overview. The term super-resolution here refers to spatial resolution, except if otherwise stated.

## 2.4.1 Multiframe Super-Resolution

The term \textit{super-resolution image reconstruction} was first used for the process of obtaining one high-resolution image from multiple observed low-resolution images. The main idea is that the low-resolution images represent different samples of the same scene and are shifted by sub-pixel translations as well as downsampled (aliased). From these multiple jittered images a high-resolution image can be inferred, if the sub-pixel shift is not an integer, in which case the images are the same and no new information can be used. With multiple takes from one camera and by estimating their relative co-registration with sub-pixel precision, this extra information makes super-resolution possible. The core of the method lies in the observation model of the imaging process. The model consists of warping, blurring and downsampling that are applied to the high-resolution image to get the low-resolution observed images. First, the parameters of the model must be determined and they can vary per image. For one image the system is underdetermined, but by adding more images the system can be inverted. Note however, that this technique requires a lot of observations and can only be used for moderate upsampling rates, as the required observations grow quadratically with the upsampling rate. Finally it may also be difficult to acquire multiple jittered images in specific cases, \textit{e.g.}, in the case of satellite imaging or if the objects in the scene move. A similar technique is also used for image restoration to recover images that are corrupted by blur (\textit{e.g.}, motion blur) or aliasing (but without changing the image size).

## 2.4.2 Single Image Super-Resolution

Single Image Super-Resolution corresponds to upsampling a single low-resolution RGB image to high \textit{spatial} resolution. This has been a popular topic for several years,
2.4. SUPER-RESOLUTION

and quite a lot of literature exists, going beyond simple interpolation.

Early methods devised clever upsampling functions by manually analysing the image statistics and using adequate regularisation to compensate for the underdetermined systems. More recently, the trend has been to use example-based methods that rely on training samples (Freeman et al., 2002). Research on image statistics suggests that image patches can be represented well as a sparse linear combination of elements from an appropriately chosen, overcomplete library. Such libraries are built with sparse signal representation from multiple low and high resolution image patches, in combination with a sparsity prior. For each image patch, the sparse representation of the low-resolution patch can be used to recover the high-resolution patch with the same coefficients. Theoretically, from compressed sensing the sparse representation can be correctly recovered from downsampled signals. This is in general true under mild downsamplings. Prominent examples of this category are given in (Yang et al., 2010, Zeyde et al., 2010, Kim and Kwon, 2010, Gao et al., 2012, Timofte et al., 2014).

In the last few years, end-to-end trained convolutional neural networks have boosted single-image super-resolution. Significant improvements have been shown by Dong et al. (2014), Kim et al. (2016a,b) that utilise deep networks. Even though the networks require a lot of resources, recent hardware advances enable real-time solutions (Shi et al., 2016).

Interestingly, standard distance measures (e.g., the root mean square error) do not seem to be the best loss function to obtain visually convincing upsampling results. Other loss functions, so-called “perceptual losses” aim for “photo-realism” (implemented with adversarial networks) and are better suited for human perception, although the actual intensity differences are higher (Ledig et al., 2017).
2.4.3 Hyperspectral Single-Image Super-Resolution

Inspired by single-image RGB upsampling, a few research articles have utilised similar ideas to enhance the spatial resolution of hyperspectral images. Akgun et al. (2005) super-resolve hyperspectral images by representing the hyperspectral observations from different wavelengths as weighted linear combinations of a small number of basis image planes. Moreover, Gu et al. (2008) integrate spatial/spectral information for super-resolution. Another natural way to increase the spatial resolution, along the same lines as multiframe super-resolution of RGB images, is to acquire multiple hyperspectral images from slightly different viewpoints and apply standard super-resolution methods (Zhang et al., 2012). Nevertheless, even recent works (Wang et al., 2017, Irmak et al., 2018) do not use example-based principles, but rather exploit the sparse properties of the hyperspectral images with low-rank methods and problem-specific regularisation. The main reason deep learning does not appear to have been used, is the lack of available images that can be acquired, and their complexity in the spectral domain.

2.4.4 Spectral Super-Resolution

There is a large body of literature on single-image super-resolution, which is however limited to enhancing the spatial domain. Single-image spectral super-resolution on the other hand addresses the complementary problem, how to increase the spectral resolution of the input image beyond the coarse RGB bands. The problem is also heavily under-constrained, where for a typical terrestrial application, the goal is to generate, for each pixel, \( \approx 30 \) spectral bands from 3 input bands. The difference is even more extreme in aerial and satellite remote sensing, because typical low spectral resolution sensors only have up to 10 spectral bands, whereas hyperspectral sensors have one order of magnitude more bands, that cover the same range.

As already mentioned, hyperspectral images can be represented in a low-dimensional subspace, meaning that “blind” spectral super-resolution is possible. Given that most scenes consist of a limited number of materials, distributed in characteristic patterns a generic prior can be learned from a suitable training set of similar images. The high-frequency spatial information and spatial context make it possible to separate the spectral signatures of different objects and materials that have similar RGB colours.

The main drawbacks of this process, compared to single-image super-resolution (Subsection 2.4.2), is that training data is limited, because hyperspectral images are not as omnipresent. On the contrary, creating training samples for the single-image super-resolution is achieved by simply downsampling existing RGB images to the desired resolution, practically for free. Another drawback is the fact that after simulating the training examples the system is “tuned” to the specific spectral response of the camera which was simulated.

In concurrent work (not included in the thesis) which was carried out in our laboratory at ETH Zurich, we have developed a spectral super-resolution method (Galliani et al.,...
Figure 2.8: A schematic overview of three different types of multi-sensor super-resolution. Blue is used for panchromatic/multispectral images and green for hyperspectral images.

Therein, we trained a CNN to predict the missing high-frequency detail of the colour spectrum observed at each pixel, in an end-to-end fashion. Further work has also been carried out by (Arad and Ben-Shahar, 2016) using a sparse dictionary of hyperspectral signatures and their corresponding RGB projections. Along the same lines, Aeschbacher et al. (2017) argue that traditional, shallow approaches, based on sparse representation perform on par with recent deep learning methods. Nevertheless, in the recent NTIRE 2018 challenge on spectral reconstruction from RGB images (Arad et al., 2018) deep learning outperformed all other methods.

### 2.4.5 Multi-sensor Super-Resolution

The fusion of a low-resolution hyperspectral image with a higher-resolution multispectral image can be seen as *hyperspectral super-resolution*, as the process enhances the spatial resolution of the hyperspectral image, by using auxiliary data (the multispectral image of higher spatial resolution in this case). Alternatively, one could see it as spectral super-resolution of the high resolution bands, supported by the low resolution ones. The auxiliary data is considered the extra source of information and this is the type of super-resolution, considered in Chapter 3. More details and related work are investigated within that chapter. A schematic overview of multi-sensor super-resolution is shown in Figure 2.8.
2.4.6 Pan-sharpening

A specific type of multi-sensor super-resolution (data fusion) that combines the spatial details of a panchromatic image and the spectral details of a multispectral image is called pan-sharpening. The panchromatic image input must have higher spatial resolution than the multispectral image. This configuration of images is a standard setup on many active satellites, such as IKONOS, GeoEye, SPOT, QuickBird, WorldView and Landsat.

The two main approaches considered for pan-sharpening are component substitution and multiresolution analysis. The former performs some kind of spectral transformation of the multispectral image, to decouple colour and brightness, and then substitutes one of these components with the high resolution image. Some well-known transformations include the intensity-hue-saturation (IHS), principal component analysis (PCA) and Gram-Schmidt (GS) spectral sharpening. The multiresolution approaches decompose the panchromatic image in multiple resolutions and inject the spatial details in the resampled multispectral bands. The spatial details that are injected can be extracted with different computational tools, e.g., the decimated wavelet transform (DWT), “a trous” wavelet transform (ATWT) and Laplacian pyramid (LP). Apart from these two main approaches, a further idea has been to use Bayesian methods for inference in a suitable statistical model of pan-sharpening. This ill-posed problem must be regularised, using for example total variation or compressive sensing.

For a review and critical comparison of pan-sharpening methods, we refer the reader to Vivone et al. (2015).

Hyperspectral Pan-sharpening. The term pan-sharpening refers usually to multispectral pan-sharpening, but recently the term hyperspectral pan-sharpening is used to describe the process of enhancing the resolution of a hyperspectral image, given a panchromatic one. Conventional pan-sharpening methods, however, are not suitable (Loncan et al., 2015). Hence, it is necessary to design hyperspectral pan-sharpening methods to address the particular problem. Among the top performers are Bayesian methods with sparsity prior, inverse methods that use total variation regularisation, and matrix factorisation methods. More details can be found in a review by Loncan et al. (2015).

2.5 Convolutional Neural Networks

Convolutional Neural Networks (CNN) are closely related to deep learning (for an introduction see Goodfellow et al., 2016). While deep learning has been extensively applied to computer vision, it is also becoming popular in the domain of remote sensing (for a review see Zhu et al., 2017). In the following, the basics concepts are briefly presented.

A CNN is a special form of multi-layer perceptron, which is particularly suitable for image and sound processing. CNNs are inspired by biological processes and exploit
the connectivity patterns found between neurons that are observed in the human visual cortex. Each neuron receives an input, performs a linear transformation, followed by a non-linearity. Neural networks are used to approximate complex arbitrary models with a sequence of simple differentiable functions. These functions are referred to as layers. Stacking multiple layers sequentially creates a deep network and machine learning with deep networks is called *deep learning*. The advantage of stacking many layers is that they collectively learn a complex representation of the input data and extract features directly from the image pixels. The rebirth of CNNs from 2012 onwards (Krizhevsky et al., 2012) was facilitated by the availability of (*i*) massive parallel computing on GPUs and (*ii*) a plethora of training data.

### 2.5.1 Forward Pass

**Dense layers.** A *dense* or *fully connected* layer has a connection between each of its neurons to all the neurons in the next layer. This layer is the same as a in a traditional multi-layer perceptron (Goodfellow et al., 2016). The operations performed by such a layer are simple linear operations, that are equivalent to matrix multiplication followed by a vector addition.

**Convolutional layers.** A convolutional layer on the other hand, performs a convolution of a kernel/filter (plus an additive bias) with the input image and thus has a limited *receptive field* compared to dense layer as one pixel only affects a specific neighborhood in the subsequent layer. This also radically limits the number of learnable parameters, with a consequence of allowing the network to be deeper with fewer parameters. Moreover, by sharing the filter weights across the image, convolution layers are tuned to the shift-invariance of the image statistics.

**Activation layers.** Typically, the power of neural networks is unleashed with the activation layers. These are non-linear functions that are applied after convolutional or fully-connected layers. Only the stacking of dense layer would act like as a single linear transformation. The most common activation function is the Rectified Linear Unit (ReLU). It simply sets to zero all the negative values of a layer: $f(x) = \max(x, 0)$. Additionally, other functions can be used to enforce non-linearity, the hyperbolic tangent $f(x) = \tanh(x)$, or the the sigmoid function $f(x) = (1 + e^{-x})^{-1}$. Both the hyperbolic tangent and the sigmoid function however, saturate towards their extreme values and hinder the gradient flow. On the contrary, ReLU does not saturate the positive responses and is very cheap to compute, hence it has become the most common non-linearity choice.

**Pooling layers.** The concept of pooling in CNNs is to enable downsampling of the image. The most common non-linear downsampling is *max pooling*. Max pooling partitions the image into a raster of continuous non-overlapping regions and for each region outputs the maximum value. The process of subsampling reduces the resolution of the
image and simultaneously improves the position invariance of the features, meaning that the exact location is less important than the relative position of the features. Importantly, every pooling layer increases the effective field of view of subsequent convolutions, giving the network the power to learn large and complex spatial relations. Additionally, it reduces the number of parameters of the network and controls overfitting. In addition to max pooling, average pooling (extracting the average value within each region) is also commonly used.

**Loss layers.** A loss layer defines the penalty between the predicted and true labels during training and is the final layer of the network. Depending on the task, various functions can be used: Euclidean loss (for real valued labels), softmax (for mutually exclusive classes) or sigmoid cross-entropy for independent probability values.

### 2.5.2 Backward Pass

The objective of training a CNN is to find a set of parameters of the layers, that minimises the loss function. Training the network involves iterating between a forward and a backward pass through the network. For this process to operate smoothly the network weights must be initialised in a way that will not restrict all of the activations to zero and will not cause the network to explode as a result of huge activations. Weight initialisation is actually quite important, as in any other non-convex optimisation. The problem of properly initialising the weights can be alleviated with the use of batch normalisation (Ioffe and Szegedy, 2015). It forces the activations to take a unit Gaussian distribution at the beginning of training. This is accomplished by using a linear transformation with learnable weights, which is differentiable. It is used commonly after the dense or convolutional layers and before the non-linearities.

The network is trained by back-propagating the errors through the network, using an algorithm called back propagation (Rumelhart et al., 1986; LeCun et al., 1989). It is a gradient based learning method that performs differentiation based on the chain rule, by taking the numerical derivative for each individual weight. Due to the large size of training sets, stochastic gradient descent is deployed, meaning that at each iteration the parameters are updated using only a small part of the training data.

As the number of trainable parameters compared to standard machine learning is massive, CNNs can show signs of overfitting, if not enough data are used. Whenever overfitting occurs, a few measures have been proposed to decrease its effect. First, one can augment the data by applying transformations to the inputs in the spatial and spectral domain. Dropout is also used to randomly deactivate some neurons during training, such that the network becomes more robust against changes in the outputs of individual neurons. Additionally, and similarly to other cases of optimisation, a regulariser is applied to keep the magnitude of the weights low. For example penalising the L2 norm encourages the network to use all the weights, rather than using only the large weights a lot, while the L1 norm favours sparse weights and this leads to resilience against noisy data. A final measure taken against overfitting is to monitor the
2.5. CONVOLUTIONAL NEURAL NETWORKS

Figure 2.9: The network architecture of AlexNet (Krizhevsky et al., 2012). It contains a series of convolutions, non-linearities, poolings and dense (fully connected) layers.

fitting error also on a separate validation set and to stop learning as soon as that error does not decrease any more (early stopping). The point of early stopping represents a trade-off between optimal learning and overfitting.

2.5.3 Applications and Architectures

LeNet-5, a simple CNN able to recognise characters and digits was proposed by LeCun et al. (1998). It achieved impressive results on the MNIST digit recognition benchmark dataset (Deng, 2012), but due to the computational complexity of training, the applications were limited at the time of the development.

Recently, one of the most important applications of CNNs has been in image recognition. The large amount of training data available through the annual “ImageNet Large Scale Recognition Challenge” (Deng et al., 2009) has facilitated the use of CNNs and in 2012 Krizhevsky et al. (2012) achieved a 16% error rate (top-5 error) with AlexNet, better than the error obtained with traditional recognition algorithms which had saturated at that time at 25%. Since then, CNNs have been the only competitive technology in that challenge, and the error has dropped to <5% with their continuous improvement. The CNNs used for image recognition usually contain among others components, convolutional layers, pooling and at the end, after flattening (reshaping) of the images, fully connected layers, see e.g., AlexNet in Figure 2.9. A notable architecture for image recognition is VGGNet (Simonyan and Zisserman, 2015), which has pushed to the extreme the philosophy of using small filter kernels and many layers, reasoning that the depth of the network is a critical component for good performance. Nevertheless in training deep networks problems are encountered which result in vanishing gradients and very slow convergence. Interestingly, ResNet developed by He et al. (2016) introduced special “skip connections” that can bypass a few layers of the networks and add residuals to the main stream (tunneling of the gradients). The skip connections alleviate the vanishing gradient problem and speed up the training significantly. Alternatively, ResNet can also be seen as an ensemble of smaller networks combined into one layout (Veit et al., 2016).

Furthermore, Long et al. (2015) used end-to-end training, from pixels-to-pixels with
application to semantic segmentation. They removed the dense layers at the end of the network (retaining a fully convolutional network) and upsampled the spatial dimensions to match those of the input. With transposed convolutions an “hourglass” type of network contains an encoder-decoder part, that is assisted by skip connections to maintain fine spatial details. Many more architectures have been proposed for semantic segmentation, most notably U-Net (Ronneberger et al., 2015), SegNet (Badrinarayanan et al., 2017), Tiramisu (Jégou et al., 2017) and RefineNet (Lin et al., 2017).

The extensive use of CNNs has also advanced natural language processing and recommendation systems. Other types of deep network architectures exist, but are not usually used for image processing, e.g., recurrent neural networks (RNNs) are designed for sequence modelling.
Chapter 3

Hyperspectral Super-Resolution with Spectral Unmixing Constraints

Charis Lanaras, Emmanuel Baltsavias, Konrad Schindler
Remote Sensing 9 (11), 1196, 2017

(Author version; for typeset version please refer to the original journal paper.)

3.1 Abstract

Hyperspectral sensors capture a portion of the visible and near-infrared spectrum with many narrow spectral bands. This makes it possible to better discriminate objects based on their reflectance spectra and to derive more detailed object properties. For technical reasons, the high spectral resolution comes at the cost of lower spatial resolution. To mitigate that problem, one may combine such images with conventional multispectral images of higher spatial, but lower spectral resolution. The process of fusing the two types of imagery into a product with both high spatial and spectral resolution is called hyperspectral super-resolution. We propose a method that performs hyperspectral super-resolution by jointly unmixing the two input images into pure reflectance spectra of the observed materials, along with the associated mixing coefficients. Joint super-resolution and unmixing is solved by a coupled matrix factorisation, taking into account several useful physical constraints. The formulation also includes adaptive spatial regularisation to exploit local geometric information from the multispectral image. Moreover, we estimate the relative spatial and spectral responses of the two sensors from the data. That information is required for the super-resolution, but often at most approximately known for real-world images. In experiments with five public datasets, we show that the proposed approach delivers up to 15% improved hyperspectral super-resolution.
CHAPTER 3. HYPERSPECTRAL SUPER-RESOLUTION

3.2 Introduction

Hyperspectral imaging, also called imaging spectroscopy, delivers image data with many (up to several hundreds) contiguous spectral bands of narrow bandwidth, on the order of a few nm. The spectral range depends on the sensor; typically it covers the visible wavelengths and some part of the near-infrared region, up to at most 2.5 μm. With hyperspectral sensors (also known as imaging spectrometers), one can more accurately distinguish visually similar surface materials, since every material has its own, characteristic reflectance spectrum. For our purposes, we in fact identify a “material” with a unique reflectance spectrum. Without reference spectra, one cannot separate two physically-different surface compounds if they always occur together in fixed proportions. Based on these spectra, one can not only detect and localise more different objects, e.g., for land cover mapping, environmental and agricultural monitoring; one can also estimate further (bio-)physical and (bio-)chemical properties. For a discussion of important applications, see, e.g., Varshney and Arora (2004).

However, as a consequence of the very narrow spectral bands, only a small amount of radiant energy is available in each band. To achieve an acceptable signal-to-noise ratio (SNR), the sensor area per pixel must therefore be large, leading to coarser geometric resolution. This is the opposite situation of multispectral cameras, where much of the spectral information is sacrificed to achieve high spatial resolution, by integrating the scene radiance over fewer, much wider spectral bands.

To get the best of both worlds, a natural possibility is to fuse a hyperspectral image (termed HSI in the following) with a multispectral image (termed MSI), resulting in a synthetic image with both high spectral and spatial resolution; see Figure 3.1. This process is referred to as hyperspectral super-resolution or hyperspectral and multispectral data fusion. For a recent overview, see Yokoya et al. (2017).

![Figure 3.1](image.png)

Figure 3.1: Hyperspectral super-resolution aims to fuse a low resolution hyperspectral image (HSI) with a high resolution multispectral image (MSI).

A starting point for this work is the observation that hyperspectral super-resolution is linked to the well-known problem of spectral unmixing: pixels in the HSI, having
a larger footprint on the ground, often include a mixture of multiple materials. The aim of unmixing is to determine, at each pixel, the observed pure material spectra (endmembers) and their proportions (abundances). When increasing the resolution, there will be fewer mixed pixels; at the same time, the endmembers and abundances should be preserved.

We follow the classical, linear spectral mixing model (Keshava and Mustard, 2002). Inspired by it, we formulate super-resolution as a factorisation of the image data into a (non-orthogonal) basis of endmembers, and the corresponding coefficients to represent the data in that basis, which can be interpreted as fractional abundances. Unmixing imposes physical constraints: in particular, neither the reflected intensity at any wavelength, nor the surface area covered by a material can ever be negative, and each observed spectrum should be completely described by a mixture of few materials. In our work, we impose the necessary constraints to ensure a physically-plausible unmixing. We point out that this does not guarantee physically correct, pure material spectra: the latter cannot be found from the data alone. Still, we will, in line with the literature, use the terms endmembers and abundances. Technically, we cast the super-resolution as a coupled, constrained matrix factorisation of the two input images into their mutually-dependent endmembers and abundances. That is, we constrain the problem as tightly as possible under the linear mixing model, thus removing slack that could potentially help to absorb non-linear effects. Still, we find that the linear approximation holds sufficiently well, such that the added stability outweighs the reduced flexibility. Our formulation delivers improved fusion results, as well as a visually convincing segmentation into spectrally-distinct materials.

To relate an MSI and HSI of the same scene, one needs the (relative) spatial and spectral responses of the two sensors. Most existing work assumes that these responses are known in advance. However, in practice, this is not always the case. On the one hand, the sensor specifications may be (partially) unknown. On the other hand, the specifications determined in the laboratory may differ from the actual behavior during deployment, particularly in orbit, e.g., due to aging and deterioration of sensor components or malfunctioning of on-board calibration devices (Wang et al., 2010). In actual fact, the overwhelming majority of experiments in hyperspectral super-resolution have so far used synthetic data, since it is otherwise very hard to generate high-quality reference data. When the input HSI and MSI are generated by synthetically degrading an existing image, the spatial and spectral response functions are known, by definition. We aim for a method that can be applied in most practical situations; therefore, we also propose methods to derive the relative spatial and spectral sensor responses directly from the data, given approximately co-registered HSI and MSI. Again, we ensure plausible (non-negative) responses, which significantly stabilises the estimation. We note that the centre of the response function is part of the relative spatial response. By estimating its remaining translation, we can correct co-registration errors, which is rather useful in practice, since sub-pixel accurate co-registration between images from different sources can be challenging.

Here, we also extend an earlier, preliminary version of our method (Lanaras et al., 2015b) to include an adaptive spatial regulariser. It is generally accepted that, in natu-
 CHAPTER 3. HYPER SPECTRAL SUPER-RESOLUTION

r al images, nearby pixel values are correlated. These spatial correlations, particularly in the high-resolution MSI, potentially contain much information that can be used to regularise the result. Hence, we include in our model a quadratic smoothing term, modulated by the edges (gradients) of the MSI, to favour spatially-smooth outputs.

Overall, our proposed method features (i) physically-constrained hyperspectral super-resolution including (ii) spatial regularisation and (iii) data-driven recovery of the necessary relative spatial and spectral response functions. In reference to the underlying mathematical optimisation technique, we call our method SupResPALM (super-resolution with proximal alternating linearised minimisation).

3.3 Related Work

The limitation that hyperspectral images can only be acquired at low spatial resolution has naturally led researchers in remote sensing and computer vision to try and fuse them with high-resolution multispectral or panchromatic images. From the point of view of image processing, the problem is a special case of image fusion (Wang et al., 2005). Perhaps the most widespread use of image fusion in remote sensing is pan-sharpening with a single, pan-chromatic high-resolution image. That method has also been applied to hyperspectral images, for a recent review, see Loncan et al. (2015). However, hyperspectral super-resolution is more general, in that the high-resolution input may have multiple (possibly overlapping) spectral bands.

Our work is closely related to methods that rely on a linear basis (linear unmixing) and some type of matrix factorisation. Kawakami et al. (2011) first learn a fixed spectral basis by unmixing the HSI via $l_1$-minimisation. Then, using this basis, they compute mixing coefficients for the MSI (here RGB) pixels, using $l_1$-minimisation to encourage sparsity. Huang et al. (2014) learn the spectral basis with SVD and solve for the MSI mixing coefficients with orthogonal matching pursuit (OMP). Akhtar et al. (2014) learn a non-negative spectral basis from the HSI, and then solve for the MSI coefficients under a sparsity constraint, again using OMP. Simoes et al. (2015) also recover a linear basis, and include a total variation regulariser to achieve spatial smoothness of the mixing coefficients. Like the previously mentioned works, they proceed sequentially and first construct a basis, which is then held fixed to solve for the coefficients in a second step.

On the contrary, Yokoya et al. (2012) also update the spectral basis. They ensure that the mixing coefficients are $\geq 0$, by unmixing both the HSI and the MSI in an iterative fashion, with non-negative matrix factorisation. Wycoff et al. (2013) also use a joint energy function for both the basis and the coefficients, which supports sparsity and non-negativity. Veganzones et al. (2016) propose an approach to handle the case where the intrinsic dimensionality of the HSI is large, by exploiting that, even then, the local subspace in a small neighbourhood is often low-dimensional. Hardie et al. (2004) and Wei et al. (2015a) allow for priors on the distribution of image intensities and do MAP inference, which, for simple priors, is equivalent to reweighting the contributions of pixels to the error function. Wei et al. (2015b) have proposed an efficient
Bayesian model, in which maximising the likelihood corresponds to solving a Sylvester equation. That work has then been generalised (Wei et al., 2016b), to be more robust with respect to the blur kernel, and also slightly more efficient. Recently, Zou and Xia (2017) use an instance of non-negative factorisation to compute endmembers and abundances with graph Laplacian regularisation, without however making use of the sum-to-one constraint.

In our work, we also rely on the factorisation of a linear mixture model, but attempt to base the super-resolution on an optimal, physically-plausible reconstruction of the endmembers and their abundances. While most state-of-the-art work uses some of the constraints that arise from the elementary physics of spectral mixing, only our model and very recent, concurrent work by Wei et al. (2016a) use all of them. The formulation of Wei et al. (2016a) leads to an alternating optimisation of fusion and unmixing, with Sylvester equation solvers; whereas we use efficient proximal mappings to impose the constraints. Moreover, to account for the influence of the constraints on the spectral basis, we update the endmembers together with the abundances, whereas (Kawakami et al., 2011; Huang et al., 2014; Akhtar et al., 2014; Simoes et al., 2015; Akhtar et al., 2015a) estimate the spectral basis in advance and then keep it fixed. Finally, while several other methods include some sort of smoothness prior, e.g., vector total variation in Simoes et al. (2015), or the $L_2$-distance to the bicubic upsampling in Wei et al. (2015b), we are not aware of any prior work that uses the MSI to obtain an adaptive regulariser.

There is not much literature about the estimation of relative MSI/HSI sensor responses. To the best of our knowledge, the issue has been investigated only as a prerequisite for hyperspectral super-resolution. Yokoya et al. (2013) assume perfect co-registration and model the spatial response as a zero-mean Gaussian blur, whose variance is estimated by maximising the cross-correlation between the HSI and downsampled-MSI gradients. They strongly constrain the spectral response to deviate only slightly from the known laboratory values. Huang et al. (2014) propose an unconstrained solution for the spectral response, which however does not appear to yield plausible results (Yokoya et al., 2013). Finally, Simoes et al. (2015) estimate both responses, assuming that the width of the MSI spectral response is known. Both response functions are regularised by quadratic regularisation of the gradient, which in our experience tends to over-smooth. Furthermore, non-negativity of the response is not enforced.

### 3.4 Problem Formulation

We are searching for an image $\hat{Z} \in \mathbb{R}^{W \times H \times B}$ that has both high spatial and high spectral resolution, with $W$, $H$ and $B$ the image width, image height and number of spectral bands respectively. For that task, we have two inputs: a hyperspectral image $\hat{H} \in \mathbb{R}^{w \times h \times B}$ with (much) lower spatial resolution, i.e. the same region in object space is covered by a smaller number of pixels: $w \ll W$ and $h \ll H$; and a multispectral image $\hat{M} \in \mathbb{R}^{W \times H \times b}$ with high spatial resolution, but a reduced number of spectral bands, $b \ll B$. 
To simplify the notation, we will write images as matrices. That is, all pixels of an image are concatenated, such that every column of the matrix corresponds to the spectral responses at a given pixel, and every row corresponds to a specific spectral band of the complete image. Accordingly, the images are written 
\[ Z \in \mathbb{R}^{B \times N_m}, \quad H \in \mathbb{R}^{B \times N_h} \text{ and } M \in \mathbb{R}^{b \times N_m}, \]
where \( N_h = wh \) and \( N_m = WH \).

In the linear mixing model (Keshava and Mustard, 2002, Bioucas-Dias et al., 2012), the intensities \( z_i \in \mathbb{R}^B \) at a given pixel \( i \) of \( Z \) are described by an additive mixture:
\[
z_i = \sum_{j=1}^{p} e_j a_{ij}, \quad Z = E A,
\] (3.1)

with a matrix \( E \equiv [e_1, e_2, \ldots, e_p] \) of pure spectra (endmembers) a matrix \( A \equiv [a_1, a_2, \ldots, a_{N_m}] \) of per-pixel mixing coefficients (abundances) and \( a_i \equiv [a_{i1}; a_{i2}; \ldots; a_{ip}] \) the abundances at pixel \( i \). By this definition, at most \( p \) endmembers (materials) are visible in the image. The spectra \( E \) act as a non-orthogonal basis to represent \( Z \) in a lower-dimensional space \( \mathbb{R}^p \), and \( \text{rank}\{Z\} \leq p \). Components of the observed spectral vectors that do not lie in the low-dimensional space are considered noise. It is thus important that the basis \( E \) is determined as well as possible, which motivates us to optimally estimate it during processing, rather than stick with initial approximation.

The actually recorded hyperspectral image \( H \) is a spatially-downsampled version of \( Z \),
\[
H \approx ZBS = EABS = E\tilde{A},
\] (3.2)
where \( B \in \mathbb{R}^{N_m \times N_m} \) is a circulant matrix that blurs \( A \) according to the hyperspectral sensor’s spatial response and \( S \in \mathbb{R}^{L_m \times L_h} \) is the downsampling (subsampling) operator that depends on the resolution difference of the two images. The \( \tilde{A} \equiv ABS \) are the abundances at the lower resolution; under a linear downsampling, simply the weighted average of the high-resolution abundances within one low-resolution pixel. We assume that the blur is the same for all the hyperspectral bands, and thus its effect is equivalent to blurring \( A \) directly.

Similarly, the multispectral image \( M \) is a spectrally-downsampled version of \( Z \),
\[
M \approx RZ = REA = \tilde{E}A,
\] (3.3)
where \( R \in \mathbb{R}^{b \times B} \) is the spectral response function of the sensor and \( \tilde{E} \equiv RE \) are the spectrally degraded endmembers (the multispectral signatures of different materials).

The spatial response function \( B \) of the hyperspectral camera and the spectral response function \( R \) of the multispectral sensor are either known from camera specifications, or they can be estimated directly from the data; see below. By combining Equations (3.2) and (3.3) the unknown \( R \) and \( B \) are related, up to noise, via:
\[
MBS = RH,
\] (3.4)
where both sides of the equation correspond to the same image with low spectral and spatial resolution, obtained by degrading the MSI and HSI, respectively. A graphical
3.5. PROPOSED SOLUTION

overview of the observation model for the relative responses is shown in Figure 3.2.

Figure 3.2: A visual representation of the underlying model used to define the relative spatial and spectral response.

3.4.1 Constraints

The core idea of the present paper is to improve super-resolution by making full use of the linear mixing model. In that model, the endmembers \( E \) in (3.1) are interpreted as reflectance spectra of individual materials, and the abundances \( A \) as the relative proportions of a pixel covered by those materials. As a consequence, the following physical constraints must hold:

\[
\begin{align*}
    a_{ij} &\geq 0 \quad \forall \ i, j \quad \text{(non-negative abundance)} \\
    1^\top A &= 1^\top \quad \text{(abundances sum to one)} \quad (3.5) \\
    0 \leq e_{ij} &\leq 1 \quad \forall \ i, j \quad \text{(non-negative, bounded reflectance)}
\end{align*}
\]

with \( e_{ij} \) and \( a_{ij} \) the elements of \( E \), respectively \( A \). \( 1 \) denotes a vector of one’s compatible with the dimensions of \( A \) and \((\cdot)^\top\) denotes the matrix transpose. The first two constraints together bound the \( l_1 \)-norm of the solution, and hence restrict the solution to a simplex. This means that the constraints already include the desired sparsity of the abundances (few materials, respectively endmembers per pixel). The elements of \( E \) have an upper bound of one, assuming that the image intensities have been rescaled to \([0 \ldots 1]\). That is, they behave like surface reflectances, assuming that there is at least one pure pixel in the image whose material is highly reflective in at least one spectral band.

3.5 Proposed Solution

3.5.1 Super Resolution

To solve for the super-resolved image \( Z \), we recover its two factors \( E \) and \( A \). From Equations (3.2), (3.3) and (3.5), we get the following constrained least-squares prob-
CHAPTER 3. HYPERSPECTRAL SUPER-RESOLUTION

\[
\min_{E, A} \frac{1}{2} \| H - E A S \|_F^2 + \frac{1}{2} \| M - R E A \|_F^2 \\
+ \frac{\lambda}{2} \| A D_h W \|_F^2 + \frac{\lambda}{2} \| A D_v W \|_F^2
\]  

(3.6a)

subject to

\[
0 \leq e_{ij} \leq 1, \quad \forall i, j
\]

(3.6b)

\[
a_{ij} \geq 0, \quad \forall i, j
\]

(3.6c)

\[
1^T A = 1^T
\]

(3.6d)

\[
\| A \|_0 \leq s
\]

(3.6e)

with \( \| \cdot \|_F \) denoting the Frobenius norm and \( \| A \|_0 \) the number of non-zero elements of \( A \). The two last (quadratic) terms of (3.6a) are related to the spatial information used to impose information from the MSI to the solution. \( D_h \in \mathbb{R}^{N_m \times N_m} \) and \( D_v \in \mathbb{R}^{N_m \times N_m} \) are two sparse matrices that compute the discrete gradients (differences) between neighbouring pixels in horizontal, respectively vertical direction. The parameter \( \lambda \) controls the strength of the regularisation, and \( W \) is a diagonal matrix of weights that reduce the smoothness prior at high-contrast edges of the MSI, described below. The constraints in Equations (3.6c) and (3.6d) together restrict the abundances \( A \) to the surface of a simplex spanned by the endmembers in \( E \), and thus also act as a sparsity prior on the per-pixel abundances. The last constraint (3.6e) is optional; it serves to further increase sparsity, if desired. The diagonal values of \( W = \text{diag}(w) \) are computed from the vector:

\[
w = \exp\left( -\frac{\left( \frac{1}{Q_{95}} \sum_{i=1}^{b} g_i \right)^2}{2\sigma^2} \right)
\]

(3.7)

where \( \sum g_i \) is the sum of the (vectorised) Sobel gradient magnitudes over all \( b \) channels and \( Q_{95} \) is the 95%-quantile of that vector. The latter normalises the summed gradient magnitudes to make them invariant against the intensity range, such that the weight decay \( \sigma \), which is a user-defined hyper-parameter, need not be adapted to those sensor- and scene-specific influences.

Empirically, solving Equation (3.6a) directly for \( E \) and \( A \) is difficult and rather unstable. The second term is strongly ill-posed w.r.t. \( E \) due to the spectral degradation \( R \), in other words only \( b \) spectral channels do not contain sufficient information to separate \( p > b \) materials. Conversely, the first term is ill-posed w.r.t. \( A \), because the hyperspectral image, after the blurring \( B \) and downsampling \( S \), contains little information how to disentangle the abundance vector of a low-resolution pixel into contributions from its \( (N_m/N_h) \) constituent high-resolution pixels. We found it advantageous to split Equation (3.6a) into a low-resolution (\( H \)) and a high-resolution (\( M \)) part and solve them by alternation. Note however, alternating between the two parts fixes their relative importance: rescaling one of the two data terms with a scalar will not affect the solution.

The low-resolution step minimises the first term of Equation (3.6a) subject to the con-
3.5. PROPOSED SOLUTION

strains on $E$,\[3.5\]
\[
\min_{E} \frac{1}{2} \|H - E\tilde{A}\|_F^2
\]
subject to $0 \leq e_{ij} \leq 1, \ \forall \ i,j.$ \ (3.8)

*i.e.* the endmembers of $H$ are updated for given low-resolution abundances $\tilde{A}$. The latter are straightforward to obtain from (preliminary estimates of) the high-resolution abundances $A$ by spatial down sampling, *c.f.* (3.2). The high-resolution step proceeds the opposite way and minimises the second term of (3.6a) under the constraints on $A$,

\[
\min_{A} \frac{1}{2} \|M - \tilde{E}A\|_F^2 + \frac{\lambda}{2} \|A D_h W\|_F^2 + \frac{\lambda}{2} \|A D_v W\|_F^2
\]
subject to $a_{ij} > 0, \ \forall \ i,j$

$1^\top A = 1^\top$

$\|A\|_0 \leq s$ \ (optional) \ (3.9)

This time the abundances at full resolution are updated for given endmembers $\tilde{E}$, which are again just spectrally downsampled version of the (preliminary) endmembers $E$ from the low-resolution step.

Optimisation Scheme

Both parts of the alternation are constrained least-squares problems. Inspired by the PALM (proximal alternating linearised minimisation) algorithm (Bolte et al., 2014), we propose to use a projected gradient method. For Equation (3.8) the following two steps are iterated for $q = 1, 2, ...$ until convergence:

\[
U^q = E^{q-1} - \frac{1}{c} (E^{q-1} \tilde{A} - H) \tilde{A}^\top
\]

\[
E^q = \text{prox}_E(U^q)
\]

with $c = \gamma_1 \|\tilde{A} \tilde{A}^\top\|_F$ a non-zero constant, and $\text{prox}_E$ a proximal operator that projects onto the constraints in Equation (3.8). What makes the algorithm attractive is that $\text{prox}_E$ is computationally very cheap: it amounts to truncating the entries of $U$ to zero from below and to one from above.

Likewise, Equation (3.9) is minimised by iterating the following steps until convergence:

\[
V^q = A^{q-1} - \frac{1}{d_1} \tilde{E}^\top (\tilde{E} A^{q-1} - M)
\]

\[
- \frac{\lambda}{d_2} A D_h W W^\top D_h^\top - \frac{\lambda}{d_3} A D_v W W^\top D_v^\top
\]

\[
A^q = \text{prox}_A(V^q)
\]

(3.11a)

(3.11b)
with \( d_1 = \gamma_2 \| \tilde{E} \tilde{E}^\top \|_F, \) \( d_2 = \gamma_2 \| WD_v D_v^\top W^\top \|_F \) and \( d_3 = \gamma_2 \| WD_h D_h^\top W^\top \|_F \) non-zero constants and \( \text{prox}_A \) a proximal operator that projects onto the constraints in (3.9). Again, the proximal operator for the simplex projection is computationally efficient, see [Condat (2016)](#). The complete optimisation scheme is given in Algorithm 1.

**Algorithm 1** Solution of minimisation problem Equation (3.6a).

**Require:** \( H, M, S, R \)

Initialise \( E^{(0)} \) with SISAL and \( \tilde{A}^{(0)} \) with SUnSAL

Initialise \( A^{(0)} \) by upsampling \( \tilde{A}^{(0)} \)

\( k \leftarrow 0 \)

while not converged do

\( k \leftarrow k + 1 \)

// low-resolution step:

\( \tilde{A} \leftarrow A^{(k-1)} \times BS \); Estimate \( E^{(k)} \) with (3.10a) and (3.10b)

// high-resolution step:

\( \tilde{E} \leftarrow R E^{(k)} \); Estimate \( A^{(k)} \) with (3.11a) and (3.11b)

end while

return \( Z = E^{(k)} A^{(k)} \)

Since Equation (3.6a) is highly non-convex, we need good initial values to start the local optimisation. We choose the SISAL algorithm [Bioucas-Dias (2009)](#) to initialise the endmembers in \( E \). SISAL robustly fits a minimum-volume simplex to the response vectors (the columns) of \( H \) with a sequence of augmented Lagrangian optimisations, and returns the vertices of the simplex as endmembers. With the initial endmembers \( E^{(0)} \) we then use SUnSAL [Bioucas-Dias and Nascimento (2008)](#) to get initial abundances. SUnSAL includes the constraints in Equations (3.6c) and (3.6d) and solves a constrained least-squares problem for \( \tilde{A}^{(0)} \), via alternating direction method of multipliers (ADMM). Finally, we initialise \( A^{(0)} \) by upsampling \( \tilde{A}^{(0)} \).

We always set \( \gamma_1 = \gamma_2 = 1.01 \). The choice of these parameters, which control the step size of the iterative minimisation, affects only the computation time, not the final result. In our experience, the optimisation exhibits monotonic convergence.

### 3.5.2 Relative Spatial Response

Using only image observations, one can generally not recover absolute sensor characteristics, but only the relative response of one sensor with respect to the other. The spectral and spatial response functions are coupled, since one must be known in order to directly solve for the other, as can be seen from Equation (3.4). In practice, we found that estimating them in two consecutive steps is sufficient. Iterating the two steps is possible, but brings no significant improvement.
3.5. **PROPOSED SOLUTION**

Estimating the spatial response function amounts to finding the discrete 2D blur kernel that maps MSI pixels onto HSI pixels. We assume a separable kernel and split it into horizontal and vertical 1D components. The 1D kernels need not be Gaussian, but are assumed to be unimodal and symmetric. In other words, the kernel has a single peak, which need not be at the centre, and descends symmetrically as one moves away from the peak. The two kernels can be different, to allow for anisotropic blur (particularly in remote sensing, there can be along-track blur and the across-track blur). We do not impose local smoothness of the kernel, since this tends to spread out the tails of the kernel function. The assumption of a separable and symmetric kernel may first seem a restriction, but it holds rather well for real imaging systems, and reduces the number of unknown coefficients, such that in practice, it stabilises the estimation. We compute the unknowns of the blur kernel jointly for all MSI bands, which corresponds to the assumption that the blur of the MSI and HSI are band-independent.

The spatial resolution difference (ratio) between the HSI and MSI is $S = \sqrt{N_m/N_h}$. For simplicity, we restrict the discussion to integer ratios $S \in \mathbb{N}_+$, such that the discrete kernel has the same values everywhere in the image. Let $R_0 \in \mathbb{R}^{b \times B}$ be an initial approximation of the relative spectral response, which reduces the number of spectral channels from $B$ to $b$. Using the approximate spectral response does not change the sharpness of the image (Yokoya et al., 2013), thus we can start by first estimating the spatial response. Let $h_0 = R_0 H \in \mathbb{R}^{b \times N_h}$ denote the image created from $H$, having the spectral channels of $M$. We seek to estimate the blur that will optimally fit $M$ to $h$ under some spatial subsampling. The search window size for the blur coefficients (in units of MSI pixels) is $W = (2k + 1)S$, where $k \in \mathbb{N}$ is determined empirically. For pixels $j = 1, \ldots, Z$ of $h$ we extract $Z$ 1D patches of size $W$ from $M$, either horizontally or vertically. $Z$ is the number of all HSI pixels for which the 1D kernel of size $W/S$ falls completely inside the HSI; meaning that a small border of width $k$ pixels in the HSI is ignored, to avoid boundary effects. We solve the following optimisation separately for each of the two 1D kernels:

$$
\begin{align*}
\min_b & \quad \|h - Lb\|_F \\
\text{subject to} & \quad Gb \geq 0,
\end{align*}
$$

where $b_i \in \mathbb{R}^W$ are the unknown 1D coefficients of the blur, $h \in \mathbb{R}^Z$ are the values of $h_0$ corresponding to the centre of the patches and $L \in \mathbb{R}^{Z \times W}$ is a matrix containing the values of $M$, as $Z$ rows of $W$-dimensional patches. $G \in \mathbb{R}^{W \times W}$ is a matrix whose rows compute the differences between all pairs of adjacent values in the blur kernel $b_i$, ordered by increasing distance from the mode (value closer to the centre of gravity of the kernel minus value of the more distant neighbour). It thus forces the coefficients to decrease with increasing distance from the mode. The row for the most distant kernel coefficient has only the respective element set to one, to ensure non-negativity of $b_i$. The ordering of rows in $G$ is arbitrary, by convention we set it such that the one-values lie on the diagonal. To find the centre of gravity (in horizontal and vertical direction), we first get a solution of Equation (3.12) with $G = I_W$, where $I$ is the identity matrix to enforce non-negativity. The final 2D blur for band $i$ is given by the outer product of the vertical and horizontal kernels, $B = b_{ver} b_{hor}^T$. 


Equation (3.12) is solved as a quadratic programme, using the interior point method (Boyd and Vandenberghe, 2004). We do not constrain the kernel coefficients to sum to 1 during the estimation (although in practice their sum always stays close to one). Instead, we normalise the final blur kernel. Note that the offset between the kernel mode and the centre of the kernel window corresponds to a global (shift) misregistration between the MSI and HSI. It is estimated as part of the relative spatial response. For more technical details, the interested reader is referred to our dedicated paper (Lanaras et al., 2016).

3.5.3 Relative Spectral Response

Here the aim is to estimate the shape and size of 1D kernels that integrate HSI bands into MSI bands. Approximate knowledge of the HSI and MSI spectral responses (e.g., from the sensor manufacturer) can help initialise and guide the estimation. To increase robustness, in this case we use the $l_1$-norm for the data fitting. Given that the HSI bands are very narrow, and assuming they cover the complete spectral range of the MSI, each MSI band $i$ can be expressed as a linear combination of HSI bands.

The estimation of the spectral response $R$ is thus independent for each MSI band (each row of $R$), which leads to the following optimisation for the unknown spectral response $r_i \in \mathbb{R}^B$ of band $i$:

$$
\min_r \| F(m_i - H^T r_i) \|_1 + \mu_i \| D r_i \|_\alpha
\text{subject to } r \succeq 0,
$$

where $m_i \in \mathbb{R}^{N_h}$ is the $i$-th MSI band, spatially downsampled with the blur $B$ (computed above) and $\mu_i \geq 0$ is a regularisation parameter to enforce smoothness of the spectral response curve for band $i$, using the finite difference operator $D \in \mathbb{R}^{B-1 \times B}$ to compute the differences between elements $r_i$ for spectrally adjacent bands. The diagonal matrix $F = \text{diag}(m_{i1}^2, \ldots, m_{iN_h}^2) \in \mathbb{R}^{N_h \times N_h}$ holds individual weights for the individual pixels $m_i$. Weights are selected such that pixels with higher intensity, and thus better SNR, in band $i$ contribute more to the estimation of $r_i$. Empirically, this weighting stabilises the solution. The type of norm $\alpha$ is selected to reflect prior knowledge about $r_i$. Steep, nearly rectangular response curves, like for instance those of the ADS80 or Landsat-8 OLI (Knight and Kvaran, 2014), require piecewise constant kernels and thus $\alpha = 1$. While for flatter, gradually changing response curves, like those of some amateur cameras, $\alpha = 2$ is preferable. (Other choices of $\alpha$ are in principle also possible, but complicate the optimisation.) We do not enforce $1^T r_i = 1$, meaning that the spectral response curves need not integrate to 1. In this way, one can compensate global radiometric differences between the MSI band and the matching HSI bands.

To solve Equation (3.13), we again use the interior point method, starting from the approximate spectral response already used in Section 3.5.2. In practice, one almost always knows the approximate spectral ranges of both MSI and HSI channels; hence, one can limit the search to a smaller subset of bands $\hat{B} < B$. The appropriate subsets $\hat{B}$ depend on the sensors used and must be specified as part of the input data. Note
that restricting the search to reasonable wavelengths only places lower and upper bounds on the response curve, its exact width need not be known. The parameters $\mu_i$ are set individually for each band, as discussed in Section 3.7.1.

### 3.6 Experiments

Before getting to the results, we introduce the five (aerial, terrestrial, and satellite) datasets used in our evaluation, explain error metrics and baselines, and summarise implementation details for SupResPALM.

#### 3.6.1 Datasets

**APEX and Pavia University**

We test the super-resolution approach with synthetic HSI and MSI images derived from two well-known open remote sensing datasets, both captured with airborne sensors. The first test images is Pavia University. One spectral channel is displayed in Figure 3.11. The image was acquired with DLR’s ROSIS sensor and has $608 \times 336$ pixels with a GSD of $\approx 1$ m. There are 103 spectral bands spanning 0.4–0.9 $\mu$m. The second dataset was acquired by APEX (Schaepman et al., 2015), a sensor developed by a Swiss-Belgian consortium on behalf of ESA. APEX covers the spectral range 0.4–2.5 $\mu$m. The image is available as Open Science Dataset, has a GSD of about 2 m and was acquired in 2011 over Baden, Switzerland. A true colour composite is shown in Figure 3.7. We crop an image of $400 \times 400$ pixels, with 211 spectral bands, after removing water absorption bands.

We follow the experimental procedure used in most of the previous work: the original hyperspectral image serves as ground truth, and the input HSI and MSI are simulated by synthetically degrading it. In all cases, we chose $B$ to be Gaussian blur, with variance depending on the resolution difference $S$. For APEX and Pavia University, we select $S = 8$ and a variance of 4 MSI pixels. For each image, we first apply the blur $B$ and then subsample with the given rate $S$ to obtain the HSI.

The MSI were created similarly, by integrating over the original spectral channels with a given spectral response $R$. In the case of Pavia University, we use the spectral response of IKONOS, as also done in Simoes et al. (2015) and Lanaras et al. (2015a), since the two sensors have a fairly good spectral overlap. On the contrary APEX cover a bigger range than existing airborne mapping cameras. We use the spectral response of ADS80, which leads to a partial spectral overlap, since ADS80 does not capture wavelengths beyond 900 nm. The area under all the spectral response curves is normalised to 1 to ensure all MSI bands have the same intensity range.
CAVE and Harvard

To further evaluate our method, we use two publicly available hyperspectral close-range databases. The first database, called CAVE ([Yasuma et al., 2008]), includes 32 indoor images showing, e.g., paintings, toys, food, etc., captured under controlled illumination. The dimensions of the images are $512 \times 512$ pixels, with 31 spectral bands, each 10 nm wide, covering the visible spectrum from 400–700 nm. The second database, called Harvard ([Chakrabarti and Zickler, 2011]), has 50 indoor and outdoor images recorded under daylight illumination, and 27 images under artificial or mixed illumination. The spatial resolution of these images is $1392 \times 1040$ pixels, with 31 spectral bands of width 10 nm, ranging from 420–720 nm. We use only the top left $1024 \times 1024$ pixels to avoid fractional coverage of the HSI pixels, in accordance with Akhtar et al. (2014). For CAVE and Harvard we use the extreme value of $S = 32$, the standard in the existing literature. The corresponding variance of the Gaussian blur is set to $16$ MSI pixels. We use the spectral response $R$ of a typical digital camera, the Nikon D700 (www.maxmax.com/spectral_response.htm).

For all the above cases, we use the known spectral ($R$) and spatial ($B$) responses in the experiments and also assume perfect co-registration, in order to obtain a fair and meaningful comparison to the super-resolution baselines, who also assume perfectly known response functions. In the simulations, we add Gaussian noise of $\text{SNR} = 30$ dB to the HSI and $\text{SNR} = 40$ dB to the MSI, to simulate independent sensor noise.

Real EO-1 Data

Finally, we add a test case where the HSI and MSI inputs are not synthesised from the same source, but are real images as captured by a contemporary earth observation platform. The HSI and the MSI were acquired on 18 June 2016 by the Hyperion, respectively ALI sensors on board USGS’s EO-1 satellite. They show the Rhine river on the border of southern Germany and France. We crop a region of $198 \times 500$ pixels, with all 198 calibrated bands of Hyperion (spectral range 0.4–2.5 mm, original GSD 30 m), and the 9 multispectral bands of ALI (GSD 30 m GSD). Since, unfortunately, ALI has only limited resolution that does not exceed Hyperion, we were obliged to synthetically downsample the HSI to a GSD of 120 m also for these experiments. On the other hand, this has the advantage that we can treat the original 30 m HSI as ground truth for evaluation.

The HSI is simulated by blurring with a Gaussian of variance of 2 pixels and subsampling with stride 4, to obtain an input HSI with GSD 120 m, four times larger than the MSI. A true colour composite of the scene is shown in Figure 3.9.

3.6.2 Error Metrics and Baselines

As primary error metric, we measure the root mean square error between the estimated high-resolution hyperspectral image $\hat{Z}$ and the ground truth $Z$, scaled to an
3.6. EXPERIMENTS

8-bit intensity range $[0 \ldots 255]$.

$$\text{RMSE} = \sqrt{\frac{1}{BN_m} \sum_{i} \sum_{j} (\hat{z}_{ij} - z_{ij})^2} = \sqrt{\frac{\|\hat{Z} - Z\|_F^2}{BN_m}} \quad (3.14)$$

As a second error measure, we compute the Erreur Relative Globale Adimensionnelle de Synthése (Wald, 2000), which is independent of the intensity units and also takes into account the GSD difference between the HSI and MSI.

$$\text{ERGAS} = \frac{100}{S} \sqrt{\frac{1}{B} \sum_{i} \frac{\text{MSE}(\hat{z}_i, z_i)}{\mu_{\hat{z}_i}^2}}, \quad (3.15)$$

where $S$ is the ratio of GSD difference of the MSI and HSI, MSE$(\hat{z}_i, z_i)$ is the mean squared error of every estimated spectral band $\hat{z}_i$ and $\mu_{\hat{z}_i}$ is the mean value of each spectral band.

Additionally, we also compute the the spectral angle mapper (SAM, Yuhas et al., 1992), which is defined as the angle in $\mathbb{R}^B$ between the estimated spectrum $\hat{z}_j$ and the ground truth spectrum $z_j$, averaged over all pixels.

$$\text{SAM} = \frac{1}{N_m} \sum \arccos \frac{\hat{z}_j^T z_j}{\|\hat{z}_j\|_2 \|z_j\|_2} \quad (3.16)$$

where $\|\cdot\|_2$ is the $l_2$ vector norm. The SAM is given in degrees, and compares only the shape of the predicted spectra, while ignoring differences in magnitude (scale of the response vector). As a final quality metric we use Q$2^s$, an extension of the Universal Image Quality Index (UQI) from single-band to multispectral/hyperspectral images, based on hypercomplex numbers (Garzelli and Nencini, 2009). The Q$2^s$ takes on values between $-1$ (worst) and $1$ (best).

As baselines to compare against, we use four state-of-the-art methods, which we term: CNMF (Yokoya et al., 2012), SNNMF (Wycoff et al., 2013), HySure (Simoes et al., 2015) and R-FUSE (Wei et al., 2016b). These baselines were chosen because for them the best results are reported in the literature, and the source code for all of them was made available by the authors. We thus run the authors’ original implementations and tune them for best performance on the datasets used in our study. Furthermore, we report the error metrics for the naively magnified HSI (using bicubic interpolation), as a baseline for upsampling without the additional information from the MSI.

3.6.3 Implementation Details

The inner loops of the two optimisation steps $(3.10a)$ and $(3.11a)$ are run until the update falls below 1%, which typically takes $\approx 10$ iterations in the early stages and drops to $\approx 2$ iterations as the alternation proceeds. The outer loop over the two steps is iterated until the overall cost $(3.6a)$ changes less than 0.01%, or for at most 2000 iterations. In simulated data the limit of 2000 is reached in several cases, while with real data the algorithm converges in a few tens of iterations. As a default setting for our method we
use smoothing with $\lambda = 0.1$, except if stated otherwise. The parameter that governs the contrast-sensitivity of the smoothing is set to $\sigma = 1.5$. Perhaps the most important user parameter is the number $p$ of endmembers. This parameter depends on the scene contents and according to our experience $p = 30$ is sufficient for most Remote Sensing images. In the case of CAVE and Harvard datasets, with much fewer bands and fewer materials per image we reduce $p$ to 10. We use the same number of basis vectors in all baselines. The final results may vary slightly, due to random sampling in the VCA initialisation (for SupResPALM and also for HySure and CNMF). Our current implementation in MATLAB 9.0 has not been optimised for speed. Computation times depend on the image size and the number of iterations, as well as the sparsity parameter $s$, if used. For the EO-1 data with dimensions $500 \times 180$ pixels and 198 channels it takes $\approx 1$ min, on a single Intel Xeon E5 3.2 GHz CPU.

3.7 Experimental Results and Discussion

We now present and discuss the empirical performance of the proposed super-resolution scheme. To keep the length of the paper reasonable, not all available results are presented, since many experiments exhibit similar trends and support the same conclusions. Complementary tables and figures are available in the supplementary material.

3.7.1 Relative Responses

CAVE Database

As baseline for this evaluation we only use HySure, because it is the only one among our baselines that can recover the relative responses from the data. To evaluate the estimation of the spectral response test on a simulated MSI from the CAVE database and compare to the actually used response function (cf. Section 3.6.1). The results are shown in Figure 3.3. The estimated response curves are in both cases fairly close to the true ones (dashed line). However, HySure returns a number of negative response values, which SupResPALM avoids by construction. The latter was run with the setting for smooth responses, $\alpha = 2$. Regarding the spatial response function, please refer to the following section.
3.7. EXPERIMENTAL RESULTS AND DISCUSSION

Figure 3.3: The estimated spectral response for SupResPALM (super-resolution with proximal alternating linearised minimisation) and HySure for the response used in the CAVE and Harvard database simulations.

Real EO-1 Data

For the EO-1 data we estimate the relative responses from the data as described in Sections 3.5.2 and 3.5.3. Since the low resolution HSI was simulated with a known blur kernel, we can evaluate the results of both SupResPALM and HySure against the ground truth, see Figure 3.4a. The estimated spatial responses can be seen in Figure 3.4b for SupResPALM and Figure 3.4c for HySure. The full blur of SupResPALM is given as the outer product of its respective horizontal and vertical components. For the size we used a radius parameter of $k = 2$, larger values give almost identical results. The spatial responses are computed over all nine spectral channels of ALI, assuming uniform blur in all channels. While HySure strongly regularises the response functions, this tends to enlarge their spatial extents and even lead to negative values in the presence of noise. On the other hand, our method is less flexible regarding the shape of the blur, but still ensures a monotonously descending kernel, without negative values. From the estimation of the spatial response, we get a translation of $[0.07, 0.11]$ MSI pixels between the two images, meaning that the two sources are coregistered very well. Still, we do account for the small shift during fusion, by using the estimated, slightly decentred blur kernel.

The spectral response is then estimated at low spatial resolution, i.e., from our input HSI and a degraded MSI with 120 m GSD, downsampled with the estimated blur (Figure 3.4). For comparison, we also ran the estimation of the spectral response at high resolution, using “ground truth” HSI and the input MSI with 30 m GSD. This control experiment excludes any possible influence of the (estimated) spatial blur, but gives practically identical results.
Figure 3.4: The spatial response of Hyperion and ALI for all nine spectral bands. (a) Ground truth. (b) Estimation by SupResPALM. The 1D kernels are shown in green/yellow with the respective coefficients, the 2D kernel in blue/yellow. (c) Estimation by HySure.

The results of both methods are shown in Figure 3.5a. Unfortunately, there is no ground truth for the spectral responses. We do have approximate values computed from the specified absolute responses of the two sensors, but these do not appear to accurately represent the actual band-pass filters used. Due to the strong $l_2$ regularisation, the response curves from HySure end to have non-zero values everywhere inside the maximum spectral range $B$ for any given channel, in some cases including implausible negative values near the spectral bounds. Our method, tuned for steep band-pass filters with $\alpha = 1$, leads to more plausible results and seems to better pick up which hyperspectral bands really contribute to each MSI band, especially for longer wavelengths, see Figure 3.5a. Note, when run with a one-normalised spatial blur kernel, SupResPALM and HySure include the band-specific radiometric scaling in the spectral response function. The approximate values from the specifications (blue line) were re-normalised per band to have the same area under the curve as SupResPALM (respectively, HySure, the estimated areas are practically identical).

In Figure 3.5b, we plot the influence of the regularisation term $\|Dr_i\|$ against its weight $\mu_i$, for five MSI (ALI) channels. The curves show a typical behavior, which we use to determine the weight: for too low values the band-to-band differences remain constant (no smoothing), for too high values they vanish (strong oversmoothing). In between there typically is a region where the curves flatten out to return similar results across a range of weights (denoted with red circles). We found that choosing weights from those intermediate plateaus, where the solution is rather stable, yields good spectral responses. In general the spectral responses remain reasonable over a wide range of values, and their exact shape has only a small effect on the final super-resolution. Even simple heuristics like using half of the maximum (unsmoothed) $\|Dr_i\|$ will give very reasonable results, see Section 3.7.2.
3.7. EXPERIMENTAL RESULTS AND DISCUSSION

Figure 3.5: (a) The relative spectral responses estimated from SupResPALM (green) and HySure (red). The approximate values given from the sensor specification are shown with a blue line. The area under the curve does not sum to one, because of global radiometric differences. (b) The regularisation term $\|D_r\|$ against its weight $\mu_i$, for five MSI (ALI) channels. Flat regions in the curves lead to stable results that empirically yield good spectral responses.

3.7.2 Super-Resolution

APEX and Pavia University

The numerical results for the two aerial datasets are displayed in Table 3.1 (best results in bold). SupResPALM achieves lower errors than all four baselines, in all metrics. The difference is more pronounced in the case of APEX, which is the harder dataset, due to the only partial spectral overlap between HSI and MSI. The effect becomes evident in Figure 3.6 where the RMSE per band is plotted for both images. For APEX (left) there is a sudden increase in the RMSE value around band 100, where the spectral sensitivity of the ADS80 ends. Of course, also our method has higher errors in this extrapolated part of the spectrum, still it achieves lower errors than the baselines. For Pavia University, the RMSE values are fairly constant throughout the spectrum, except near the two ends of the spectral coverage, in which wavelengths the channels are noisier.

<table>
<thead>
<tr>
<th>Method</th>
<th>APEX</th>
<th>Pavia University</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>ERGAS</td>
</tr>
<tr>
<td>SNNMF</td>
<td>8.82</td>
<td>3.23</td>
</tr>
<tr>
<td>HySure</td>
<td>8.75</td>
<td>3.15</td>
</tr>
<tr>
<td>R-FUSE</td>
<td>9.29</td>
<td>3.53</td>
</tr>
<tr>
<td>CNMF</td>
<td>9.17</td>
<td>3.35</td>
</tr>
<tr>
<td>SupResPALM</td>
<td>8.23</td>
<td>3.02</td>
</tr>
</tbody>
</table>

Table 3.1: Quantitative results for APEX and Pavia University. SAM, spectral angle mapper.
Moreover, we visualise the errors of all methods, in two different bands. We do this for APEX, since it is the more challenging scene, reconstruction errors are bigger, and the scene has a larger variety of distinguishable surface materials. In Figure 3.7 we plot the errors, in 8-bit range, for wavelengths 552 and 1506 nm. The first wavelength corresponds to a low reconstruction error (overlap with MSI), while the second one lies outside the MSI range and has high reconstruction error.

Figure 3.6: Per-band RMSE for all spectral images of APEX and Pavia University. (a) APEX; (b) Pavia University.

Figure 3.7: Top row: True colour image for APEX and ground truth spectral images for wavelengths 552 and 1506 nm. Second and third rows: The reconstruction error for each method at the mentioned wavelengths. Note the difference in the colour scale between the two rows.
Moreover, we plot the cumulative histogram of per-pixel residuals. In Figure 3.8 all pixels of a scene are sorted by their RMSE (norm of residual vector across all channels), in ascending order. Horizontal cuts correspond to the number of pixels below a fixed RMSE tolerance, vertical cuts correspond to the minimum error one must accept when using only a fixed number of “best” pixels. From these graphs it can be seen that our method has the largest amount of pixels with the low reconstruction errors. Note the different scaling of the axes; the gains are actually greater for APEX than for Pavia.

Figure 3.8: Cumulative histogram of per-pixel RMSE. (a) Pavia University; (b) APEX.

CAVE and Harvard

For CAVE and Harvard we focus on the numerical results. Please refer to Lanaras et al. (2015b) for visualisations. Tables 3.2 and 3.3 show the error metrics (best results in bold) that each method achieved in both full datasets. We only report average and median RMSE, ERGAS, SAM and \( Q_2^n \) values, as well as the number of images that each method reconstructed best. Complete per-image numbers are available in the supplementary material. Again SupResPALM outperforms all baselines in a clear majority of images. The errors and differences are greater on CAVE for all methods, which has a larger number and variability of materials per image. We also report for how many images each method returned the best result, in each error metric. That number gives an intuition whether a method is consistently better than another one across different images.

Real EO-1 Data

The main challenge of this dataset, compared to the simulated data, is that the input images originate from two different sensors with individually different, non-Gaussian noise. Moreover, like in the APEX scene, the spectral ranges of the HSI and the MSI do not fully overlap. The numerical results of the super-resolution are given in Table 3.4. We test SupResPALM in four configurations, where we compare different values for the sparsity of the endmembers and the spatial regularisation. Empirically, the spatial regularisation has only a small influence, whereas limiting the average number of active endmembers per pixel with the sparsity constraint, Equation (3.6e), produces a noticeable increase in reconstruction error. In this previous experiment the reconstruction is...
### Table 3.2: Quantitative results for the CAVE Database.

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>ERGAS</th>
<th>SAM</th>
<th>Q2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aver. Med.</td>
<td>3.15</td>
<td>1.81</td>
<td>0.927</td>
<td>0.934</td>
</tr>
<tr>
<td>1st</td>
<td>1.87</td>
<td>1.32</td>
<td>0.924</td>
<td>0.953</td>
</tr>
<tr>
<td>2nd</td>
<td>2.88</td>
<td>1.93</td>
<td>0.927</td>
<td>0.956</td>
</tr>
<tr>
<td>3rd</td>
<td>3.17</td>
<td>1.84</td>
<td>0.926</td>
<td>0.958</td>
</tr>
<tr>
<td>SupResPALM</td>
<td>1.81</td>
<td>1.35</td>
<td>0.925</td>
<td>0.961</td>
</tr>
<tr>
<td>Bic. upsampling</td>
<td>1.81</td>
<td>1.32</td>
<td>0.924</td>
<td>0.953</td>
</tr>
</tbody>
</table>

### Table 3.3: Quantitative results for the Harvard Database.

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>ERGAS</th>
<th>SAM</th>
<th>Q2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aver. Med.</td>
<td>3.15</td>
<td>1.81</td>
<td>0.927</td>
<td>0.934</td>
</tr>
<tr>
<td>1st</td>
<td>1.87</td>
<td>1.32</td>
<td>0.924</td>
<td>0.953</td>
</tr>
<tr>
<td>2nd</td>
<td>2.88</td>
<td>1.93</td>
<td>0.927</td>
<td>0.956</td>
</tr>
<tr>
<td>3rd</td>
<td>3.17</td>
<td>1.84</td>
<td>0.926</td>
<td>0.958</td>
</tr>
<tr>
<td>SupResPALM</td>
<td>1.81</td>
<td>1.35</td>
<td>0.925</td>
<td>0.961</td>
</tr>
<tr>
<td>Bic. upsampling</td>
<td>1.81</td>
<td>1.32</td>
<td>0.924</td>
<td>0.953</td>
</tr>
</tbody>
</table>

Table 3.2: Quantitative results for the CAVE Database.

Table 3.3: Quantitative results for the Harvard database.
done with relative response functions estimated from the data with SupResPALM. We further test the SupResPALM super-resolution with HySure responses and vice versa, to separate the estimation of the response functions from the super-resolution itself, Table 3.5. Not all baselines allow one to input the relative responses. SNNMF has memory issues and cannot handle the estimated spatial response function, whereas CNMF assumes a fixed spatial response and lets the user chose only the spectral response. In R-FUSE it is possible to input both response functions, so we include it in Table 3.5. It turns out that small variations in the relative sensor characteristics do not have a significant impact on the super-resolution. While SupResPALM super-resolution, with different reasonable response functions, outperforms HySure and R-FUSE by a clear margin.

The reconstructed Hyperion images of both methods, as well as the ground truth, are shown as colour composite in Figure 3.9. SupResPALM produces visibly fewer and weaker artifacts than HySure and R-FUSE. Nevertheless, some areas do exhibit spectral distortions, e.g., the small lake above the river in the middle of the image. These qualitative and quantitative results are, in our view, a much stronger indication than purely synthetic experiments that SupResPALM can be applied to real data.

### Table 3.4: Quantitative results of SupResPALM for Real EO-1 data.

<table>
<thead>
<tr>
<th>EO-1 Real Data</th>
<th>Method</th>
<th>RMSE</th>
<th>ERGAS</th>
<th>SAM</th>
<th>Q2\textsuperscript{P}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bicubic upsampling</td>
<td>5.99</td>
<td>12.58</td>
<td>4.06</td>
<td>0.547</td>
</tr>
<tr>
<td></td>
<td>SupResPALM ( \lambda = 0, s = 5N_m )</td>
<td>3.63</td>
<td>13.49</td>
<td>3.01</td>
<td>0.765</td>
</tr>
<tr>
<td></td>
<td>SupResPALM ( \lambda = 0.01, s = 5N_m )</td>
<td>3.61</td>
<td>13.52</td>
<td>3.10</td>
<td>0.764</td>
</tr>
<tr>
<td></td>
<td>SupResPALM ( \lambda = 0 )</td>
<td>3.48</td>
<td>13.50</td>
<td>2.85</td>
<td>0.774</td>
</tr>
<tr>
<td></td>
<td>SupResPALM ( \lambda = 0.01 )</td>
<td><strong>3.39</strong></td>
<td>13.58</td>
<td><strong>2.80</strong></td>
<td><strong>0.776</strong></td>
</tr>
</tbody>
</table>

Figure 3.9: Colour composite of Hyperion bands 15, 23, 30 corresponding to RGB. (Top Left) The ground truth HSI. (Bottom Left) Reconstruction of R-FUSE with the HySure relative responses. (Top Right) Reconstruction of HySure. (Bottom Right) Reconstruction of SupResPALM with \( \lambda = 0.01 \).
Table 3.5: Comparison of SupResPALM and HySure on real EO-1 data, separating estimation of response functions and super-resolution. Best results with bold, second best in Italics.

<table>
<thead>
<tr>
<th></th>
<th>Relative Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HySure</td>
</tr>
<tr>
<td>R-FUSE</td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>3.89</td>
</tr>
<tr>
<td>ERGAS</td>
<td>13.81</td>
</tr>
<tr>
<td>SAM</td>
<td>3.40</td>
</tr>
<tr>
<td>$Q^2_n$</td>
<td>0.747</td>
</tr>
<tr>
<td>HySure</td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>4.77</td>
</tr>
<tr>
<td>ERGAS</td>
<td>14.18</td>
</tr>
<tr>
<td>SAM</td>
<td>3.97</td>
</tr>
<tr>
<td>$Q^2_n$</td>
<td>0.698</td>
</tr>
<tr>
<td>SupResPALM</td>
<td></td>
</tr>
<tr>
<td>$\lambda = 0.01$</td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>3.34</td>
</tr>
<tr>
<td>ERGAS</td>
<td>13.73</td>
</tr>
<tr>
<td>SAM</td>
<td>2.75</td>
</tr>
<tr>
<td>$Q^2_n$</td>
<td>0.775</td>
</tr>
</tbody>
</table>

3.7.3 Discussion

Spectral Unmixing

Introducing the spectral unmixing constraints has proven to give good numerical results and plausible mixtures of materials. In Figure 3.10 are the abundances of selected endmembers, which correspond to easily recognisable surface materials observed in the APEX scene. Even though these are not always “pure materials” in the physical sense, due to non-linearities like inter-reflections, specularities and shadows, they look realistic, and comparatively clean. This confirms the conventional wisdom that the LMM is sufficient for many hyperspectral imaging problems, including in particular super-resolution. We also note that enforcing non-negativity of the solution and the sum-to-one constraint has very pragmatic advantages for further processing. Without the explicit constraints, artifacts do occur, like negative reflectance values, or pixels with reflectance zero in all bands. These mistakes may disturb further processing. As a simple example, for the spectral power zero the computation of the SAM will fail (for our baselines, we have excluded such pixels from the error computation).

SupResPALM has one main user parameter, namely $p$, the number of the endmembers (the subspace dimension). Note that classic dimensionality estimation methods do not apply in our case, where both the basis and the coefficients are constrained. If the main goal is super-resolution (and not accurate and unambiguous unmixing), one can in our experience use the upper bounds $p = 30$ for remote sensing scenes with hundreds of bands, and $p = 10$ for close-range scenes with $\approx 30$ bands.
3.7. EXPERIMENTAL RESULTS AND DISCUSSION

Effect of the Sparsity Term

As discussed in Section 3.5, our framework allows one to explicitly limit the average number of active endmembers per pixel, via Equation (3.6e). Although the simplex constraint favours sparse solutions, sometimes it may be desired to enforce stronger sparsity, by suppressing small abundance values. Sparser solutions will typically yield more realistic endmembers and abundances, by eliminating “garbage collection” endmembers that fit systematic effects like shading variations. However, in general, they will produce worse reconstruction results, because eliminating systematic errors obviously benefits the super-resolution, even if done by “misusing” the linear mixing model. For instance, a single, complex material can exhibit spectral shifts due to shading effects and give rise to two slightly different endmembers. Eliminating one of them improves the correspondence between endmembers and materials, but increases the super-resolution error. In Table 3.6 we show a quantitative evaluation on APEX and Pavia University for different \( s/N_m \) ratios ranging from 2–6. Both images initially have \( \approx 20 \) average active endmembers per-pixel with basic SupResPALM. As we push down that number, the accuracy slowly drops, as expected. While the drop for Pavia University becomes rather severe, APEX can be super-resolved acceptably well with only \( 3N_m \) non-zero abundances (RMSE still lower than several baselines). We attribute this behavior to the scene content of Pavia University, where many small objects exist that cause mixed pixels.

The extra sparsity is also applied to the CAVE database (Table 3.7) with the \( s/N_m \) ratio between six and two. Since the CAVE images are taken from close range and have a limited number of mixed pixels, restricting \( s = 3N_m \) still gives a better reconstruction than most baselines, while maintaining a plausible explanation about the scene’s endmembers and abundances. A similar effect is observed for the Harvard database. For

Figure 3.10: Abundances of different endmembers observed for the APEX scene. (Top) Scene in natural colours, meadows/lawn, tennis court (clay), (bottom) beach volleyball courts (sand), tree canopies, water and dark surfaces.
CHAPTER 3. HYPERSPECTRAL SUPER-RESOLUTION

details, see the supplementary material. Again, increased sparsity through a stricter $s$ appears to reduce over-fitting of the unmixing and lead to physically more plausible endmembers by suppressing overly small abundances. However, as shown in Tables 3.4, 3.6 and 3.7, this will usually not improve the super-resolution.

Table 3.6: Effect of the sparsity parameter $s$ for APEX and Pavia University. The first row is the selected sparsity level (average number of non-zero endmembers per pixel). The values in the first column are without the additional sparsity term.

<table>
<thead>
<tr>
<th>$s/N_m$</th>
<th>APEX</th>
<th>Pavia University</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>ERGAS</td>
</tr>
<tr>
<td></td>
<td>8.23</td>
<td>3.02</td>
</tr>
<tr>
<td>–</td>
<td>8.29</td>
<td>3.06</td>
</tr>
<tr>
<td>6</td>
<td>8.46</td>
<td>3.14</td>
</tr>
<tr>
<td>5</td>
<td>8.68</td>
<td>3.24</td>
</tr>
<tr>
<td>4</td>
<td>9.07</td>
<td>3.42</td>
</tr>
<tr>
<td>3</td>
<td>10.31</td>
<td>4.01</td>
</tr>
<tr>
<td>2</td>
<td>2.28</td>
<td>0.79</td>
</tr>
<tr>
<td>6</td>
<td>2.61</td>
<td>0.93</td>
</tr>
<tr>
<td>5</td>
<td>2.79</td>
<td>0.98</td>
</tr>
<tr>
<td>4</td>
<td>3.18</td>
<td>1.09</td>
</tr>
<tr>
<td>3</td>
<td>4.29</td>
<td>1.36</td>
</tr>
<tr>
<td>2</td>
<td>7.87</td>
<td>2.26</td>
</tr>
</tbody>
</table>

Table 3.7: Effect of the sparsity parameter $s$ for CAVE. The first row is the selected sparsity level (average number of non-zero endmembers per pixel). The values in the first column are without the additional sparsity term.

<table>
<thead>
<tr>
<th>$s/N_m$</th>
<th>CAVE Database</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
</tr>
<tr>
<td></td>
<td>3.15</td>
</tr>
<tr>
<td>–</td>
<td>3.18</td>
</tr>
<tr>
<td>6</td>
<td>3.23</td>
</tr>
<tr>
<td>5</td>
<td>3.32</td>
</tr>
<tr>
<td>4</td>
<td>3.77</td>
</tr>
<tr>
<td>3</td>
<td>5.18</td>
</tr>
</tbody>
</table>

Effect of the Spatial Regularisation

We have also tested the spectral unmixing approach without the spatial regularisation on APEX and Pavia University images (Table 3.8). We conjecture that spatial smoothing may in fact not be all that important, if the super-resolution is sufficiently constrained by other means. In particular, we found that the simplex constraint (3.6c) and (3.6d), that abundances must be non-negative and sum to 1 very effectively stabilises the prediction, such that further regularisation may not be necessary; while they have not been used in this form before. As it happens, our method performs almost equally well with $\lambda = 0$ (no spatial regularisation). There are differences, but they are very small, and not always in favour of spatial smoothing, see Table 3.8. However, in the EO-1 experiment the spatial regularisation does improve the super-resolution, see Table 3.4. We believe this is due to the fact that a realistic (slightly spatially variant) misalignment is present between the EO-1 images that where actually acquired by distinct sensors, the effect of which can be mitigated by moderate smoothing. While further work is needed to clarify the role of smoothing, at this point we recommend to retain it for real applications with separate MSI and HSI sensors.
Table 3.8: Effect of the spatial regularisation on APEX and Pavia.

<table>
<thead>
<tr>
<th>Method</th>
<th>APEX</th>
<th>Pavia University</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>ERGAS</td>
</tr>
<tr>
<td>SupResPALM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>λ = 0.1</td>
<td>8.23</td>
<td>3.02</td>
</tr>
<tr>
<td>λ = 0</td>
<td><strong>8.19</strong></td>
<td>2.98</td>
</tr>
</tbody>
</table>

Denoising

Hyperspectral images are often quite noisy, mainly due to sensor limitations and atmospheric effects. Figure 3.11 shows the first band (≈400 nm) of the Pavia University image. On the left is the original, noisy “ground truth” image, on the right the super-resolution reconstruction by SupResPALM. Obviously, much of the noise has been removed, because the high-frequency content of the reconstruction is obtained from the MSI, which is less corrupted as the spectral integration cancels noise (no matter whether it is done in software or by wider band-pass filters). We point out that this rather beneficial effect does not show up in the quantitative results: on the contrary, it increases the deviation from the noisy ground truth (for all methods). The effect is even stronger for Hyperion (EO-1), where many of the short-wave infrared bands have very low power and are very noisy, see Figure 3.12. At the top is an amplified colour composite of bands 125, 173 and 181 of the high resolution HSI, below is our reconstruction. The fusion has produced a much clearer image, with clear and homogeneous geographic features that are barely observable in the original. Similar results have also been reported by Cerra et al. (2014) in the context of hyperspectral image denoising.

Figure 3.11: Pavia University. (Left) The first channel of the ground truth image corresponding to approximately 400 nm. (Right) The first channel reconstructed from SupResPALM. The ground truth exhibits strong noise, whereas in the reconstructed image, it is heavily reduced.
Figure 3.12: EO-1. (Top) Colour composite of the ground truth image for the 1397-, 1881- and 1961-nm wavelengths. (Bottom) Same colour composite of reconstructed image from our method. The ground truth exhibits strong noise, whereas in the reconstructed image, it is heavily reduced.

3.8 Conclusions

We have proposed a method for hyperspectral super-resolution. Its basic idea is to jointly solve the spectral unmixing problem for both input images. Linking super-resolution to unmixing allows one to constrain the solution to conform with the elementary physics of optical imaging. We argue that this might be a more suitable way to regularise the inherently ill-posed super-resolution task. Given a spatially-coarse hyperspectral image and a spectrally-coarse, high-resolution MSI image, we estimate an image with high spatial and spectral resolution. As a side effect, our method additionally delivers a set of spectral endmembers and physically-plausible decomposition of each pixel into those endmembers. Under the linear mixing model, the proposed approach boils down to two coupled, constrained least-squares problems, which can be solved reliably with a projected gradient scheme. We found that updating also the basis, rather than only the mixing coefficients, during super-resolution gives better results, and recommend it over the sequential processing of many existing methods. Additionally, we have experimented with a spatial smoothness term driven by the gradients of the high-resolution MSI. Spatial regularisation is typically the first and most obvious strategy used to regularise ill-posed image processing problems. However, we found that, while it is straight-forward to include in our method, it only marginally improves the result. We thus do not generally recommend it; it seems that if one has access to other, less “diffuse” a priori knowledge like our unmixing constraints, explicit spatial smoothness may not be as important. Finally, we have expanded our method to work with real data, where the spatial and spectral response functions are not perfectly known, by estimating them from the images themselves. Again, we prefer to restrict the inference to plausible solutions, in this case symmetric, unimodal responses, but avoid arbitrary priors chosen for convenience (such as Gaussian responses). In simulations on four public datasets, the proposed SupResPALM method
3.8. CONCLUSIONS

has shown excellent performance. Moreover, we have applied our method to real remote sensing data from two different satellite sensors (Hyperion and ALI), where it also worked well. While the EO-1 satellite carrying Hyperion and ALI has recently been decommissioned, a number of missions are underway that will bring new hyperspectral sensors to space in the next few years (e.g., EnMAP, PRISMA, HISUI). These missions will also benefit from improved data fusion and super-resolution. A worthwhile extension of the presented method will be to deal with cases where the HSI and MSI have not been acquired from the same platform. In that situation, additional problems arise, such as precise co-registration, handling differing viewing angles and handling temporal changes between different acquisition times.

A limitation of the current method is that linear mixing does not adequately capture the actual behavior of some important materials, in particular high vegetation including for example forest and many agricultural crops. In the current framework, non-linear BRDF effects like shading, specular reflection, cast shadows and direction-dependent reflectance are phenomenologically dissipated with additional “virtual” endmembers. For highly non-linear materials or when correct unmixing matters, one may have to extend the model to explicitly include more complex shading behavior. Furthermore, it is known that the spatial response of many sensors depends on the wavelength and sometimes also on the location in the field of view, whereas we assume a spatially isotropic and band-independent blur. Again, this does not seem to negatively affect super-resolution; still, it may be desirable to recover more detailed sensor characteristics, especially when using them in the further analysis of the data.
Chapter 4

Super-Resolution of Multispectral Multiresolution Images from a Single Sensor

Charis Lanaras, Jose Bioucas-Dias, Emmanuel Baltsavias, Konrad Schindler
IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1505-1513, 2017

(Author version; for typeset version please refer to the original conference paper.)

4.1 Abstract

Some remote sensing sensors, acquire multispectral images of different spatial resolutions in variable spectral ranges (e.g., Sentinel-2, MODIS). The aim of this research is to infer all the spectral bands, of multiresolution sensors, in the highest available resolution of the sensor. We formulate this problem as a minimisation of a convex objective function with an adaptive (edge-reserving) regulariser. The data-fitting term accounts for individual blur and downsampling per band, while the regulariser “learns” the discontinuities from the higher resolution bands and transfers them to other bands. We also observed that the data can be represented in a lower-dimensional subspace, reducing the dimensionality of the problem and significantly improving its conditioning. In a series of experiments with simulated data, we obtain results that outperform state-of-the-art, while showing competitive qualitative results on real Sentinel-2 data.
4.2 Introduction

In remote sensing there are a growing number of sensors that acquire multi-spectral images in which the spatial resolution (or Ground Sampling Distance – GSD) varies across different spectral bands. This is mainly the case for satellite sensors and prominent examples are MODIS, ASTER, VIIRS, Worldview-3 and Sentinel-2 (S2). The underlying reason is that design considerations, sensor hardware limitations, and further influences like atmospheric absorption, necessitate the use of different spatial resolutions for various channels, so as to achieve a satisfactory signal-to-noise ratio (SNR). It is unlikely that all such resolution differences will go away with hardware improvements, thus it is natural to try and improve the resolution of the coarser bands computationally, by exploiting the structure in the optical signal.

Without loss of generality, we focus here on Sentinel-2, which records 13 bands at 3 different (spatial) resolutions, with good geometric and spectral details. A list of the S2 bands is given in Table [4.1]. The frequent revisit rate, global access and freely available data constitute a great utility for a wide range of applications based on remote sensing. Sentinel-2 data is, among others, used for the monitoring of vegetation, soil, water cover, inland waterways and coastal areas, as well as for the estimation of geophysical variables.

![Sentinel-2 input](image1.png) ![Result of SupReME](image2.png)

Figure 4.1: Results on real Sentinel-2 images with a false colour composite of bands (B5, B6, B7). Left: The input 20 m resolution bands. Right: The super-resolved image.

The aim of the present work is to increase the spatial resolution of lower-resolution bands, such that all bands have the same, maximal resolution. This is obviously desirable from a product point of view, since then all further down-stream products can be
4.3. RELATED WORK

derived at high spatial resolution and can create new opportunities, e.g., better observation of clouds/ice/snow and better estimation of climate variables. The environment and climate applications can especially benefit from this approach, since they require many bands in the infrared spectrum.

Interestingly, this problem can be seen as an extension of pan-sharpening, with the main differences that (i) there can be more than one channel at the highest resolution, and (ii) the high-resolution bands need not spectrally overlap the lower-resolution ones. These two conditions of classical pan-sharpening, however, are not met by any of the mentioned instruments. Some recent work has explored an intermediate solution, where a single high-resolution band (a “virtual panchromatic image”) is synthesized and used as input to a standard pan-sharpening approach. We take one step further and disconnect from the classic pan-sharpening paradigm, integrating channels of different spectral sensitivity and spatial resolution into a compact imaging model.

**Contribution.** We propose a high-quality solution for this “smart upsampling”. Our method relies strongly on the observation model of the imaging (blurring and downsampling) process that generates the low-resolution images. Given the nature of blurring and downsampling, the inversion of the model is an ill-posed problem. To overcome this, we exploit the fact that spectral bands are correlated and can thus be represented in a lower-dimensional subspace, where most of the energy is contained. We learn this subspace from the input data, to reduce the number of unknowns and further stabilise the computation. Moreover, we fully exploit the textural information from the high resolution bands to tailor a regulariser, by encoding the discontinuities of the data and propagating the spatial information to the lower resolution bands. The main underlying assumption is that the discontinuities are likely to be located in the same location across all bands. Given the quadratic data term and the adaptive quadratic regulariser we formulate a convex problem to invert the observation model, and propose an efficient numerical solver. The proposed method performs, in a single step, a super-resolution for all lower resolution bands to the maximum resolution. We term our method **SupReME** (SUPer-REsolution for multispectral Multiresolution Estimation).

4.3 Related Work

The problem of enhancing the spatial resolution based on spectral bands from the same sensor has been approached for (at least) MODIS, ASTER, and more recently for Sentinel-2 and VIIRS. Tonooka (2005) performs super-resolution for the thermal infrared and shortwave infrared bands of ASTER, using the visible and near infrared bands, by means of spectral similarity. Similarly, Aiazzi et al. (2005) super-resolve the ASTER TIR channels by injecting spatial detail from the VNIR channels using a Generalised Laplacian Pyramid method. Also to super-resolve the ASTER TIR channels, Fasbender et al. (2008) propose a general Bayesian data fusion approach. Thermal sharpening of VIIRS data was investigated by Picaro et al. (2016), by considering various “thermal sharpening” methods.
A few studies also increase the resolution of the 500 m bands in MODIS. Sirguey et al. (2008) use wavelet-based multiresolution analysis, based on the ARSIS concept (Ranchin and Wald, 2000), in an "injection-type" method, which was applied for sub-pixel monitoring of seasonal snow cover (Sirguey et al., 2009). Trishchenko et al. (2006) proposed a non-linear regression and normalisation to preserve radiometric consistency of the super-resolved channels. This method was further applied for producing clear-sky, cloud and cloud shadow masks at 250 m resolution (Luo et al., 2008). In technically similar work, Wang et al. (2015) propose a pan-sharpening method termed ATPRK (area-to-point regression kriging) that also starts from regression modeling, then residual upsampling in order to comply with the spectral properties of the lower-resolution bands.

The same method (ATPRK) has been applied to sharpen the 20 m bands of S2 (Wang et al., 2016) with the help of the 10 m bands. Since ATPRK only accepts one high resolution image, it is proposed to use either the spectrally nearest 10 m band, or an average of all 10 m bands. Du et al. (2016) compare four different pan-sharpening methods to sharpen the SWIR band (B11) of S2. Their application is to compute the Modified Normalised Difference Water Index (MNDWI) at 10 m resolution, for monitoring of water bodies. Moreover, Vaiopoulos and Karantzalos (2016) compared 21 different fusion algorithms to sharpen the VNIR and SWIR bands of Sentinel-2. Surprisingly, they report that standard bicubic interpolation outperforms sophisticated pan-sharpening methods in terms of the QNR metric (Quality with No Reference, Alparone et al. (2008)). They believe that this is, at least in part, due to a bias in the QNR metric. Most recently, Brodu (2017) introduced a so-called "geometry of scene elements" to unmix (super-resolve) the low-resolution bands of S2, while preserving their overall reflectance. This is the only work we are aware of that also enhances the low resolution bands (60 m).

We take a somewhat different approach: We set up a joint model of the physical imaging process in all the different channels, and invert that model with variational calculus.

### 4.4 Problem Formulation

Without loss of generality, we present our model for the case of S2. Our inputs have \( L_1 = 4 \) high resolution bands with GSD 10 m, \( L_2 = 6 \) medium resolution bands (20 m), and \( L_6 = 3 \) low resolution bands (60 m), in total \( L = L_1 + L_2 + L_6 \) spectral bands. The output shall have 10 m resolution for all bands, thus the upsampling factors are \( r_1 = 1 \), \( r_2 = 2 \), \( r_6 = 6 \), and a fixed image area contains \( n = n_1 = n_1/r_1^2 \) high-resolution pixels, \( n_2 = n/r_2^2 \) medium-resolution pixels, or \( n_6 = n/r_6^2 \) low-resolution pixels. To derive our method it is convenient to vectorise the images: the pixel intensities of each individual band are collected into a vector \( y_i \), and the bands are then concatenated (in arbitrary, fixed order) into \( y = (y_1, y_2, \ldots, y_L) \in \mathbb{R}^{L_1n_1+L_2n_2+L_6n_6} \). Similarly, the unknown output image is \( x = (x_1, x_2, \ldots, x_L) \in \mathbb{R}^{Ln} \), except that now all bands have the same, maximal, resolution, \( x_i \in \mathbb{R}^n \). Accordingly, the output bands can also be reformatted into a matrix when necessary, \( X = [x_1^T; x_2^T; \ldots; x_L^T] \in \mathbb{R}^{L \times n} \), and we have \( x = \text{vec}(X^\top) \).
The input and output images are, up to noise, related through the following observation model:

\[ y = MBx , \quad (4.1) \]

where \( M \in \mathbb{R}^{(L_1n_1 + L_2n_2 + L_6n_6) \times L_n} \) and \( B \in \mathbb{R}^{L_n \times L_n} \) are two block diagonal matrices. Every sub-block, in both of them, acts on one spectral band. For \( M \), the blocks represent the sampling of \( x \) to obtain \( y \), i.e., blocks are identity matrices for high-resolution channels, and downsampling functions (masks) for other channels, i.e., a subset of rows of the identity matrix. The blur matrix \( B \) is a block-circulant-circulant-block (BCCB) matrix, where each block represents a 2D cyclic convolution, associated with the point spread function (PSF) of the corresponding band at the resolution of \( x \) (highest spatial resolution). The blur can be different in every spectral band¹ and is assumed to be spatially invariant. The model in eq. (4.1) in its raw form is obviously ill-posed, since there are fewer observations than unknowns.

### 4.4.1 Subspace representation

It has often been observed that multi-spectral image data are correlated and can be projected into a subspace of lower dimensionality without losing information. For the 13 bands of S2, we find that > 99% of the signal energy is retained in the \( p = 6 \) largest components of the correlation-based eigen-decomposition. The dimensionality reduction significantly reduces the number of unknowns, and proves to be a (implicitly or explicitly used) key ingredient for the envisaged super-resolution. Formally, the columns of \( X \) (i.e., spectral vectors) live in a subspace spanned by the columns of \( U \in \mathbb{R}^{L \times p} \) and thus we may write:

\[ X = UZ , \quad (4.2) \]

where \( Z \in \mathbb{R}^{p \times n} \) are the representation coefficients with respect to \( U \). We assume that \( U \) is semi-unitary. Vectorising the matrices yields

\[ x = \text{vec}(X^\top) = \text{vec}(IZ^\top U^\top) = (U \otimes I)\text{vec}(Z^\top) = (U \otimes I)z , \quad (4.3) \]

where \( I \) is an identity matrix with suitable dimensions.

With the dimensionality reduction, the observation model of eq. (4.1) becomes

\[ y = MB(U \otimes I)z . \quad (4.4) \]

Note that, algebraically, the reduced problem is no longer ill-posed, as long as \( pn < L_1n_1 + L_2n_2 + L_6n_6 \). However, it is still very ill-conditioned. A direct solution is not practical, as it is extremely sensitive to the presence of noise (even with low magnitude).

**Estimation of the subspace.** To compute the subspace \( U \), we need to have access to \( X \), respectively to the span(\( X \)). However, we only have a version of \( X \) that is blurred and downsampled, namely \( y \). Our objective is to estimate the \( \text{span}(X) \), using a blurred

¹This is why we use the vector notation \( x \) and \( y \). Matrix format is impractical if one allows for band-specific blur.
version of \(X, XQ\), where \(Q \in \mathbb{R}^{n \times n}\) is a blurring matrix representing a 2D cyclic convolution. The main motivation behind this is that the blurring operator \(Q\) does not affect the span of \(X\), i.e., \(\text{span}(X) = \text{span}(XQ)\), given that it is a linear combination of the columns of \(X\) and that \(\text{rank}(Q) \geq p\). To do so (i) we upsample all the bands of \(y\) to the same high resolution, using bicubic interpolation and (ii) we blur each band, such that the blur of all the bands is more or less the same, i.e. equivalent to the strongest blur. This image serves as the best approximation of \(XQ\). Next, we perform singular value decomposition analysis on the blurred data. We retain the first \(p\) left singular vectors (with decreasing order of singular values) as the columns of \(U\). We assume in this way that the columns of \(U\) span the same subspace as the columns of \(X\).

### 4.5 Proposed Solution

To invert our image model, we solve the optimisation

\[
\min_z \frac{1}{2} \| MB(U \otimes I)z - y \|^2 + \lambda \phi_{w, q}(D_h z, D_v z),
\]

where \(\phi_{w, q}\) is a regularisation term, based on weights \(w, q\) (see Sec. 4.5.1) and \(\lambda\) is the regularisation strength. \(D_h, D_v \in \mathbb{R}^{Ln \times Ln}\) are two block-diagonal linear operators (each with identical blocks) that approximate horizontal and vertical derivatives of the images in \(z\). For simplicity, we treat these matrices with periodic boundary conditions as cyclic convolutions. For the regulariser, we choose the quadratic form

\[
\phi_{w, q}(z) = \sum_{i=1}^{p} \sum_{j=1}^{n} \left\{ q_i w_j (H_h z_i)_{ij}^2 + q_i w_j (H_v z_i)_{ij}^2 \right\},
\]

where the index \(i\) runs over the subspace dimensions and \(j\) over all pixels for basis vector \(i\). \(H_h, H_v \in \mathbb{R}^{n \times n}\) are the individual blocks of the finite difference operators (derivatives) \(D_h\) and \(D_v\), respectively, such that \(D_h = I \otimes H_h\) and \(D_v = I \otimes H_v\).

To solve the minimisation eq. (4.5), we use C-SALSA [Afonso et al., 2010], an instance of ADMM, as follows:

\[
\min_{z, v_1, v_2, v_3} \frac{1}{2} \| MBv_1 - y \|^2 + \lambda \phi_{w, q}(v_2, v_3)
\]

subject to \(v_1 = (U \otimes I)z\)

\[
\begin{aligned}
v_2 &= D_h z \\
v_3 &= D_v z
\end{aligned}
\]

This splitting makes it possible to solve each individual problem in a much easier way.
The Augmented Lagrangian of the above problem is:

\[
\mathcal{L}(z, v_1, v_2, v_3, d_1, d_2, d_3) = \frac{1}{2} \| MBv_1 - y \|^2 + \frac{\mu}{2} \| (U \otimes I)z - v_1 - d_1 \|^2 \\
+ \lambda \phi_{w,q}(v_2, v_3) + \frac{\mu}{2} \| D_hz - v_2 - d_2 \|^2 + \frac{\mu}{2} \| D_vz - v_3 - d_3 \|^2,
\]

where \( d_1, d_2 \) and \( d_3 \) are the scaled Lagrange multipliers and \( \mu \) is a positive weight. The solution of eq. (4.8) with respect to \( z \) is

\[
z = (I + D_hD_h^\top + D_vD_v^\top)^{-1} \left\{ (U \otimes I)(v_1 + d_1) + D_h^\top(v_2 + d_2) + D_v^\top(v_3 + d_3) \right\},
\]

given that \( U^\top U = I \). The matrix to be inverted is block diagonal, from the definition of \( D_h \) and \( D_v \). We can therefore solve separately for each subspace dimension of \( z \).

Even so, the numerical inversion of a block element would be impossible on a normal computer in terms of memory and operation count. The trick is to exploit the structure of \( D_h \) and \( D_v \). They are both BCCB matrices and thus, one can compute the solution in the frequency domain (Simões et al., 2015). In the following we minimise the Lagrangian eq. (4.8) with respect to \( v_1 \):

\[
v_1 = (B^\top M^\top MB + \mu I)^{-1} \{ B^\top M^\top y + \mu ((U \otimes I)z - d_1) \}.
\]

Recently, it has been shown that systems of equations involving this type of matrices can be efficiently solved in the frequency domain (Wei et al., 2016b). The inversion is done independently for each diagonal block, i.e., each dimension of the subspace. The unknown \( v_2 \) and \( v_3 \) in eq. (4.8) is obtained from

\[
v_2, v_3 \in \arg \min_{v_2, v_3} \lambda \phi_{w,q}(v_2, v_3) + \frac{\mu}{2} \| D_hz - v_2 - d_2 \|^2 + \frac{\mu}{2} \| D_vz - v_3 - d_3 \|^2.
\]

The solution is uncoupled with respect to \( v_2 \) and \( v_3 \) and can be computed for each element as

\[
v2_{ij} = \frac{\mu(H_hz_i - d_{2i})}{\mu + 2\lambda w_{i}d_{j}}, \quad v3_{ij} = \frac{\mu(H_vz_i - d_{3i})}{\mu + 2\lambda w_{i}d_{j}},
\]

where \( i \) represents the \( i \)th image in \( z \) and \( j \) each pixel.

Finally, we update the Lagrange multipliers as:

\[
d_1^{(k)} = d_1^{(k-1)} - ((U \otimes I)z - v_1) \\
d_2^{(k)} = d_2^{(k-1)} - (D_hz - v_2) \\
d_3^{(k)} = d_3^{(k-1)} - (D_vz - v_3).
\]

The complete optimisation scheme is summarised in Algorithm 2. Initial values for the parameters \( v_1, v_2, v_3, d_1, d_2 \) and \( d_3 \) can be chosen arbitrarily, since the problem is convex. The convergence to the global minimum is guaranteed for any \( \mu > 0 \), see Afonso et al. (2010).
Algorithm 2: SupReME. Solver for the optimisation of (4.5).

Require: \( \text{data:} \ y, \ \text{sensor blurs:} \ B, \ \text{regularisation parameter} \ \lambda, \ \text{weights} \ q \)

Estimate the subspace \( U \) and weights \( w \) from eq. (4.14, 4.15)

\[
k \leftarrow 0
\]

Initialise \( v_1^{(0)}, v_2^{(0)}, v_3^{(0)}, d_1^{(0)}, d_2^{(0)} \) and \( d_3^{(0)} \)

while not converged do

\[
k \leftarrow k + 1
\]

Estimate \( z^{(k)} \) with eq. (4.9)

Estimate \( v_1^{(k)}, v_2^{(k)}, v_3^{(k)} \) with eq. (4.10, 4.12)

Estimate \( d_1^{(k)}, d_2^{(k)}, d_3^{(k)} \) with eq. (4.13)

end while

return \( x = (U \otimes I)z \)

4.5.1 Adapting the spatial regularisation

For the quadratic smoothing used as spatial regulariser, we can introduce weights \( w \) for each pixel, so as to reduce smoothing across discontinuities. We take

\[
w = \exp\left(-\frac{g_{\text{max}}^2}{2\sigma_s}\right)
\]  (4.14)

and define \( g_{\text{max}} = \max(g(x_1), \ldots, g(x_L)) \), of the high resolution bands, where the \( \max(\cdot) \) operator is applied element-wise (per pixel), \( g(\cdot) \) is the Prewitt image gradient magnitude. The weights eq. (4.14) are in the range \([0 \ldots 1]\). We found that very strong edges are downweighted too strongly, and truncate values smaller than 0.5:

\[
w = \max(0.5, w).
\]  (4.15)

Empirically, this improves the result.

By their definition, the subspace coefficients have different value ranges. The first few dimensions (basis vectors) cover most of the energy (information), and thus are – relatively – less affected by noise, whereas in the last dimensions the noise is more dominant. It is therefore necessary to apply progressively stronger regularisation. We simply reweight the regulariser with \( q \in \mathbb{R}^p \), in an heuristic fashion. At this stage the best setting is

\[
q = [1 \ 1.5 \ 4 \ 8 \ 15 \ 15 \ 20].
\]  (4.16)

We tried using principal component based schemes, but it was not leading to better results compared to the heuristic approach, that performed well across all tested datasets.

4.5.2 Implementation details

Our method requires the PSF of each band. The variance of the Gaussian blur (termed sdf) included in \( B \) differs for each band, and is computed from the calibrated MTF (Modulation Transfer Function) supplied by ESA as part of the meta-data
4.6. EXPERIMENTAL RESULTS

(Clerc and MPC Team [2017]):

\[ \text{sdf} = r_i \sqrt{\frac{-2 \ln(\text{mtf})}{\pi^2}}, \]  

(4.17)

with \( i = 1, 2, 6 \) for the different resolutions. The highest resolution bands are assumed to have a point response (\( \text{sdf} = 0 \)). Empirically, the estimation is not very sensitive to small variations of the blur. For better numerical conditioning we also normalise the images before the processing, such that their mean squared intensities are 1 (normalisation has been undone for the qualitative results below). As mentioned before, the blur \( B \) and the difference matrices \( D_h \) and \( D_v \) have periodic boundary conditions. For this reason we remove a border of 18 pixels, to suppress artifacts originating from the periodic boundaries. The subspace dimension is set to \( p = 7 \), as it was able to capture the spectral variability in all our experiments. The value \( \mu = 0.2 \) is kept constant for the whole processing, and the normalisation of the weights \( w \) is set to \( \sigma_s = 1 \).

As convergence criterion for the optimiser, we iterate until the residuals associated with the variable splitting fall below 0.001, or at most 200 iterations. With this criterion we have obtained a solution for every scenario we tested. For an image of dimensions 180×180 pixels the program runs on average less than 10 seconds on an Intel Xeon E5 3.2 GHz CPU, in a MATLAB implementation.

4.6 Experimental Results

Quality indices. As a primary quality metric for the evaluation we use the signal-to-reconstruction error (SRE), given in dB. It is computed from the reconstructed image \( \hat{x} \) and the ground truth \( x \) as:

\[ \text{SRE} = 10 \log_{10} \frac{\mu_x^2}{\|x - \hat{x}\|^2/n}, \]  

(4.18)

where \( \mu_x \) is the average value of \( x \). As a complementary metric we also report the Spectral Angle Mapper (SAM, [Yuhas et al., 1992]) for each image, defined as

\[ \text{SAM} = \frac{1}{n(L_2 + L_6)} \sum \arccos \frac{X_{j}X_{\hat{j}}}{\|X_{j}\|2 \|X_{\hat{j}}\|2}, \]  

(4.19)

where \( X_{j}, \hat{X}_{j} \) are the spectral values of the \( L_2 \) and \( L_6 \) bands at pixel \( j \), of the ground truth and reconstruction respectively. The SAM is computed only for the super-resolved bands (per pixel) and is averaged over the whole image. It is given in degrees.

The QNR value has also been used to judge super-resolution results [Vaiopoulos and Karantzalos, 2016]. However, as mentioned in Sec. 4.3 the result can be misleading. The QNR was primarily designed to evaluate pan-sharpening results and appears unsuitable to compare conceptually different types of algorithms [Vivone et al., 2015].
Table 4.2: Numerical results on simulated data, SRE in dB and SAM (last column) in degrees. The input APEX data are in their original resolution (2 m). Note that, the resolution mentioned does not correspond with the actual resolution of the image, but gives an overview of the resolution of the S2 bands. Best results in bold font.

<table>
<thead>
<tr>
<th>Method</th>
<th>B1 (60m)</th>
<th>B5 (20m)</th>
<th>B6 (20m)</th>
<th>B7 (20m)</th>
<th>B8a (20m)</th>
<th>B9 (20m)</th>
<th>B11 (20m)</th>
<th>B12 (20m)</th>
<th>SAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicubic upsampling</td>
<td>8.39</td>
<td>14.97</td>
<td>15.88</td>
<td>15.86</td>
<td>15.93</td>
<td>12.66</td>
<td>19.55</td>
<td>17.12</td>
<td>7.80</td>
</tr>
<tr>
<td>MTF-GLP-HPM-PP</td>
<td>–</td>
<td>19.38</td>
<td>21.66</td>
<td>21.48</td>
<td>22.42</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>ARTPK</td>
<td>–</td>
<td>25.29</td>
<td>29.80</td>
<td>30.35</td>
<td>31.08</td>
<td>–</td>
<td>20.93</td>
<td>20.45</td>
<td>–</td>
</tr>
<tr>
<td>SupReME</td>
<td>22.01</td>
<td>25.16</td>
<td>27.87</td>
<td>31.29</td>
<td>32.12</td>
<td>19.78</td>
<td>21.42</td>
<td>21.16</td>
<td>4.63</td>
</tr>
</tbody>
</table>

Baselines. As baselines to compare against, we use the following 4 methods. As the simplest solution to the problem we use a bicubic upsampling, implemented in MATLAB. The second baseline is MTF-GLP-HPM-PP, which is an injection-based pan-sharpening method, initially presented by Lee and Lee (2010). That method performed best against other pan-sharpening methods in Vaiopoulos and Karantzalos (2016) and we follow the same experimental procedure as described there. The third baseline, termed SuperRes, is the method of Brodu (2017). The last baseline is ARTPK Wang et al. (2016), which reported the best results for our task that we are aware of. From the two variants described in that publication (“selected” and “synthesised”), we compare against “synthesised”, which is the average of all high resolution bands and gave the best numerical results. For all baselines, code has been made available, and we use those original implementations.

4.6.1 Simulated data

For the simulated data, we use the Open Science Dataset of APEX Schaepman et al. (2015), available online[^2]. APEX is a hyperspectral sensor that covers the wavelengths from 0.4 to 2.5 µm and can be used to simulate S2 images, given the nominal spectral response of S2 (Table 4.1). We took care to model the spectral downgrading as well as possible, by using a few more bands, if the inputs were noisy. From the APEX scenes, we created two sets of inputs with different resolutions, and consequently different levels of object detail. The first set has maximum resolution ≈2 m, which is the nominal resolution of the APEX image. The second one has maximum resolution ≈10 m, for a strict simulation of S2. The corresponding medium and low resolution images are 2 and 6 times degraded respectively, by pixel aggregation with the known PSF of S2. Band B10 is discarded from all further processing. That band is primarily there to detect Cirrus clouds, but it does not show any ground structures, and is always very noisy (both in real S2 data and when created from APEX). The dimensions of the image used are in both cases 180 × 180. Gaussian noise of SNR = 40 dB has been added to all simulated images. In the current version of SupReME we use \( \lambda = 0.005 \) for APEX.

[^2]: http://www.apex-esa.org/content/free-data-cubes
Table 4.3: Numerical results on simulated data, SRE in dB and SAM (last column) in degrees. The input APEX data has been downsampled to match the true GSD of S2 bands. Best results in bold font.

<table>
<thead>
<tr>
<th>Method</th>
<th>B1</th>
<th>B5</th>
<th>B6</th>
<th>B7</th>
<th>B8a</th>
<th>B9</th>
<th>B11</th>
<th>B12</th>
<th>SAM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>60m</td>
<td>20m</td>
<td>20m</td>
<td>20m</td>
<td>20m</td>
<td>60m</td>
<td>20m</td>
<td>20m</td>
<td></td>
</tr>
<tr>
<td>Bicubic upsampling</td>
<td>8.35</td>
<td>16.46</td>
<td>17.59</td>
<td>17.51</td>
<td>17.54</td>
<td>13.20</td>
<td>16.70</td>
<td>13.47</td>
<td>5.77</td>
</tr>
<tr>
<td>MTF-GLP-HPM-PP</td>
<td>–</td>
<td>2.53</td>
<td>23.37</td>
<td>18.71</td>
<td>23.92</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>ARTPK</td>
<td>–</td>
<td>25.47</td>
<td>30.84</td>
<td>31.34</td>
<td>31.84</td>
<td>–</td>
<td>21.09</td>
<td>17.14</td>
<td>–</td>
</tr>
<tr>
<td>SupReME</td>
<td>20.64</td>
<td>26.36</td>
<td>29.85</td>
<td>34.26</td>
<td>34.23</td>
<td>23.69</td>
<td>19.95</td>
<td>16.41</td>
<td>2.71</td>
</tr>
</tbody>
</table>

Tabulated numerical results for 2 m and 10 m super-resolution are presented in Tables 4.2 and 4.3, respectively. The best results are marked in bold font. ARTPK and SupReME perform best, with SupReME dominating in bands B1, B7, and B8a and ARTPK performing well in B6. Most challenging of all are bands B11 and B12, which lie in the SWIR range, far away from all high-resolution bands. B1 and B9 also have slightly weaker results, simply because the original (input) bands are of low resolution. To be fair, SAM is reported in Tables 4.2 and 4.3 for methods that reconstruct all channels. Here, SupReME always delivers the best result. That means that the errors are distributed more evenly across the 8 bands compared to other methods. This is especially useful for applications where spectral properties of materials (their proportions) are important. To compare to ARTPK in terms of SAM, we only evaluate the spectral angle on the 6 bands with 20 m resolution. In both cases SupReME is considerably better and achieves a smaller spectral angle: in the full resolution APEX (2 m), 3.30 vs 4.43 and in the reduced resolution APEX (10 m), 2.31 vs 2.98.

Figure 4.2 visually compares the results of all baselines to the 10 m ground truth, using a false colour composite of bands (B8a, B6, B5). As expected, Bicubic upsampling and SuperRes lack high-frequency detail and appear blurry. MTF-GLP-HPM-PP is better, but exhibits a number of artifacts. ARTPK and SupReME are clearly best and are visually almost indistinguishable from the ground truth. Figure 4.3 compares the two top-performing methods ARTPK and SupReME to the 2 m ground truth on a scene detail, using a composite of bands (B12, B11, B7). SupReME in this case gives the best result, whereas ARTPK introduces artifacts at spectral discontinuities, c.f. the stadium tartan or the grass area near the top left corner. Notice that both reconstruction results are significantly sharper than the ground truth, likely due to the imaging system of APEX. Its VNIR and SWIR sensors have rather different PSF. In contrast, super-resolution transfers high-resolution detail from spectral bands in different wavelengths, which are sharp. In that sense, computational super-resolution can also be seen as a way to remedy wavelength-dependent blur in imaging systems that record all pixels at the same nominal resolution, by exploiting correlations across spectral bands.
CHAPTER 4. SUPER-RESOLUTION OF MULTIRESOLUTION IMAGES

4.6.2 Real Sentinel-2 data

Complementary to simulated data, we use real Sentinel-2a data, acquired on 29th of April 2016 over the Czech Republic in the vicinity of Prague. In this work we use a cut-out of $180 \times 180$ pixels that contains a variety of scene elements, close to the Prague airport, see Figure 4.4. Since we have no access to the true high-resolution information, we cannot perform any quantitative evaluation.

For this data we use a stronger regularisation $\lambda = 0.01$. The visual results in Figure 4.1 and 4.4 show two different false colour composites, with bands (B5, B6, B7) and (B12, B9, B11). In both cases the result of SupReME is compared to the input lower resolution images, resized with nearest neighbour to be able to overlay them with equal pixel size. One clearly sees that, also for real S2 data, the result is much sharper than the input, and does not seem to suffer from any visible spectral (colour) distortion.

4.7 Discussion

So far, work on S2 super-resolution has largely disregarded the potential of the 60 m channels. Even though these bands are primarily intended for atmospheric corrections,
4.7. DISCUSSION

Figure 4.3: Comparison of ARTPK and SupReME to the ground truth (APEX with 2 m resolution). The false colour composite is created with bands (B12, B11, B7).

Figure 4.4: Qualitative results on real S2 images. Left: The scene in true colours (10 m GSD). Centre and Right: A colour composite of bands (B12, B9, B11) before and after the super-resolution.

we find that they can be upsampled to 10 m, with an accuracy similar to that of the 20 m bands. Although upsampling 1 pixel to 36 may at first sight seem unrealistic, the results indicate that the spectral correlations are strong enough to do it, and these bands could be used to derive high-resolution information, rather than only for calibrating the other high-resolution bands.

We have observed that our approach is not independent of the texture scale (respectively, object size) in the input image. Thus, its performance might change across different sensor resolutions, or across scenes with very different object scales, even if recorded with the same sensor. In future work it should therefore be investigated how well our method generalises to much lower image resolutions (e.g., MODIS). Such images have much smaller homogeneous areas and correspondingly more high-frequency texture. We note that, also for such low-resolution, global imagery, super-resolution may have interesting uses. E.g., enhancing the even lower resolution of some spectral bands to 250 m could make quite a difference for coastlines and coastal mapping.
In our current setup, we have not attempted to automatically set the parameters $p, \lambda$ and $q$. Fixed, empirically chosen values gave good results, but for larger areas of interest it may become necessary to choose them automatically, or even to adaptively change the subspace to best represent the local spectral distribution.

### 4.8 Conclusions

We have presented a novel method to deal with the different resolutions in contemporary multi-spectral satellite sensors. Our method integrates information from all channels, independent of their input resolution, into one joint, convex optimisation. The model is computationally efficient, and super-resolves all lower-resolution channels to the highest available resolution, by transferring high-frequency information from multiple, non-overlapping high-resolution bands to the entire data cube. The method has been tested on both simulated data and real Sentinel-2 imagery, and achieves state-of-the-art results.

Future extensions shall include locally adaptive subspace projection to deal with strongly varying, locally heterogeneous scene content. Eventually, we aim to super-resolve entire S2 tiles ($100 \times 100 \text{ km}^2$). Moreover, as longer time series become available, we would like to investigate high-resolution change detection. It is an interesting open question how to best integrate time series analysis with super-resolution.

### Acknowledgements

The authors acknowledge financial support from the Swiss National Science Foundation (SNSF), project no. 200021_162998 and from Portuguese Fundação para a Ciência e a Tecnologia (FCT) under the grants UID/EEA/5008/2013 and ER-ANETMED/0001/2014.

### 4.9 Supplementary Information: Not available in the published manuscript

In this chapter, we present further information on the solution of eq. (4.10) from Chapter 4. It can be seen as supplementary material to the paper. We recall the solution for $v_1$ as:

$$v_1 = (B^\top M^\top MB + \mu I)^{-1}\{B^\top M^\top y + \mu((U \otimes I)z - d_1)\}.$$

(4.20)

Here we will focus on how to invert the matrix on the left. From the properties of $B$ and $M$ we know that the matrix to be inverted is block diagonal and thus can be inverted individually per block. Effectively, the inversion of each band is independent of the other
bands. We follow a general case below for a band, with a dimension of $m_1$ rows, $m_2$ columns and a downsampling rate of $d$.

According to [Wei et al. (2015b)] the bluring operation can be efficiently expressed in the Fourier space, as $B = FDF^H$ and $B^\top = FD^*F^H$, where $F$ is the 2D DFT (Direct Fourier Transform) matrix, $(\cdot)^*$ denotes the conjugate operator and $(\cdot)^H$ denotes the conjugate transposed of a matrix. The conjugate transposed is used for complex matrices instead of the transpose. Note that the matrix $D$ is diagonal and simplifies the computations. The expression becomes:

\[
(B^\top M^\top M B + \mu I)^{-1} \quad (4.21)
\]

\[
=(F D^*F^H M^\top M F D^*F^H + \mu I)^{-1}. \quad (4.22)
\]

As we are working with a vectorised image, $F$ is composed of $F_1 \in \mathbb{C}^{m_1 \times m_1}$ and $F_2 \in \mathbb{C}^{m_2 \times m_2}$ the respective 1D DFT matrices under the expression $F = F_2 \otimes F_1$. $F_1$ acts on the columns (with length $m_1$) of the image and $F_2$ acts on the rows (with length $m_2$). Similarly, $M = M_2 \otimes M_1$, where $M_1 \in \mathbb{R}^{m_1 \times m_1}$ downsamples rows and $M_2 \in \mathbb{R}^{m_2 \times m_2}$ columns respectively. The symbol $\otimes$ depicts the Kronecker product of two matrices. Part of eq. (4.22) becomes (with Kronecker product properties):

\[
(F^H M^\top M F) \quad (4.23)
\]

\[
=(F_2 \otimes F_1)^H (M_2 \otimes M_1)^\top (M_2 \otimes M_1) (F_2 \otimes F_1) \quad (4.24)
\]

\[
=(F_2^H M_2^\top M_2 F_2) \otimes (F_1^H M_1^\top M_1 F_1) \quad (4.25)
\]

We recall the lemma 2 of [Wei et al. (2015b)] where $F^H M^\top M F = \frac{1}{d} (1_d \otimes I_m)(1_d^\top \otimes I_m)$, valid for 1D DFT matrices. Substituting this into eq. (4.25) we get:

\[
\left(\frac{1}{d^2} (1_d \otimes I_{m_2})(1_d^\top \otimes I_{m_2}) \otimes (1_d \otimes I_{m_1})(1_d^\top \otimes I_{m_1}) \right)
\]

\[
\left\{ (1_d \otimes I_{m_2}) \otimes (1_d \otimes I_{m_1}) \right\} \quad (4.26)
\]

\[
= \frac{1}{d^2} (1_d \otimes I_{m_2})(1_d^\top \otimes I_{m_2}) \otimes (1_d \otimes I_{m_1})(1_d^\top \otimes I_{m_1}) \quad (4.27)
\]

At this point we define the matrix $K = (1_d \otimes I_{m_2}) \otimes (1_d \otimes I_{m_1})$, with $K \in \mathbb{R}^{d^2 m_2 m_1 \times m_1 m_2}$. The expression in eq. (4.22) can now be written as:

\[
(F D^*F^H M^\top M F D^*F^H + \mu I)^{-1} \quad (4.29)
\]

\[
= \left( \frac{1}{d^2} F D^* F^H K^\top D^*F^H + \mu I \right)^{-1}. \quad (4.30)
\]

In the following we exploit the Woodbury matrix identity:
\[(A + UV)^{-1} = A^{-1} - A^{-1}U(I + VA^{-1}U)^{-1}VA^{-1}\]. Eq. (4.30) now is:

\[
\begin{align*}
\frac{1}{d^2}FD^*KK^HD^H + \mu I
\end{align*}
\]

\[
= I/\mu - FD^*K(d^2I + K^TD^HD^*K/\mu)^{-1}K^TD^H/\mu^2
\]

\[
= I/\mu - FD^*K(d^2I + K^TD^2K/\mu)^{-1}K^TD^H/\mu^2
\]

\[
= I/\mu - FD^*K(d^2I + K^TD^2K/\mu)^{-1}K^TD^H/\mu^2
\]

(4.34)

In the above we used the property of the DFT matrices \(F^H F = I\) and exploited the fact that the product of \(D\) and its conjugate \(D^*\) is \(D^2\). By studying the structure of \(K\) we observe that the effect of \(K^T Q K\) on any diagonal matrix \(Q\) (representing an image) is taking the sum over distinct \(m_1/d \times m_2/d\) image patches (e.g., in MATLAB this can be implemented with the built-in function \texttt{im2col}, with the option 'distinct', and in Python with the function \texttt{view.as_blocks} of the scikit-image package). The result of the matrix multiplication \(K^T Q K\) is a diagonal matrix, that makes the inversion of \((d^2I + K^TD^2K/\mu)\) as simple as inverting each diagonal element. On the contrary, the effect of \(KQK^T\) on any diagonal matrix \(Q\) (representing an image) is to replicate the contents of \(Q\) \(m_1/d \times m_2/d\) multiple times to get an image of size \(m_1 \times m_2\).

Eventually, the inversion of eq. (4.21) is multiplication in the Fourier space according to eq. (4.34)
Chapter 5

Super-Resolution of Sentinel-2 Images: Learning a Globally Applicable Deep Neural Network

Charis Lanaras, Jose Bioucas-Dias, Silvano Galliani, Emmanuel Baltsavias, Konrad Schindler
ISPRS Journal of Photogrammetry and Remote Sensing 146, 305–319, 2018
(Author version; for typeset version please refer to the original journal paper.)

5.1 Abstract

The Sentinel-2 satellite mission delivers multi-spectral imagery with 13 spectral bands, acquired at three different spatial resolutions. The aim of this research is to superresolve the lower-resolution (20 m and 60 m Ground Sampling Distance – GSD) bands to 10 m GSD, so as to obtain a complete data cube at the maximal sensor resolution. We employ a state-of-the-art convolutional neural network (CNN) to perform end-to-end upsampling, which is trained with data at lower resolution, i.e., from 40→20 m, respectively 360→60 m GSD. In this way, one has access to a virtually infinite amount of training data, by downsampling real Sentinel-2 images. We use data sampled globally over a wide range of geographical locations, to obtain a network that generalises across different climate zones and land-cover types, and can super-resolve arbitrary Sentinel-2 images without the need of retraining. In quantitative evaluations (at lower scale, where ground truth is available), our network, which we call DSen2, outperforms the best competing approach by almost 50% in RMSE, while better preserving the spectral characteristics. It also delivers visually convincing results at the full 10 m GSD.
5.2 Introduction

Several widely used satellite imagers record multiple spectral bands with different spatial resolutions. Such instruments have the considerable advantage that the different spectral bands are recorded (quasi-) simultaneously, thus with similar illumination and atmospheric conditions, and without multi-temporal changes. Furthermore, the viewing directions are (almost) the same for all bands, and the co-registration between bands is typically very precise. Examples of such multi-spectral, multi-resolution sensors include: MODIS, VIIRS, ASTER, Worldview-3 and Sentinel-2. The resolutions between the spectral bands of any single instrument typically differ by a factor of about 2–6. Reasons for recording at varying spatial resolution include: storage and transmission bandwidth restrictions, improved signal-to-noise ratio (SNR) in some bands through larger pixels, and bands designed for specific purposes that do not require high spatial resolution (e.g., atmospheric corrections). Still, it is often desired to have all bands available at the highest spatial resolution, and the question arises whether it is possible to computationally super-resolve the lower-resolution bands, so as to support more detailed and accurate information extraction. Such a high-quality super-resolution, beyond naive interpolation or pan-sharpening, is the topic of this paper. We focus specifically on super-resolution of Sentinel-2 images.

Sentinel-2 (S2) consists of two identical satellites, 2A and 2B, which use identical sensors and fly on the same orbit with a phase difference of 180 degrees, decreasing thus the repeat and revisit periods. The sensor acquires 13 spectral bands with 10 m, 20 m and 60 m resolution, with high spatial, spectral, radiometric and temporal resolution, compared to other, similar instruments. More details on the S2 mission and data are given in Section 5.4. Despite its recency, S2 data have been already extensively used. Beyond conventional thematic and land-cover mapping, the sensor characteristics also favour applications like hydrology and water resource management, or monitoring of dynamically changing geophysical variables.

E.g., Mura et al. (2018) exploit S2 to predict growing stock volume in forest ecosystems. Castillo et al. (2017) compute the Leaf Area Index (LAI) as a proxy for above-ground biomass of mangrove forests in the Philippines. Similarly, Clevers et al. (2017) retrieve LAI and leaf and canopy chlorophyll content of a potato crop. Delloye et al. (2018) estimate nitrogen uptake in intensive winter wheat cropping systems by retrieval of the canopy chlorophyll content. Paul et al. (2016) map the extent of glaciers, while Toming et al. (2016) map lake water quality. Immel et al. (2016) have demonstrated the use of S2 data for crop and tree species classification, and Pesaresi et al. (2016) for detecting built-up areas. The quality, free availability and world-wide coverage make S2 an important tool for (current and) future earth observation, which motivates this work.

Obviously, low-resolution images can be upsampled with simple and fast, but naive methods like bilinear or bicubic interpolation. However, such methods return blurry images with little additional information content. More sophisticated methods, including ours, attempt to do better and recover as much as possible of the spatial detail, through a “smarter” upsampling that is informed by the available high-resolution bands. Here,
we propose a (deep) machine learning approach to multi-spectral super-resolution, using convolutional neural networks (CNNs). The goal is to surpass the current state-of-the-art in terms of reconstruction accuracy, while at the same time to preserve the spectral information of the original bands. Moreover, the method shall be computationally efficient enough for large-area practical use. We train two CNNs, one for super-resolving 20 m bands to 10 m, and one for super-resolving 60 m bands to 10 m. Our method, termed DSen2, implicitly captures the statistics of all bands and their correlations, and jointly super-resolves the lower-resolution bands to 10 m GSD. See an example in Fig. 5.1. True to the statistical learning paradigm, we learn an end-to-end-mapping from raw S2 imagery to super-resolved bands purely from the statistics over a large amount of image data. Our approach is based on one main assumption, namely that the spectral correlation of the image texture is self-similar over a (limited) range of scales. I.e., we postulate that upsampling from 20 m to 10 m GSD, by transferring high-resolution (10 m) details across spectral bands, can be learned from ground truth images at 40 m and 20 m GSD; and similarly for the 60 m to 10 m case. Under this assumption, creating training data for supervised learning is simple and cheap: we only need to synthetically downsample original S2 images by the desired factor, use the downsampled version as input to generate original data as output.

In this way, one gains access to large amounts of training data, as required for deep learning: S2 data are available free of charge, covering all continents, climate zones, biomes and land-cover types. Moreover, we assert that the high-capacity of modern deep neural networks is sufficient to encode a super-resolution mapping which is valid across the globe. Fig. 5.2 and 5.3 show various land-cover types and geographical/climatic areas used for training and testing. It is likely that even better results could be achieved, if a user focusing on a specific task and geographic region retrains the proposed networks with images from that particular environment. In that case, one
can start from our trained network and fine-tune the network weights with appropriate training sites. However, our experiments show that even a single model, trained on a selected set of representative sites world-wide, achieves much better super-resolution than prior state-of-the-art methods for independent test sites, also sampled globally. That is, our network is not overfitted to a particular context (as often the case with discriminative statistical learning), but can be applied worldwide.

Extensive experimental tests at reduced scale (where S2 ground truth is available) show that our single, globally applicable network yields greatly improved super-resolution of all S2 bands to 10 m GSD. We compare our method to four other methods both quantitatively and qualitatively. Our approach achieves almost 50% lower RMSE than the best competing methods, as well as > 5 dB higher signal-to-reconstruction-error ratio and >30% improvement in spectral angle mapping. The performance difference is particularly pronounced for the Short-Wave Infrared (SWIR) bands and the 60 m ones, which are particularly challenging for super-resolution. For completeness, we also provide results for three “classical” pan-sharpening methods on the 20 m bands, which confirm that pan-sharpening cannot compete with true multi-band super-resolution methods, including ours. Importantly, we also train a version of our network at half resolution (80 → 40 m) and evaluate its performance on 40 → 20 m test data. While there is of course some loss in performance, the CNN trained in this way still performs significantly better than all other methods. This supports our assertion that the mapping is to a large extent scale-invariant and can be learned from training data at reduced resolution – which is important for machine learning approaches in general, beyond our specific implementation.

Summarising our contributions, we have developed a CNN-based super-resolution algorithm optimised for (but conceptually not limited to) S2, with the following characteristics: (i) significantly higher accuracy of all super-resolved bands, (ii) better preservation of spectral characteristics, (iii) favourable computational speed when run on modern GPUs, (iv) global applicability for S2 data without retraining, according to our (necessarily limited) tests, (v) generic end-to-end system that can, if desired, be retrained for specific geographical locations and land-covers, simply by running additional training iterations, (vi) free, publicly available source code and pre-trained network weights, enabling out-of-the-box super-resolution of S2 data.
5.3 Related work

Enhancing the spatial resolution of remotely sensed multi-resolution images has been addressed for various types of images and sensors, including for example ASTER (Tonooka, 2005, Fasbender et al., 2008), MODIS (Trishchenko et al., 2006, Sirguey et al., 2008), and VIIRS (Picaro et al., 2016). In the following, we differentiate three types of methods: pan-sharpening per band, inverting an explicit imaging model, and machine learning approaches. The first group increases the spatial resolution independently for each target band, by blending in information from a spectrally overlapping high-resolution band. It is therefore essentially equivalent to classical pan-sharpening, applied separately to the spectral region around each high-resolution band. Such an approach relies on the assumption that for each relevant portion of the spectrum there is one high-resolution band (in classical pan-sharpening the “panchromatic” one), which overlaps, at least partially, with the lower-resolution bands to be enhanced. That view leads directly to the inverse problem of undoing the spatial blur from the panchromatic to the lower-resolution texture. A number of computational tools have been applied ranging from straight-forward component substitution to multiresolution analysis, Bayesian inference and variational regularisation. For a few representative examples we refer to (Choi et al., 2011), (Lee and Lee, 2010) and (Garzelli et al., 2008). A recent review and comparison of pan-sharpening methods can be found in Vivone et al. (2015). The pan-sharpening strategy has also been applied directly to Sentinel-2, although the sensor does not meet the underlying assumptions: as opposed to a number of other earth observation satellites (e.g., Landsat 8) it does not have a panchromatic band that covers most of the sensor’s spectral range. In a comparative study Vaiopoulos and Karantzalos (2016) evaluate 21 pan-sharpening algorithms to enhance the 20 m visible and near infrared (VNIR) and short wave infrared (SWIR) bands of Sentinel-2, using heuristics to select or synthesise the “panchromatic” input from the (in most cases non-overlapping) 10 m bands. Wang et al. (2016) report some of the best results in the literature for their ATPRK (Area-To-Point Regression Kriging) method, which includes a similar band selection, performs regression analysis between bands at low resolution, and applies the estimated regression coefficients to the high-resolution input, with appropriate normalisation. Park et al. (2017) propose a number of modifications to optimise the band selection and synthesis, which is then used for pan-sharpening with component substitution and multiresolution analysis. Du et al. (2016), having in mind the monitoring of open water bodies, have tested four popular pan-sharpening methods to sharpen the B11 SWIR band of S2, in order to compute a high-resolution the normalised differential water index (NDWI). Further in this direction, Gasparovic and Jogun (2018) used five different pan-sharpening methods to enhance the resolution of the 20 m bands. Their goal was to investigate the effect of the sharpened images on a land-cover classification compared to naive nearest neighbour upsampling. Interestingly, the classification results improved for most of the methods.

The second group of methods attacks super-resolution as an inverse imaging problem under the variational, respectively Bayesian, inference frameworks. These model-based methods are conceptually appealing in that they put forward an explicit ob-
servation model, which describes the assumed blurring, downsampling, and noise processes. As the inverse problem is ill-posed by definition, they also add an explicit regulariser (in Bayesian terms an “image prior”). The high-resolution image is then obtained by minimising the residual error w.r.t. the model (respectively, the negative log-likelihood of the predicted image) in a single optimisation for all bands simultaneously. Brodi (2017) introduced a method that separates band-dependent spectral information from information that is common across all bands, termed “geometry of scene elements”. The model then super-resolves the low-resolution bands such that they are consistent with those scene elements, while preserving their overall reflectance. Lanaras et al. (2017b) adopt an observation model with per-band point spread functions that account for convolutional blur, downsampling, and noise. The regularisation consists of two parts, a dimensionality reduction that implies correlation between the bands, and thus lower intrinsic dimension of the signal; and a spatially varying, contrast-dependent penalisation of the (quadratic) gradients, which is learned from the 10 m bands. SMUSH, introduced in Paris et al. (2017), adopts an observation model similar to Lanaras et al. (2017b), but employs a different, patch-based regularisation that promotes self-similarity of the images. The method proceeds hierarchically, first sharpening the 20 m bands, then the coarse 60 m ones.

The third group of super-resolution methods casts the prediction of the high-resolution data cube as a supervised machine learning problem. In contrast to the two previous groups, the relation between lower-resolution input to higher-resolution output is not explicitly specified, but learned from example data. Learning methods (and in particular, deep neural networks) can thus capture much more complex and general relations, but in turn require massive amounts of training data, and large computational resources to solve the underlying, extremely high-dimensional and complex optimisation. We note that the methods described in the following were designed with the classic pan-sharpening problem in mind. Due to the generic nature of end-to-end machine learning, this does not constitute a conceptual problem: in principle, they could be retrained with different input and output dimensions. Obviously, their current weights are not suitable for Sentinel-2 upsampling. To the best of our knowledge, we are the first to apply deep learning to that problem. Masi et al. (2016) adapt a comparatively shallow three-layer CNN architecture originally designed for single-image (blind) super-resolution. They train pan-sharpening networks for Ikonos, GeoEye-1 and WorldView-2. Yang et al. (2017) introduced PanNet, based on the high-performance ResNet architecture (He et al., 2016). PanNet starts by upsampling the low-resolution inputs with naive interpolation. The actual network is fed with high-pass filtered versions of the raw inputs and learns a correction that is added to the naively upsampled images. PanNet was trained for Worldview-2, Worldview-3, and Ikonos. More recently, this concept has been further exploited in Scarpa et al. (2018). Learning based pan-sharpening networks are trained with relatively small amounts of data, presumably because of the high data cost. In this context, we point out that with deep learning one need not specify sensor characteristics like for instance spectral response functions. Rather, the sensor properties are implicit in the training data.

1 In CNN terminology, adding the upsampled input constitutes a “skip connection”.
Table 5.1: The 13 Sentinel-2 bands.

<table>
<thead>
<tr>
<th>Band</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
<th>B5</th>
<th>B6</th>
<th>B7</th>
<th>B8</th>
<th>B8a</th>
<th>B9</th>
<th>B10</th>
<th>B11</th>
<th>B12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centre wavelength [nm]</td>
<td>443</td>
<td>490</td>
<td>560</td>
<td>665</td>
<td>705</td>
<td>740</td>
<td>783</td>
<td>842</td>
<td>865</td>
<td>945</td>
<td>1380</td>
<td>1610</td>
<td>2190</td>
</tr>
<tr>
<td>Band width [nm]</td>
<td>20</td>
<td>65</td>
<td>35</td>
<td>30</td>
<td>15</td>
<td>15</td>
<td>20</td>
<td>115</td>
<td>20</td>
<td>20</td>
<td>30</td>
<td>90</td>
<td>180</td>
</tr>
<tr>
<td>Spatial Resolution [m]</td>
<td>60</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>10</td>
<td>20</td>
<td>60</td>
<td>60</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

Example-based super-resolution has been investigated in computer vision and image processing (e.g., [Freeman et al., 2002]), but mainly for single-image super-resolution. I.e., enhancing the spatial resolution of a single (RGB) image with the help of a prior learned from a suitable training set. The rise of deep learning has also advanced single-image super-resolution ([Lim et al., 2017], [Kim et al., 2016a]). Moreover, such super-resolution has been applied to Sentinel-2 and Landsat-8 images ([Pouliot et al., 2018]). All these works have in common that they predict images of higher spatial resolution, meaning that what is learned is a generic prior on the local structure of high-resolution images; whereas our method increases resolution of particular bands in a more informed and more accurate manner, by transferring the texture from available high-resolution bands; effectively learning a prior on the correlations across the spectrum (or, equivalently, on the high-resolution structure of some bands conditioned on the known high-resolution structure of other bands).

5.4 Input data

We use data from the ESA/Copernicus satellites Sentinel 2A and 2B. They were launched on June 23, 2015 and March 7, 2017, respectively, with a design lifetime of 7.25 years, potentially extendible up to 5 additional years. The two satellites are identical and have the same sun-synchronous, quasi-circular, near-polar, low-earth orbit with a phase difference of 180 degrees. This allows the reduction of the repeat (and revisit) periods from 10 to 5 days at the equator. The satellites systematically cover all land masses except Antarctica, including all major and some smaller islands. The main sensor on the satellites is a multispectral imager with 13 bands. Their spectral characteristics and GSDs are shown in Table 5.1. Applications of the 10 m and 20 m bands include: general land-cover mapping, agriculture, forestry, mapping of biophysical variables (for instance, leaf chlorophyll content, leaf water content, leaf area index), monitoring of coastal and inland waters, and risk and disaster mapping. The three bands with 60 m GSD are intended mainly for water vapour, aerosol corrections and cirrus clouds estimation. In actual fact they are captured at 20 m GSD and are down-sampled in software to 60 m, thus increasing the SNR. The first 10 bands cover the VNIR spectrum and are acquired by a CMOS detector for two bands (B3 and B4) with 2-line TDI (time delay and integration) for better signal quality. The last 3 bands cover

---

the SWIR spectrum and are acquired by passively cooled HgCdTe detectors. Bands B11 and B12 also have staggered-row, 2-line TDI. The swath width is 290 km. Intensities are quantised to 12 bit and compressed by a factor ≈2.9 with a lossy wavelet method (depending on the band). Empirical data quality has been quantified as: absolute geolocation accuracy (without ground control) of 11 m at 95.5% confidence, absolute radiometric uncertainty (except B10) <5%, and SNR values comply to the specifications with >27% margin.

Clerc and MPC Team (2018) report on further aspects of S2 data quality. The mean pairwise co-registration errors between spectral bands are 0.14–0.21 pixels (at the lower of the two resolutions) for S2A and 0.07–0.18 pixels for S2B, 99.7% confidence. This parameter is important for our application: good band-to-band co-registration is important for super-resolution, and S2 errors are low enough to ignore them and proceed without correcting band-to-band offsets. Moreover, data quality is very similar for Sentinel-2A and 2B, so that no separate treatment is required. B10 (in an atmospheric absorption window, included for cirrus clouds detection) has comparatively poor radiometric quality and exhibits across-track striping artifacts, and is excluded from many aspects of quality control. For that reason we also exclude it.

Potential sensor issues that could impair super-resolution would mainly be band-to-band misregistration (which is very low for S2), radiometric or geometric misalignments within a band (which do not seem to occur), and moving objects such as airplanes (which are very rare). The data thus fulfills the preconditions for super-resolution, and we did not notice any effects in our results that we attribute to sensor anomalies.

S2 data can be downloaded from the Copernicus Services Data Hub, free of charge. The data comes in tiles (granules) of 110 × 110 km² (~800MB per tile). For processing, we use the Level-1C top-of-atmosphere (TOA) reflectance product, which includes the usual radiometric and geometric corrections. The images are geocoded and orthorectified using the 90m DEM grid (PlanetDEM³) with a height (LE95) and planimetric (CE95) accuracy of 14 m and 10 m, respectively. We note that a refinement step for the Level-1C processing chain is planned, which shall bring the geocoding
accuracy between different passes to <0.3 pixels at 95% confidence, which will allow high-accuracy multi-temporal analysis.

In this study, we use data from both Sentinel 2A and 2B, acquired between December 2016 and November 2017, respectively July 2017 and November 2017. Fig. 5.3 shows the locations of the tiles used. They have been picked randomly, aiming for a roughly even distribution on the globe and for variety in terms of climate zone, land-cover and biome type. To simplify implementation and testing, we chose only tiles with no undefined (“black background”) pixels. Pointers to the exact tiles are included in our publicly available implementation (see below). Using this wide variety of scenes, we aim to train a globally applicable super-resolution network that can be applied to any S2 scene.

5.5 Method

We adopt a deep learning approach to Sentinel-2 super-resolution. The rationale is that the relation between the multi-resolution input and a uniform, high-resolution output data cube is a complex mixture of correlations across many (perhaps all) spectral bands, over a potentially large spatial context, respectively texture neighbourhood. It is thus not obvious how to design a suitable prior (regulariser) for the mapping. On the contrary, the underlying statistics can be assumed to be the same across different Sentinel-2 images. We therefore use a CNN to directly learn it from data. In other words, the network serves as a big regression engine from raw multi-resolution input patches to high-resolution patches of the bands that need to be upsampled. We found that it is sufficient to train two separate networks for the 20 m and 60 m bands. I.e., the 60 m resolution bands, unsurprisingly, do not contribute information to the upsampling from 20 to 10 m.

We point out that the machine learning approach is generic, and not limited to a specific sensor. For our application the network is specifically tailored to the image statistics of Sentinel-2. But the sensor-specific information is encoded only in the network weights, so it can be readily retrained for a different multi-resolution sensor.

5.5.1 Simulation process

CNNs are fully supervised and need (a lot of) training data, i.e., patches for which both the multi-resolution input and the true high-resolution output are known. Thus, a central issue in our approach is how to construct the training, validation and testing datasets, given that ground truth with 10 m resolution is not available for the 20 m and 60 m bands. Even with great effort, e.g., using aerial hyper-spectral data and sophisticated simulation technology, it is at present impossible to synthesise such data with the degree of realism necessary for faithful super-resolution. Hence, to become practically viable, our approach therefore requires one fundamental assumption: we posit that the transfer of spatial detail from high-resolution to low-resolution bands is scale-invariant and that it depends only on the relative resolution difference, but not on the absolute
GSD of the images. *i.e.* the relations between bands of different resolutions are *self-similar* within the relevant scale range. Note however, we require only a weak form of self-similarity: it is not necessary for our network to learn a “blind” generative mapping from lower to higher resolution. Rather, it only needs to learn how to transfer high frequency details from existing high-resolution bands. The literature on self-similarity in image analysis supports such an assumption (*e.g.*, Shechtman and Irani, 2007, Glasner et al., 2009), at least over a certain scale range. We emphasise that for our case, the assumption must hold only over a limited range up to $6 \times$ resolution differences, *i.e.*, less than one order of magnitude. In this way, virtually unlimited amounts of training data can be generated by synthetically downsampling raw Sentinel-2 images as required.

For our purposes, the scale-invariance means that the mappings between, say, $20 \rightarrow 10\text{ m}$ and $40 \rightarrow 20\text{ m}$ are roughly equivalent. We can therefore train our CNN on the latter and apply it to the former. If the assumed invariance holds, the learned spatial-spectral correlations will be correct. To generate training data with a desired scale ratio $s$, we downsample the original S2 data, by first blurring it with a Gaussian filter of standard deviation $\sigma = 1/s$ pixels, emulating the modulation transfer function ($mtf$) of S2. From the Data Quality Report (Clerc and MPC Team, 2018) we get a range of 0.44–0.55 for the point spread function ($psf$) of the bands, given the relation $psf = \sqrt{-2 \log(mtf)/\pi^2}$. Then we downsample by averaging over $s \times s$ windows, with $s = 2$ respectively $s = 6$. The process of generating the training data is schematised in Fig. 5.4. In this way, we obtain two datasets for training, validation and testing. The first dataset consists of “high-resolution” images at 20 m GSD and “low-resolution” images of 40 m GSD, created by downsampling the original 10 m and 20 m bands by a factor of 2. It serves to train a network for $2 \times$ super-resolution. The second one consists of images with 60 m, 120 m and 360 m GSD, downsampled from the original 10 m, 20 m and 60 m data. This dataset is used to learn a network for $6 \times$ super-resolution. We note that, due to unavailability of 10 m ground truth, quantitative analysis of the results must also be conducted at the reduced resolution. We chose the following strategy: to validate the self-similarity assumption, we train a network at quarter-resolution $80 \rightarrow 40$ m as well as one at half-resolution $40 \rightarrow 20$ m and verify that both achieve satisfactory performance on the ground truth 20 m images. To test the actual application scenario, we then apply the $40 \rightarrow 20$ m network to real S2 data to get $20 \rightarrow 10$ m super-resolution. However, the resulting 10 m super-resolved bands can only be checked by visual inspection.
5.5.2 20m and 60m resolution networks

To avoid confusion between bands and simplify notation, we collect bands that share the same GSD into three sets $A = \{B2, B3, B4, B8\}$ (GSD=10 m), $B = \{B5, B6, B7, B8a, B11, B12\}$ (GSD=20 m) and $C = \{B1, B9\}$ (GSD=60 m). The spatial dimensions of the high-resolution bands in $A$ are $W \times H$. Further, let $y_A \in \mathbb{R}^{W \times H \times 4}$, $y_B \in \mathbb{R}^{W/2 \times H/2 \times 6}$, and $y_C \in \mathbb{R}^{W/6 \times H/6 \times 2}$ denote, respectively, the observed intensities of all bands contained in sets $A$, $B$ and $C$. As mentioned above, we train two separate networks for different super-resolution factors. This reflects our belief that self-similarity may progressively degrade with increasing scale difference, such that $120 \to 60$ m is probably a worse proxy for $20 \to 10$ m than the less distant $40 \to 20$ m.

The first network upsamples the bands in $B$ using information from $A$ and $B$:

$$T_2 : \mathbb{R}^{W \times H \times 4} \times \mathbb{R}^{W/2 \times H/2 \times 6} \to \mathbb{R}^{W \times H \times 6}$$

$$\quad (y_A, y_B) \mapsto x_B,$$  \hspace{1cm} (5.1a)

where $x_B \in \mathbb{R}^{W \times H \times 6}$ denotes the super-resolved 6-band image with GSD 10 m. The second network upsamples $C$, using information from $A$, $B$ and $C$:

$$S_6 : \mathbb{R}^{W \times H \times 4} \times \mathbb{R}^{W/2 \times H/2 \times 6} \times \mathbb{R}^{W/6 \times H/6 \times 2} \to \mathbb{R}^{W \times H \times 2}$$

$$\quad (y_A, y_B, y_C) \mapsto x_C,$$  \hspace{1cm} (5.2a)

with $x_C \in \mathbb{R}^{W \times H \times 2}$ again the super-resolved 2-band image of GSD 10 m.

5.5.3 Basic architecture

Our network design was inspired by EDSR (Lim et al., 2017), state-of-the-art in single-image super-resolution. EDSR follows the well-known ResNet architecture (He et al., 2016) for image classification, which enables the use of very deep networks by using the so-called “skip connections”. These long-range connections bypass portions of the network and are added again later, such that skipped layers only need to estimate the residual w.r.t. their input state. In this way the average effective path length through the network is reduced, which alleviates the vanishing gradient problem and greatly accelerates the learning.

Our problem however, is different from classical single-image super-resolution. In the case of Sentinel-2, the network does not need to hallucinate the high-resolution texture only on the basis of previously seen images. Rather, it has access to the high-resolution bands to guide the super-resolution, i.e., it must learn to transfer the high-frequency content to the low-resolution input bands, and do so in such a way that the resulting (high-resolution) pixels have plausible spectra. Contrary to EDSR, where the upsampling takes place at the end, we prefer to work with the high (10 m) resolution from the beginning, since some input bands already have that resolution. We thus start by upsampling the low-resolution bands $y_B$ and $y_C$ to the target resolution (10 m) with simple bilinear interpolation, to obtain $\tilde{y}_B \in \mathbb{R}^{W \times H \times 6}$ and $\tilde{y}_C \in \mathbb{R}^{W \times H \times 2}$. The inputs
Algorithm 3 DSen2. Network architecture.

Require: high-resolution bands (A): $y_A$, low-resolution bands (B, C): $y_k$, feature dimensions $f$, number of ResBlocks: $d$

# Cubic interpolation of low resolution:
Upsample $y_k$ to $\tilde{y}_k$

# Concatenation:
$x_0 := [y_A, \tilde{y}_k]$

# First Convolution and ReLU:
$x_1 := \max(\text{conv}(x_0, f), 0)$

# Repeat the ResBlock module $d$ times:
for $i = 1$ to $d$ do
  $x_i = \text{ResBlock}(x_{i-1}, f)$
end for

# Last Convolution to match the output dimensions:
# where $b_{last}$ is either $6 (T_{2x})$ or $2 (S_{6x})$
$x_{d+1} := \text{conv}(x_d, b_{last})$

# Skip connection:
$x := x_{d+1} + \tilde{y}_B (T_{2x})$ or $x := x_{d+1} + \tilde{y}_C (S_{6x})$

return $x$

and outputs depend on whether the network $T_{2x}$ or $S_{6x}$ is used. To avoid confusion we define the set $k$ of low-resolution bands as either $k = \{B\}$ or $k = \{B, C\}$. Such that the input is $y_k$, and the addition (skip connection) to the output is $\tilde{y}_B$ for $T_{2x}$, respectively $\tilde{y}_C$ for $S_{6x}$. The proposed network architecture consists mainly of convolutional layers, ReLU non-linearities and skip connections. A graphical overview of the network is given in Fig. 5.5 and 5.6, pseudo-code for the network specification is given in Algorithm 3.

The operator $\text{conv}(x, f_{out})$ represents a single convolution layer, i.e., a multi-dimensional convolution of image $z$ with kernel $w$, followed by an additive bias $b$:

$$v = \text{conv}(x, f_{out}) := w \ast z + b \quad (5.3)$$

where $\ast$ is the convolution operator. The convolved image $v$ has the same spatial dimensions $(w \times h)$ as the input, as we use zero-padded convolution. The convolution kernels $w$ have dimensions $(k \times k)$. We always use $k = 3$, in line with the recent literature, which suggests that many layers of small kernels are preferable. The output feature dimension $f_{out}$ of the convolution (number of filters) depends only on $w$ and is required as an input. $f_{out}$ can be chosen for each convolutional layer, and constitutes a hyper-parameter of the network. Its selection is further discussed in Sec. 5.5.4. The input feature dimensions $f_{in}$ (depth of the filters) depend only on the input image $z$. The weights $w$ and $b$ are the free parameters learned during training and ultimately what the network has to learn.

The rectified linear unit (ReLU) is a simple non-linear function that truncates all nega-
5.5. METHOD

5.5.1. Residual Blocks

A residual block $v = \text{ResBlock}(z, f)$ is defined as a series of layers that operate on an input image $z$ to generate an output $z_4$, then adds that output to the input image (Fig. 5.6):

\begin{align}
    z_1 &= \text{conv}(z, f) \quad \# \text{convolution} \\
    z_2 &= \max(z_1, 0) \quad \# \text{ReLU layer} \\
    z_3 &= \text{conv}(z_2, f) \quad \# \text{convolution} \\
    z_4 &= \lambda \cdot z_3 \quad \# \text{residual scaling} \\
    v &= z_4 + z \quad \# \text{skip connection}
\end{align}

$\lambda$ is a custom layer (5.5d) that multiplies its input activations (multi-dimensional images) with a constant. This is also termed residual scaling and greatly speeds up the training of very deep networks (Szegedy et al., 2017). In our experience residual scaling is crucial and we always use $\lambda = 0.1$. As a alternative, we also tested the more common Batch Normalisation (BN), but found that it did not improve accuracy or training time, while increasing the parameters of the network. Also, Lim et al. (2017) report that BN normalises the features and thus reduces the range flexibility (the actual reflectance) of the images. Within each ResBlock module we only include a ReLU after the first convolution, but not after the second, since our network shall learn corrections to the bilinearly upsampled image, which can be negative. Within our network design, the ResBlock module can be repeated as often as desired. We show experiments with two different numbers $d$ of ResBlocks: 6 and 32. The final convolution at the head of

\[ \begin{array}{c}
\text{Concatenation} \\
\text{conv} \\
\text{ReLU} \\
\text{ResBlock} \\
\vdots \hspace{1cm} d \text{ times} \hspace{1cm} \vdots \\
\text{ResBlock} \\
\text{conv} \\
\text{Addition} \\
x_B \\
\end{array} \]
the network, after all ResBlocks, reduces the output dimension to $b_{last}$, such that it matches the number of the required output bands ($x_B$ and $x_C$). So $f_{out} = b_{last} = 6$ for $T_{2 \times}$, and $f_{out} = b_{last} = 2$ for $S_{6 \times}$ is used. 

A particularity of our network architecture is a long, additive skip connection directly from the rescaled input to the output (Fig. 5.5). This means that the complete network in fact learns the additive correction from the bilinearly upsampled image to the desired output. The strategy to predict the differences from a simple, robust bilinear interpolation, rather than the final output image, helps to preserve the radiometry of the input image.

### 5.5.4 Deep and very deep networks

Finding the right size and capacity for a CNN is largely an empirical choice. Conveniently, the CNN framework makes it possible to explore a range of depths with the same network design, thus providing an easy way of exploring the trade-off between small, efficient models and larger, more powerful ones. Also in our case, it is hard to know in advance how complex the network must be to adequately encode the super-resolution mapping. We introduce two configurations of our ResNet architecture, a deep (DSen2) and a very deep one (VDSen2). The names are derived from Deep Sentinel-2 and Very Deep Sentinel-2, respectively. For the deep version we use $d = 6$ and $f = 128$, corresponding to 14 convolutional layers, respectively 1.8 million tunable weights. For the very deep one we set $d = 32$ and $f = 256$, leading to 66 convolutional layers and a total of 37.8 million tunable weights. DSen2 is comparatively small for a modern CNN. The design goal here was a light network that is fast in training and prediction, but still reaches good accuracy. VDSen2 has a lot higher capacity, and was designed with maximum accuracy in mind. It is closer in terms of size and training time to modern high-end CNNs for other image analysis tasks (Simonyan and Zisserman, 2015; He et al., 2016; Huang et al., 2017), but is approximately two times slower and five times slower in both training and prediction respectively, compared to its shal-
lower counterpart (DSen2). Naturally, one can easily construct intermediate versions by changing the corresponding parameters $d$ and $f$. The optimal choice will depend on the application task as well as available computational resources. On the one hand, the very deep variant is consistently a bit better, while training and applying it is not more difficult, if adequate resources (i.e., high-end GPUs) are available. However, the gains are small compared to the $20 \times$ increase in free parameters, and it is unlikely that going even deeper will bring much further improvement.

### 5.5.5 Training details

As loss function we use the mean absolute pixel error ($L^1$ norm) between the true and the predicted high-resolution image. Interestingly, we found the $L^1$ norm to converge faster and deliver better results than the $L^2$ norm, even though the latter serves as error metric during evaluation. Most likely this is due to the $L^1$ norm’s greater robustness of absolute deviations to outliers. We did observe that some Sentinel-2 images contain a small number of pixels with very high reflectance, and due to the high dynamic range these reach extreme values without saturating.

Our learning procedure is standard: the network weights are initialised to small random values with the HeUniform method (He et al., 2015), and optimised with stochastic gradient descent (where each gradient step consists of a forward pass to compute the current loss over a small random batch of image patches, followed by backpropagation of the error signal through the network). In detail, we use the Adam variant of SGD (Kingma and Ba, 2014) with Nesterov momentum (Dozat, 2015). Empirically, the proposed network architecture converges faster than other ones we experimented with, due to the ResNet-style skip connections.

Sentinel-2 images are too big to fit them into GPU memory for training and testing, and in fact it is unlikely that long-range context over distances of a kilometre or more plays any significant role for super-resolution at the 10 m level. With this in mind, we train the network on small patches of $w \times h = (32 \times 32)$ for $T_{2\times}$, respectively $(96 \times 96)$ pixels for $S_{6\times}$. We note that this corresponds to a receptive field of several hundred metres on the ground, sufficient to capture the local low-level texture and potentially also small semantic structures such as individual buildings or small waterbodies, but not large-scale topographic features. We do not expect the latter to hold much information about the local pixel values, instead there is a certain danger that the large-scale layout of a limited training set is too unique to generalise to unseen locations.

As our network is fully convolutional, it can process input images of arbitrary spatial extent $w \times h$ (after padding to a multiple of the patch size). The tile size in the prediction step is limited only by the on-board memory on the GPU. To avoid boundary artifacts from tiling, adjacent tiles are cropped with an overlap of 2 low-resolution input pixels, corresponding to $40 \text{ m}$ for $T_{2\times}$, respectively $120 \text{ m}$ for $S_{6\times}$. 
### 5.6 Experimental results

#### 5.6.1 Implementation details

As mentioned before, we aim for global coverage. We therefore sample 60 representative scenes from around the globe, 45 for training and 15 for testing. For $T_{2\times}$ we sample 8000 random patches per training image, for a total of 360,000 patches. For $S_{6\times}$, we sample 500 patches per image for a total of 22,500 (note that each patch covers a $9\times$ larger area in object space and has $9\times$ more high-resolution pixels than for $T_{2\times}$). Out of these patches 90% are used for training the weights, the remaining 10% serve as validation set, see Table 5.2. To test the networks, we run both on the 15 test images, each with a size of $110\times110\text{km}^2$, which corresponds to $5,490\times5,490$ pixels at 20 m GSD, or $1,830\times1,830$ pixels at 60 m GSD.

Each network is implemented in the Keras framework [Chollet et al., 2015], with TensorFlow as back-end. Training is run on a NVIDIA Titan Xp GPU, with 12 GB of RAM, for approximately 3 days. The mini-batch size for SGD is set to 128 to fit into GPU memory. The initial learning rate is $lr = 1e^{-4}$ and it is reduced by a factor of 2 whenever the validation loss does not decrease for 5 consecutive epochs. For numerical stability we divide the raw $0 – 10,000$ reflectance values by 2000 before processing.

#### 5.6.2 Baselines and evaluation metrics

As baselines, we use the methods of [Lanaras et al., 2017b] – termed SupReME, [Wang et al., 2016] – termed ATPRK, and [Brodu, 2017] – termed Superres. Moreover, as elementary baseline we use bicubic interpolation, to illustrate naive upsampling without considering spectral correlations. Note, this also directly shows the effect of our network, which is trained to refine a bilinearly upsampled image. The input image sizes for the baselines were chosen to obtain the best possible results. SupReME showed the best performance when run with patches of 256, respectively 240 for $T_{2\times}$ and $S_{6\times}$. We speculate that this may be due to the subspace projection used within SupReME, which can better adapt to the local image content with moderate tile size. The remaining baselines performed best on full images. The parameters for all baselines were set...
as suggested in the original publications. This lead to rather consistent results across the test set.

The main evaluation metric of our quantitative comparison is the root mean squared error (RMSE), estimated independently per spectral band:

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum (\hat{x} - x)^2},
\]

where \( \hat{x} \) is each reconstructed band (vectorised), \( x \) is the vectorised ground truth band and \( n \) the number of pixels in \( x \). The unit of the Sentinel-2 images is reflectance multiplied by 10,000, however, some pixels on specularities, clouds, snow etc. exceed 10,000. Therefore, we did not apply any kind of normalisation, and report RMSE values in the original files’ value range, meaning that a residual of 1 corresponds to a reflectance error of \( 10^{-4} \).

Depending on the scene content, some images have higher reflectance values than others, and typically also higher absolute reflectance errors. To compensate for this effect, we also compute the signal to reconstruction error ratio (SRE) as additional error metric, which measures the error relative to the power of the signal. It is computed as:

\[
\text{SRE} = 10 \log_{10} \frac{\mu^2}{\| \hat{x} - x \|^2 / n},
\]

where \( \mu_x \) is the average value of \( x \). The values of SRE are given in decibels (dB). We point out that using SRE, which measures errors relative to the mean image intensity, is better suited to make errors comparable between images of varying brightness. Whereas the popular peak signal to noise ratio (PSNR) would not achieve the same effect, since the peak intensity is constant. Moreover, we also compute the spectral angle mapper (SAM), i.e., the angular deviation between true and estimated spectral signatures (Yuhas et al., 1992). We compute the SAM for each pixel and then average over the whole image. The values of SAM are given in degrees. This metric is complementary to the two previous ones, and quite useful for some applications, in that it measures how faithful the relative spectral distribution of a pixel is reconstructed, while ignoring absolute brightness. Finally, we report the universal image quality index (UIQ) (Wang and Bovik, 2002). This metric evaluates the reconstructed image in terms of luminance, contrast, and structure. UIQ is unitless and its maximum value is 1.

### 5.6.3 Evaluation at lower scale

Quantitative evaluation on Sentinel-2 images is only possible at the lower scale at which the models are trained. I.e., \( T_2 \times \) is evaluated on the task to super-resolve \( 40 \rightarrow 20 \) m, where the 40 m low-resolution and 20 m high-resolution bands are generated by synthetically degrading the original data — for details see Sec. [5.5.1]. In the same way, \( S_6 \times \) is evaluated on the super-resolution task from \( 360 \rightarrow 60 \) m. Furthermore, to support the claim that the upsampling function is to a sufficient degree scale-invariant, we also run a test where we train \( T_2 \times \) on the upsampling task from \( 80 \rightarrow 40 \) m,
Table 5.3: Aggregate results for 2× upampling of the bands in set $B$, evaluated at lower scale (input 40 m, output 20 m). Best results in bold.

<table>
<thead>
<tr>
<th>Training</th>
<th>RMSE</th>
<th>SRE</th>
<th>SAM</th>
<th>UIQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicubic</td>
<td>123.5</td>
<td>25.3</td>
<td>1.24</td>
<td>0.821</td>
</tr>
<tr>
<td>ATPRK</td>
<td>116.2</td>
<td>25.7</td>
<td>1.68</td>
<td>0.855</td>
</tr>
<tr>
<td>SupReME</td>
<td>69.7</td>
<td>29.7</td>
<td>1.26</td>
<td>0.887</td>
</tr>
<tr>
<td>Superres</td>
<td>66.2</td>
<td>30.4</td>
<td>1.02</td>
<td>0.915</td>
</tr>
<tr>
<td>DSen2 (ours)</td>
<td>40 → 20</td>
<td>34.5</td>
<td>36.0</td>
<td>0.78</td>
</tr>
<tr>
<td>VDSen2 (ours)</td>
<td>40 → 20</td>
<td>33.7</td>
<td>36.3</td>
<td>0.76</td>
</tr>
<tr>
<td>DSen2 (ours)</td>
<td>80 → 40</td>
<td>51.7</td>
<td>32.6</td>
<td>0.89</td>
</tr>
<tr>
<td>VDSen2 (ours)</td>
<td>80 → 40</td>
<td>51.6</td>
<td>32.7</td>
<td>0.88</td>
</tr>
</tbody>
</table>

and then test that network to the 40 → 20 m upsampling task. In the following, we separately discuss the $T_{2\times}$ and $S_{6\times}$ networks.

$T_{2\times}$ — 20 m bands We start with results for the $T_{2\times}$ network, trained for super-resolution of actual S2 data to 10 m. Average results over all 15 test images and all bands in $B = \{B5, B6, B7, B8a, B11, B12\}$ are displayed in Table 5.3. The state-of-the-art methods SupReME and Superres perform similar, with Superres slightly better in all error metrics. DSen2 reduces the RMSE by 48% compared to the previous state-of-the-art. The other error measures confirm this gulf in performance (>5 dB higher SRE, 24% lower SAM). VDSen2 further improves the results, consistently over all error measures (except UIQ, where their scores are exactly the same). Relative to the leap from the best baseline to DSen2 the differences may seem small, but note that 0.3 dB would still be considered a marked improvement in many image enhancement tasks. Interestingly, ATPRK and SupReME yield rather poor results for SAM (relative spectral fidelity). Among the baselines, only SupReME beats bicubic upsampling. Our method again wins comfortably, more than doubling the margin between the strongest competitor SupReME and the simplistic baseline of bicubic upsampling.

In the second test, we train an auxiliary $T_{2\times}$ network on 80 → 40 m instead of the 40 → 20 m, but nevertheless evaluate it on the 20 m ground truth (while the model has never seen a 20 m GSD image). Of course this causes some drop in performance, but the performance stays well above all baselines, across all bands. i.e., the learned mapping is indeed sufficiently scale-invariant to beat state-of-the-art model-based approaches, which by construction should not depend on the absolute scale. For our actual setting, train on 40 → 20 m then use for 20 → 10 m, one would expect even a smaller performance drop (compared to train on 80 → 40 m then use for 40 → 20 m), because of the well-documented inverse relation between spatial frequency and contrast in image signals (e.g., [Ruderman, 1994], [van der Schaaf and van Hateren, 1996], [Srivastava et al., 2003]). This experiment justifies our assumption, at 2× reduced res-olution, that training 40 → 20 m super-resolution on synthetically degraded images is a reasonable proxy for the actual 20 → 10 m upsampling of real Sentinel-2 images. We note that this result has potential implications beyond our specific CNN approach. It
5.6. EXPERIMENTAL RESULTS

Table 5.4: Per-band values of RMSE, SRE and UIQ, for 2× upsampling. Values are averages over all test images. Evaluation at lower scale (input 40 m, output 20 m). Best results in bold.

<table>
<thead>
<tr>
<th></th>
<th>B5</th>
<th>B6</th>
<th>B7</th>
<th>B8a</th>
<th>B11</th>
<th>B12</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RMSE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bicubic</td>
<td>105.0</td>
<td>138.1</td>
<td>159.3</td>
<td>168.3</td>
<td>92.4</td>
<td>78.0</td>
</tr>
<tr>
<td>ATPRK</td>
<td>89.4</td>
<td>119.1</td>
<td>136.5</td>
<td>147.4</td>
<td>113.3</td>
<td>91.7</td>
</tr>
<tr>
<td>SupReME</td>
<td>48.1</td>
<td>70.2</td>
<td>78.6</td>
<td>82.9</td>
<td>76.5</td>
<td>61.7</td>
</tr>
<tr>
<td>Superres</td>
<td>50.2</td>
<td>66.6</td>
<td>76.8</td>
<td>82.0</td>
<td>66.9</td>
<td>54.5</td>
</tr>
<tr>
<td>DSen2</td>
<td>27.7</td>
<td>37.6</td>
<td>42.8</td>
<td>43.8</td>
<td>29.0</td>
<td>26.2</td>
</tr>
<tr>
<td>VDSen2</td>
<td>27.1</td>
<td>37.0</td>
<td><strong>42.2</strong></td>
<td><strong>43.0</strong></td>
<td><strong>28.0</strong></td>
<td><strong>25.1</strong></td>
</tr>
</tbody>
</table>

| **SRE** |        |        |        |        |        |        |
| Bicubic | 25.1   | 25.6   | 25.4   | 25.5   | 26.3   | 24.0   |
| ATPRK   | 26.6   | 26.9   | 26.7   | 26.6   | 24.7   | 22.7   |
| SupReME | 31.2   | 31.0   | 31.0   | 31.2   | 27.9   | 26.1   |
| Superres | 31.3   | 31.7   | 31.4   | 31.4   | 29.1   | 27.2   |
| DSen2   | 36.2   | 36.5   | 36.5   | 36.9   | 36.3   | 33.6   |
| VDSen2  | **36.5** | **36.8** | **36.7** | **37.1** | **36.7** | **34.0** |

| **UIQ** |        |        |        |        |        |        |
| Bicubic | 0.811  | 0.801  | 0.802  | 0.806  | 0.857  | 0.847  |
| ATPRK   | 0.889  | 0.881  | 0.891  | 0.883  | 0.789  | 0.795  |
| SupReME | 0.889  | 0.890  | 0.894  | 0.894  | 0.878  | 0.879  |
| Superres | 0.918  | 0.920  | 0.921  | 0.919  | 0.904  | 0.905  |
| DSen2 (ours) | **0.943** | **0.942** | **0.942** | **0.935** | **0.943** | **0.940** |
| VDSen2 (ours) | 0.939 | **0.944** | 0.938 | **0.943** | **0.946** | 0.935 |

validates the general procedure to train on lower-resolution imagery, that has been synthesised from the original sensor data. That procedure is in no way specific to our technical implementation, and in all likelihood also not to the sensor characteristics of Sentinel-2.

Tables 5.4 and Fig. 5.7 show detailed per-band results. The large advantage for our method is consistent across all bands, and in fact particularly pronounced for the chal-
Table 5.5: Full results and detailed RMSE, SRE and UIQ values per spectral band. The results are averaged over all images for the 6× upsampling, with evaluation at lower scale (input 360 m, output 60 m). Best results in bold.

<table>
<thead>
<tr>
<th></th>
<th>B1</th>
<th></th>
<th>B9</th>
<th></th>
<th>Average</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>SRE</td>
<td>UIQ</td>
<td>RMSE</td>
<td>SRE</td>
<td>UIQ</td>
</tr>
<tr>
<td>Bicubic</td>
<td>171.8</td>
<td>22.3</td>
<td>0.404</td>
<td>148.7</td>
<td>17.1</td>
<td>0.368</td>
</tr>
<tr>
<td>ATPRK</td>
<td>162.9</td>
<td>22.8</td>
<td>0.745</td>
<td>127.4</td>
<td>18.0</td>
<td>0.711</td>
</tr>
<tr>
<td>SupReME</td>
<td>114.9</td>
<td>25.2</td>
<td>0.667</td>
<td>56.4</td>
<td>24.5</td>
<td>0.819</td>
</tr>
<tr>
<td>Superres</td>
<td>107.5</td>
<td>24.8</td>
<td>0.566</td>
<td>92.9</td>
<td>20.8</td>
<td>0.657</td>
</tr>
<tr>
<td>DSen2</td>
<td>33.6</td>
<td>35.6</td>
<td>0.912</td>
<td>30.9</td>
<td>29.9</td>
<td>0.886</td>
</tr>
<tr>
<td>VDSen2</td>
<td>27.6</td>
<td>37.9</td>
<td>0.921</td>
<td>24.4</td>
<td>32.3</td>
<td>0.899</td>
</tr>
</tbody>
</table>

Lenging extrapolation to B11 and B12. We point out that the RMSE values for B6, B7 and B8a are higher than for the other bands (with all methods). In these bands also the reflectance is higher. The relative errors, as measured by SRE, are very similar. Among our two networks, VDSen2 holds a moderate, but consistent benefit over its shallower counterpart across all bands, in both RMSE and SRE. In terms of UIQ, they both rank well above the competition, but there is no clear winner. We attribute this to limitations of the UIQ metric, which is a product of three terms and thus not overly stable near its maximum of 1.

It is interesting to note that the baselines exhibit a marked drop in accuracy for bands B11 and B12, whereas our networks reconstruct B11 as well as other bands and show only a slight drop in relative accuracy for B12. These two bands lie in the SWIR (>1.6 µm) spectrum, far outside the spectral range covered by the high-resolution bands (0.4–0.9 µm). Especially ATPRK performs poorly on B11 and B12. The issue is further discussed in Sec. 5.6.5.

In Fig. 5.8 we compare reconstructed images to ground truth for one of the test images. Yellow denotes high residual errors, dark blue means zero error. For bands B6, B7, B8a and B11 all baselines exhibit errors along high-contrast edges (the residual images resemble a high-pass filtering), meaning that they either blur edges or exaggerate the contrast. Our method shows only traces of this common behaviour, and has visibly lower residuals in all spectral bands.

**S_{6×} — 60 m bands** For 6× super-resolution we train a separate network, using synthetically downgraded images with 60 m GSD as ground truth. The baselines are run with the same settings as before (i.e., jointly super-resolving all input bands), but only the 60 m bands \( C = \{B1, B9\} \) are displayed. Overall and per-band results are given in Table 5.5. Once again, our DSen2 network outperforms the previous state-of-the-art by a large margin, reducing the RMSE by a factor of \( \approx 3 \). For the larger upsampling factor, the very deep VDSen2 beats the shallower DSen2 by a solid margin, reaching about 20% lower RMSE, respectively 2.3 dB higher SRE.

Among the baselines, SupReME this time exhibits better overall numbers than Su-
Figure 5.8: Absolute differences between ground truth and $2 \times$ upsampled result at 20 m GSD. The images show (absolute) reflectance differences on a reflectance scale from 0 to 10,000. Top, left to right: RGB (B2, B3, B4) image, colour composites of bands (B5, B6, B7), and of bands (B8a, B11, B12). The image depicts the Siberian tundra near the mouth of the Pur River.
Figure 5.9: Absolute differences between ground truth and $6 \times$ upsampled result at 60 m GSD. The images show (absolute) reflectance differences on a reflectance scale from 0 to 10,000. Top: True scene RGB (B2, B3, B4), and false colour composite of B1 and B9. This image depicts Berg River Dam in the rocky Hotentots Holland, east of Cape Town, South Africa.

perres, thanks to it clearly superior performance on the B9 band. Contrary to the $2 \times$ super-resolution, all baselines improve SAM compared to simple bicubic interpolation. Our method again is the runaway winner, with VDSen2 reaching 65% lower error than the nearest competitor SupReME. Looking at the individual bands, all methods perform better (relative to average radiance) on B1 than on B9. The latter is the most challenging band for super-resolution, and the only one for which our SRE drops below 33 dB, and our UIQ below 0.9. It is worth noticing, that in this more challenging $6 \times$ super-resolution, our method brings a bigger improvement compared to the state-of-the-art baselines in $2 \times$ super-resolution.

We also present a qualitative comparison to ground truth, again plotting absolute residuals in Fig. 5.9. As for 20 m, the visual impression confirms that DSen2 and VDSen2 clearly dominate the competition, with much lower and less structured residuals.

### 5.6.4 Evaluation at the original scale

To verify that our method can be applied to true scale Sentinel-2 data, we super-resolve the same test images as before, but feed the original images, without synthetic down-sampling, to our networks. As said before, we see no way to obtain ground truth data for a quantitative comparison, and therefore have to rely on visual inspection. We plot
5.6. EXPERIMENTAL RESULTS

Figure 5.10: Results of DSen2 on real Sentinel-2 data, for $2 \times$ upsampling. From left to right: True scene RGB in 10 m GSD (B2, B3, B4), Initial 20 m bands, Super-resolved bands (B12, B8a and B5 as RGB) to 10 m GSD with DSen2. Top: An agricultural area close to Malmö in Sweden. Middle: A coastal area at the Shark Bay, Australia. Bottom: Central Park at Manhattan, New York, USA. Best viewed on computer screen.
Figure 5.11: Results of DSen2 on real Sentinel-2 data, for $6 \times$ upsampling. From left to right: True scene RGB (B2, B3, B4), Initial 60 m bands, Super-resolved bands (B9, B9 and B1 as RGB) with DSen2. Top: London Heathrow airport and surroundings. Middle: The foot of Mt. Aso, on Kyushu island, Japan. Bottom: A glacier in Greenland. Best viewed on computer screen.
5.6. EXPERIMENTAL RESULTS

the upsampled results next to the low-resolution inputs, in Fig. [5.10] for $2 \times$ upsampling and in Fig. [5.11] for $6 \times$ upsampling. For each upsampling rate, the figures show 3 different test locations with varying land cover. Since visualisation is limited to 3 bands at a time, we pick bands (B5, B8a, B12) for $2 \times$ upsampling. For $6 \times$ upsampling we show both bands (B1, B9). In all cases the super-resolved image is clearly sharper and brings out additional details compared to the respective input bands. At least visually, the perceptual quality of the super-resolved images matches that of the RGB bands, which have native 10 m resolution.

5.6.5 Suitability of pan-sharpening methods

As discussed earlier, there is a conceptual difference between multi-spectral super-resolution and classical pan-sharpening, in that the latter simply “copies” high-frequency information from an overlapping or nearby high-resolution band, but cannot exploit the overall reflectance distribution across the spectrum. Still, it is a-priori not clear how much of a practical impact this has, therefore we also test three of the best-performing pan-sharpening methods in the literature, namely PRACS [Choi et al., 2011], MTF-GLP-HPM-PP [Lee and Lee, 2010] and BDSD [Garzelli et al., 2008]. Quantitative error measures for the $2 \times$ case are given in Table 5.6. Pan-sharpening requires a single “panchromatic” band as high-resolution input. The combinations that empirically worked best for our data were the following: For the near-infrared bands B6, B7 and B8a, we use the broad high-resolution NIR band B8. As panchromatic band for B5 we use B2, which surprisingly worked better than the spectrally closer B8, and also slightly better than other visual bands. While for the SWIR bands there is no spectrally close high-resolution band, and the best compromise appears to be the average of the three visible bands, $\frac{1}{3}(B2+B3+B4)$.

For bands B5, B6, B7 and B8 the results are reasonable: the errors are higher than those of the best super-resolution baseline (and consequently $2-3 \times$ higher than with our networks, c.f. Table 5.4), but lower than naive bicubic upsampling. This confirms that there is a benefit from using all bands together, rather than the high-frequency data from only one, arbitrarily defined “panchromatic” band.

On the contrary, for the SWIR bands B11 and B12 the performance of pan-sharpening drops significantly, to a point that the RMSE drops below that of bicubic interpolation (and similar for SRE). As was to be expected, successful pan-sharpening is not possible with a spectrally distant band that has very different image statistics and local appearance. Moreover, pan-sharpening is very sensitive to the choice of the “panchromatic” band. We empirically picked the one that worked best on average, but found that, for all tested methods, there isn’t one that performs consistently across all test images. This is particularly evident for MTF-GLP-HPM-PP. Even with the best pan-band we found (the average of the visible bands), it reconstructed reasonable SWIR bands for some images, but completely failed on others, leading to excessive residuals.

While it may be possible to improve pan-sharpening performance with some sophisticated, perhaps non-linear combination for the pan-band, determining that combination is a research problem on its own, and beyond the scope of this paper.
Table 5.6: Results of well-known pan-sharpening methods. RMSE, SRE and UIQ values per spectral band averaged over all images for the $2 \times$ upsampling, with evaluation at lower scale (input 40 m, output 20 m). Best results in bold.

<table>
<thead>
<tr>
<th></th>
<th>B5</th>
<th>B6</th>
<th>B7</th>
<th>B8a</th>
<th>B11</th>
<th>B12</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RMSE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bicubic</td>
<td>105.0</td>
<td>138.1</td>
<td>159.3</td>
<td>168.3</td>
<td>92.4</td>
<td><strong>78.0</strong></td>
<td>123.5</td>
</tr>
<tr>
<td>PRACS</td>
<td>99.3</td>
<td>148.1</td>
<td>99.2</td>
<td>104.2</td>
<td>290.0</td>
<td>320.0</td>
<td>176.8</td>
</tr>
<tr>
<td>MTF-GLP-HPM-PP</td>
<td>91.0</td>
<td><strong>66.5</strong></td>
<td>77.6</td>
<td>82.7</td>
<td><strong>78.7</strong></td>
<td>240.6</td>
<td>106.2</td>
</tr>
<tr>
<td>BDSD</td>
<td><strong>64.7</strong></td>
<td>84.2</td>
<td><strong>76.0</strong></td>
<td><strong>78.8</strong></td>
<td>93.4</td>
<td>79.4</td>
<td><strong>79.4</strong></td>
</tr>
<tr>
<td><strong>SRE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bicubic</td>
<td>25.1</td>
<td>25.6</td>
<td>25.4</td>
<td>25.5</td>
<td><strong>26.3</strong></td>
<td>24.0</td>
<td>25.3</td>
</tr>
<tr>
<td>PRACS</td>
<td>24.0</td>
<td>24.2</td>
<td>28.7</td>
<td>29.0</td>
<td>19.5</td>
<td>14.4</td>
<td>23.3</td>
</tr>
<tr>
<td>MTF-GLP-HPM-PP</td>
<td>28.0</td>
<td><strong>30.7</strong></td>
<td>30.5</td>
<td>30.7</td>
<td>28.0</td>
<td>23.0</td>
<td><strong>28.5</strong></td>
</tr>
<tr>
<td>BDSD</td>
<td><strong>28.3</strong></td>
<td>29.2</td>
<td><strong>31.1</strong></td>
<td><strong>31.5</strong></td>
<td><strong>26.3</strong></td>
<td>23.9</td>
<td>28.4</td>
</tr>
<tr>
<td><strong>UIQ</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bicubic</td>
<td>0.811</td>
<td>0.801</td>
<td>0.802</td>
<td>0.806</td>
<td>0.857</td>
<td>0.847</td>
<td>0.821</td>
</tr>
<tr>
<td>PRACS</td>
<td>0.836</td>
<td>0.858</td>
<td>0.882</td>
<td>0.881</td>
<td>0.791</td>
<td>0.773</td>
<td>0.837</td>
</tr>
<tr>
<td>MTF-GLP-HPM-PP</td>
<td><strong>0.893</strong></td>
<td><strong>0.898</strong></td>
<td><strong>0.909</strong></td>
<td><strong>0.909</strong></td>
<td><strong>0.877</strong></td>
<td><strong>0.881</strong></td>
<td><strong>0.895</strong></td>
</tr>
<tr>
<td>BDSD</td>
<td>0.866</td>
<td>0.892</td>
<td><strong>0.909</strong></td>
<td>0.908</td>
<td>0.858</td>
<td>0.848</td>
<td>0.880</td>
</tr>
</tbody>
</table>

For readability, the pan-sharpening results are displayed in a separate table. We note for completeness that, among the super-resolution baselines (Tables 5.3 and 5.4), ATPRK is technically also a pan-sharpening method, but includes a mechanism to automatically select one of several high-resolution channels as its the “panchromatic” input. We categorise it as super-resolution, since its creators also intend and apply it for that purpose. It can be seen in Table 5.4 that ATPRK actually also exhibits a distinct performance drop for bands B11 and B12.

Overall, we conclude that pan-sharpening cannot substitute qualified super-resolution, and is not suitable for Sentinel-2. Nevertheless, we point out that in the literature, the difficulties it has especially with bands B11 and B12 is sometimes masked, because many papers do not show the individual per-band errors.

5.7 Discussion

5.7.1 Different network configurations

The behaviour of our two tested network configurations is in line with the recent literature: networks of moderate size (by today’s standards), like DSen2, already perform fairly well. Across a wide range of image analysis tasks from denoising to instance-level semantic segmentation and beyond, CNNs with around 10–20 layers have redefined the state-of-the-art. Over the last few years, improvements to the network architecture
have enabled training of very deep networks with even more (in some cases >100) layers, like VDSen2. Empirically, these models tend to raise the bar even further, but the gains are less dramatic, as adding more and more layers faces diminishing returns. Somewhat surprisingly, even the very deep models with tens of millions of free parameters do not show a strong tendency to overfit, if designed correctly. We note that our networks differ from the prevalent design for high-level analysis (semantic segmentation, depth estimation, etc.). These normally have an “hourglass” structure with an encoder part that successively increases the receptive field (respectively, reduces the spatial resolution) via pooling operations, followed by a decoder part that restores the original resolution via transposed convolutions. We refrain from pooling, since it carries the danger of degrading local detail, while conversely a fairly small neighbourhood is, in our view, sufficient to determine the local spectral relations.

What is the “right” depth for super-resolution? As usual in such cases, there is no single answer, since this depends on the specific application (e.g., variability of the land-cover, available computational resources, update frequency, etc.). As a general guideline, we find that, with adequate hardware at hand, there is no disadvantage in using VDSen2. It is neither more difficult to use nor more brittle to train from the perspective of the user. While it does consistently produce super-resolved images with lower residuals, especially for the challenging 6× upsampling. If hardware resources (especially GPU memory) are limited, or very large interest regions must be processed in a short time, it may nevertheless be better to work with DSen2. The results are still very good, and in certain cases, e.g., if only 2× upsampling is needed and/or the spectral variability in the interest region is not too high will probably match the performance of a deeper architecture. Importantly, intermediate variants are also possible: if one aims for the highest possible quality under limited resources, it may make sense to chose a number of ResBlocks between the \( d = 6 \) of DSen2 and the \( d = 32 \) of VDSen2. In fact, for “easy” land-cover or if maximal accuracy is not needed (e.g., for visualisation) it may well be possible to remove another 1 or 2 ResBlocks from DSen2 and still obtain satisfactory results.

### 5.7.2 Timing

As in general for deep learning, training a network is computationally demanding and takes time (often several days, see sec. 5.5.5), but the single forward pass to super-resolve a new image is very fast. We note that long training times are usually required only once, when training from scratch. Refining/adapting an existing network with further training data is a lot less costly. Our pretrained networks can serve as a starting point.

In Table 5.7 the runtimes of all tested methods are presented for super-resolving all 20 m bands of a complete Sentinel-2 tile (10,980×10,980 pixels). The baselines are only available as CPU code, and in some cases not easy to parallelise, whereas CNNs are almost always run on GPUs – in fact, their current revival was, to a large part, triggered by the advent of parallel computing on GPUs. We therefore show both processing times. The comparison is indicative and not meant to claim our method is a
Table 5.7: Runtimes for super-resolving the six 20 m bands of a standard Sentinel-2 tile (10980 × 10980 pixels, ≈120 Mpix).

<table>
<thead>
<tr>
<th>Method</th>
<th>CPU time</th>
<th>GPU time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicubic</td>
<td>≪1 min</td>
<td>–</td>
</tr>
<tr>
<td>ATPRK</td>
<td>149 min</td>
<td>–</td>
</tr>
<tr>
<td>SupReME</td>
<td>123 min</td>
<td>–</td>
</tr>
<tr>
<td>Superres</td>
<td>315 min</td>
<td>–</td>
</tr>
<tr>
<td>DSen2 (ours)</td>
<td>130 min</td>
<td>3 min</td>
</tr>
<tr>
<td>VDSen2 (ours)</td>
<td>≈30 h 14 min</td>
<td></td>
</tr>
</tbody>
</table>

lot faster than the baselines: modern CNN frameworks are highly optimised, whereas the baselines are research implementations with much potential for speed-ups. Still, the numbers are useful to show that CNN-based super-resolution is fairly efficient, and clearly fast enough to be used in practice without much further code optimisation. For the Comparison, we used an Intel(R) Xeon(R) CPU E5-1650 v3 @ 3.50GHz, respectively an NVIDIA Titan Xp GPU. On a desktop computer with a single GPU, DSen2 super-resolves a complete Sentinel-2 tile to 10 m in 3 minutes, and VDSen2 in 14 minutes. We note that hardware producers are working on specialised tensor processing hardware that is optimised for deep learning (rather than gaming and computer graphics), and can be expected to further speed up CNNs. We do point out that if no powerful GPU is available, very deep networks are not viable. On the contrary, DSen2 takes ≈2 hours of CPU time and is comparable with the fastest baseline method.

5.7.3 Open-source publication of our models

The publication of this paper includes open, publicly available implementations of our models, at: https://github.com/lanha/DSen2. We provide the python source files (in Keras format) for the network specifications as well as the training procedure. Moreover, we also provide the already trained weights used in all our experiments. These shall enable out-of-the-box super-resolution of Sentinel-2 images world-wide, with minimal knowledge of neural network tools. Of course, if a study is focussed only in a specific geographic location, biome or land-cover type, even better result can be expected by training the network only with images showing those specific conditions. The literature suggests that in that case, it may be best to start from our globally trained network and fine-tune it through further training iterations on task-specific imagery.

In the future, we hope to also integrate our method into the SNAP toolbox for Sentinel-2 processing, so as to use our super-resolution instead of naive upsampling within the processing pipeline. A word of caution: our weights are trained only on real Sentinel-2 images, and their excellent performance is to a large part due to the fact that they are optimised specifically for the image statistics of the input data. They are therefore not suitable for processing data from other sensors, or other processing levels of Sentinel-2.
5.8 Conclusions

We have described a tool to super-resolve (“sharpen”) the lower-resolution (20 m and 60 m) bands of Sentinel-2 to a uniform 10 m GSD data cube. Our method uses two deep convolutional neural networks to jointly learn the mapping from all input bands to the $2 \times$, respectively $6 \times$ super-resolved output bands. To train the network, we make the empirically plausible assumption that the correct way of transferring high-frequency information across spectral bands is invariant over a range of scales. In this way, we can synthesise arbitrary amounts of training data with known ground truth from the Sentinel-2 archive. We sample a large and varied global dataset that, according to our experiments, yields a super-resolution tool that generalises to unseen locations in different parts of the world.

The super-resolution network shows excellent performance, reducing the RMSE of the prediction by 50% compared to the best competing methods; respectively, increasing the SRE by almost 6 dB. Qualitative results from different land-cover types, biomes and climate zones confirm the good performance also on full-resolution S2 images. Moreover, the method is also fast enough for practical large-scale applications, computation times are on the order of a few minutes for a complete, 120 MPix Sentinel-2 tile.

While in our work we have focussed on Sentinel-2, the networks are learned end-to-end from image data and thus completely generic. We are confident that they can be retrained for super-resolution of different multi-resolution multi-spectral sensors. We make our software and models available as open-source tools for the remote sensing community.

5.9 Acknowledgments

The authors acknowledge financial support from the Swiss National Science Foundation (SNSF), project No. 200021_162998, Fundação para a Ciência e a Tecnologia, Portuguese Ministry of Science, Technology and Higher Education, projects UID/EEA/50008/2013 and ERANETMED/0001/2014.
Chapter 6

Conclusions and Outlook

6.1 Summary

The results and findings of this thesis are the outcome of a 4½ year’s study on the spatial and spectral resolution enhancement of remote sensing images. The major milestones of the research are represented by the “core” technical chapters of the dissertation (c.f. Chapters 3, 4 and 5). The high-level target of this dissertation was to study the potential of combining high spectral and high spatial resolution and propose effective new methods and algorithms. The main difficulties of increasing the spatial resolution are the degradations in the spatial domain that occur during image acquisition, namely blurring and downsampling. Furthermore, the images are affected by spectral degradation caused by the spectral mixture of materials. Hence, the observed signals are entangled, leading to challenging inverse problems. To overcome these difficulties the present thesis proposes three different methods to tackle the two problems of spectral and spatial resolution enhancement. These are all related in terms of the objective and final product.

The first problem deals with the fusion of hyperspectral and multispectral images to enhance the spatial resolution of the hyperspectral image. The method developed in this dissertation (Chapter 3) is called SupResPALM and was experimentally tested on simulated hyperspectral remote sensing datasets (University of Pavia and an APEX scene), close-range hyperspectral images (CAVE and Harvard) and real data from the EO-1 satellite (Hyperion and ALI sensors).

The second problem tackled in the dissertation is the super-resolution of the lower resolution bands within an image that has multiple resolutions. Typical examples of this configuration are sensors onboard satellites, such as ASTER, Worldview-3 and Sentinel-2. To solve this, Chapter 4 presented a variational approach, minimising an objective function to fit the data to a physical model with adequate regularisation. This method is called SupReME and was tested on simulated data created from real hyperspectral images, so that ground truth exists. The method was also tested on real Sentinel-2 data. The thesis proposes a second solution to the multiresolution sensor setup (Chapter 5). Contrary to an explicit physical model of the image acquisition pro-
cess at multiple resolutions, the upsampling of the lower-resolution bands is learned from data. A CNN is designed to infer all the bands at maximum resolution. As this method, named DSen2 (Deep Sentinel-2), encompasses all steps from inputs to outputs, the trained network only works for the sensor it was trained for. Nonetheless, the method is not limited to any particular sensor and in principle can be applied to any other sensor as long as sufficient training data can be obtained. For this project, Sentinel-2 data were used, both in true scale and also simulated (downsampled) to enable the comparison with ground truth. Data from 60 different Sentinel-2 scenes were selected from around the globe located in multiple climate zones, encompassing different land-cover types to train (45 scenes) and validate (15 scenes) the method.

6.2 Discussion of Contributions

As previously stated, the objective of this thesis was to combine high spatial and spectral resolution in the context of remote sensing. Throughout the study, we have pointed out how these advances benefit remote sensing. In the following the most important points are highlighted.

6.2.1 Hyperspectral Super-Resolution

In order to fuse hyperspectral and multispectral images, we use a joint unmixing process for the two images (Chapter 3). The constraints of spectral unmixing helped to stabilise the solution and also to impose sparsity on the representation coefficients (abundances). Further, increasing the sparsity in accordance with the linear mixing model is possible, but this does not improve the super-resolution results. Moreover, the spatial regularisation introduced to preserve the discontinuities seemed only to have a minor effect, implying that the simplex constraint (both ANC and ASC) is powerful enough to guide the fusion.

We developed methodology for estimating the relative responses of the sensors from the data themselves. Even though estimates from the manufacturers sometimes exist, we don’t take these as granted. The problem of perfectly aligning the two images is quite challenging, because of the different spectral and spatial resolutions. Within Chapter 3 a small translational mis-registration is corrected through the estimation of the spatial blur of the hyperspectral image. This allows the model to cope with any remaining errors that still remain after preprocessing the images. Finally, the application of the method to real images validates the proposed approach beyond that of simulations alone.

6.2.2 Super-Resolution of Multispectral Multiresolution Images

Chapter 4 presented a new method to super-resolve all bands of a multiresolution sensor to the highest resolution, with an explicit forward observation model; a problem
especially relevant for spaceborne sensors. This approach is similar to the previous one (see Subsec. 6.2.1), with the difference, that the low and high spatial resolution components have only a few spectral bands per resolution, compared to hyperspectral images that usually have more than 100. Additionally, there is often no spectral overlap between the high and low resolution bands.

As all the bands observe the same scene, it is assumed that data “lives” in a subspace of the underlying full space, and this significantly reduces the complexity. The forward model is then inverted with a series of operations that are equivalent to convolutions, while the formulation and solution is generic and not restricted to any specific sensor. Thus, it can easily be used for any multiresolution sensor, out of the box. An interesting example would be its extension to WorldView-3 images, given the fact that this satellite’s data is very expensive and learning based techniques (such as the one in Chapter 5 cannot make use of extensive training data.

### 6.2.3 Super-Resolution of Sentinel-2 Images

The multiresolution configuration of Sentinel-2 proved to be well suited for training a globally applicable deep neural network. The focus of Chapter 5 is to tailor the super-resolution to Sentinel-2, contrary to the generic case mentioned before (Chapter 4). Therein, a series of hand-crafted convolutions were applied iteratively, until the problem converged. However, with a neural network many more convolutions are learned directly from the input data. Increasing the number of convolution layers of a moderately deep network (with up to $20 \times$ more parameters) does not seem to significantly improve the performance. On the other hand, even with a network of $\approx 40$ million parameters no signs of overfitting appear.

Striving to improve the spatial resolution of all of the bands of S2, the method produces state-of-the-art results, even though training has to be carried out at a different scale than the scale of the inferred images. Importantly, we show that a globally valid set of network parameters can be learned. Still, users can fine-tune these weights to their specific region or application, if desired.

### 6.3 Discussion of Limitations

This dissertation, as demonstrated in the previous section, has contributed to the knowledge on spectral and spatial super-resolution. There is however, some room for further improvements as there are a number of factors that present some weaknesses and shortcomings in the approaches used. Here the most important aspects of the limitations are discussed.

**Ground truth.** One of the most important limitations in this dissertation is the lack of ground truth for experiments on real datasets. To facilitate quantitative evaluation, we revert to simulated data as mentioned in all the core chapters. The main cause of this
problem is, that we would need a sensor with a higher spatial and spectral resolution than that which is already available as input – but the lack of such a sensor is precisely the motivation of the work. In most of the cases this is not even technically possible and even if it is possible other limitations exist with comparing the higher spatial and spectral resolution image, e.g., there might be temporal differences. Furthermore, the manual creation of ground truth, e.g., with spectro-radiometers on the ground, is unrealistic for the large areas that would have to be captured. We emphasise that this strengthens the case for computational super-resolution, since it confirms that there is no other realistic way to obtain the desired high spectral and spatial resolution imagery.

**Multi-temporal sensing.** Only few methods of hyperspectral super-resolution have actually been applied to real data. In most cases the observation model assumes no multi-temporal changes in the scene, i.e., the images were captured under the same conditions. In practice, many configurations do not capture hyperspectral and multispectral images at the same time. Taking images at different times will alter the visual appearance of the scene. Especially changes in illumination can occur very fast and also objects might move, e.g., cars. Other greater temporal sensing differences are observed for vegetation (seasonal variation, deforestation, fire damage) and even changes to man-made structures (new structures, demolitions). All these temporal changes cannot easily be combined in any one model and this is a challenging problem for future research. Ferraris et al. (2017) have examined ways of identifying changes by using images of different spatial and spectral resolutions.

**Co-registration.** The methods presented rely on a certain level of preprocessing of the input images. Possibly the most important prerequisite is co-registration. In simulated data, mis-registration is not present, while with real data, the proposed approaches have proven very brittle in the presence of small co-registration errors. Thus the input images are either corrected in advance, or the co-registration is estimated in conjunction with the spatial blur. This problem arises mainly for images acquired from two different platforms, as they are not captured under the same conditions. Various geometric distortions between the different images can be present, e.g., translation, rotation, affine transformations, or displacements due to DEM (Digital Elevation Model) errors. Moreover, depending on how the geo-coding of the images has been performed and the image perspective, non-linear (geometric) effects might also be present. When the data are captured from the same platform (as is always the case for the multiresolution sensors), the data are usually perfectly co-registered, or at most a translational correction is required. All the data used in the thesis have been processed with the techniques mentioned in Section 2.2.

**Physical model.** The physical model that describes the forward problem of degrading the images relies on some assumptions that are not always true. A question arises as to whether the forward model really captures accurately the behaviour of the real world. This also questions how realistic simulated images are compared to real images. The spatial degradation can potentially vary from band to band, while at the
same time band-to-band mis-registration can occur. These effects are present in the free APEX dataset, used in Chapter 3. In this case the widely used model is not valid and it is unknown what influence the deviations might have. In addition, the very convenient subspace used for hyperspectral images (Chapter 3) and for multispectral images (Chapter 4) limits the super-resolution result into having the intrinsic dimensionality of the subspace and not the full dimensions. I.e., the span of the spectral vectors is equal to the subspace dimension, which could, for instance, suppress subtle patterns that are important for pixel classification. Using a subspace can be questionable in the case of multispectral images, where there are fewer bands and the validity of the subspace projection has not been studied extensively in the literature. Finally, a small number of pixels with very different spectral signatures are likely to be excluded from the subspace representation.

**Processing speed.** An issue which arises with the implementations presented in Chapters 3 and 4 is the long time that the inference takes to converge. While this might not be a problem for small images and specific projects, when moving towards a global scale, computation time is a limiting factor. As mentioned in Chapter 5, traditional techniques are up to an order of magnitude slower than optimised, highly parallel code that runs on high-end GPUs. Furthermore, the hyperspectral super-resolution (c.f. Chapter 3) may also need to run in real-time on limited resources, e.g., for applications on smart mobile devices.

6.4 Outlook

The developments presented in this dissertation should help to raise awareness of the progress in the fields of computer vision and machine learning. Most of the solutions in this thesis are inspired from adaptations of methods that are state-of-the-art in these fields. In the following, a few promising directions for future research are discussed, based on the methodological and practical findings of the current dissertation. These include open problems, proposals for future research and potential solutions to the limitations encountered w.r.t. new technologies, methodology and data.

6.4.1 Future Directions

**Big data.** The advances in sensor and satellite technology have increased the amount and resolution of data, and these are growing very fast. The Landsat satellite series has been observing the earth for more than 45 years, since 1972, with a global coverage occurring every 16 days with Landsat-8. Together with the exciting Sentinel-2 mission, the global coverage now occurs in less than 5 days. Moreover, Planet Labs\(^1\) with a big constellation of small satellites (called “Doves”) provide global coverage of the earth almost everyday at a 3-5 m GSD. “SkySat” is a constellation that

\(^{1}\)https://www.planet.com/
provides sub-meter resolution images of the earth with sub-weekly temporal coverage. This data is available for research in an open data access policy. These constitute only a few examples of the currently ongoing explosion in remote sensing data. With the huge amounts of global data, the bottleneck is to convert the raw data into useful geoinformation. Due to the sheer amount of data, interpreting it to extract geometric and thematic information can only be achieved with almost fully automatic processing. Automating the image interpretation process is presently the main methodological challenge of image-based earth observation, see, e.g., [Lillesand et al.] (2014). The goal is to be able to collaboratively monitor and connect processes that seemingly do not depend on each other, but have a linked behaviour.

The advance of modern machine learning has opened up new ways of handling and exploiting huge amounts of data. Especially for applications that deal with multi-temporal data, RNNs have the ability to explicitly handle time series and dynamic events. This is particularly beneficial for filling gaps in existing data, due to, e.g., clouds (Sano et al., 2007) or limited temporal coverage. Moreover, it enables the prediction of related future events, e.g., crop yields and climate monitoring. The successful estimation of such parameters can have significant, positive impacts for society and the economy.

**New sensor era.** As discussed in Section 2.1.1, there are a number of new missions planned for hyperspectral image acquisition in the near future. These will pose new challenges for hyperspectral super-resolution, depending also on the specific configuration of sensor complements. Most of the future missions will feature a panchromatic band. However, it may be useful to include the possibility of super-resolution already in the design stage and tailor all its characteristics accordingly. Research towards an optimised sensor setup will enable the exploitation of the structure of the signal by combining all the bands, to guarantee a superior result. This procedure is already a norm for creating pan-sharpened images, for which the sensor observes a higher resolution panchromatic band.

**Cloud computing services.** The ability of single users to process gigantic datasets is limited. Recently, big international technology companies provide users and scientists tools that support the scientific analysis of petabyte-scale geospatial data. Google Earth Engine\(^2\) is an online platform that stores and organises satellite imagery and also provides convenient tools to browse and search datasets. With these tools global-scale data mining is promoted and algorithms can be globally applied to the vast amount of available data, without any need to download or store data at the user’s side. All the computations are done in the cloud and the results are almost instantly presented to the user. Moreover, the data collection contains all the Sentinel-2 and Landsat collections, many MODIS products, precipitation data and elevation data among others. The Google Earth Engine will soon also be able to run modern deep learning models after the TensorFlow environment has been connected with the platform. TensorFlow is a computer library that provides efficient tools for deep learning. Except for Google,

\(^2\)[https://earthengine.google.com/](https://earthengine.google.com/)
also Amazon provides a cloud computing service, *Amazon Web Services* (AWS). This platform similarly contains open geospatial data including the Landsat 8 and Sentinel-2 archives. Various plugins and services use this platform. In general, beyond these platforms built for geospatial data analysis and visualisation, cloud services enable large scale computations, including deep learning. Cloud computing is gaining ground and can be expected to eventually lead to novel applications yet to be discovered.

**Efficient hyperspectral processing.** Hyperspectral cameras can capture images in real-time rates *(e.g., Imec[] and Cubert[])* and perform also real-time sequential inference, leading to highly redundant, 4-dimensional cubes. This extra information could offer a dynamic analysis of the full temporal series. The process of multi-temporal unmixing *(Zurita-Milla et al., 2011)* can potentially benefit from the knowledge of multi-frame super-resolution techniques.

The fast, low-latency analysis of hyperspectral images could promote the use of cheap sensors on phones and other small devices. The super-resolution of a low resolution hyperspectral image with the normal high-resolution RGB camera of the phone could be used for real-time analysis, *e.g.*, one could image to take a picture of products in the supermarket to assess their state of maturity. Moreover, such sensor could be also placed in home devices *(e.g., microwave ovens)* that monitor the cooking of the food to advance their cooking strategies.

**Climate change.** Remote sensing plays a very important role for the observation of climate change. Its continuous earth observation for over 40 years and its ability to further monitor and process the global information acquired is crucial for further decision making. In combination with the always increasing resolution and data volume, scientists will be able to analyse microeffects and identify the deeper causes of different geophysical phenomena in more detail. Climatologists will then hopefully gain an improved understanding of smaller systems and provide new insights into the global climate change.

Appendix A

Bibliography


Appendix B

List of Acronyms

ADMM  Alternating Direction Method of Multipliers
ADS80  Airborne Digital Sensor 80
ALI    Advanced Land Imager
ALS    Alternating Least Squares
ANC    Abundance Non-negativity Constraint
APEX   Airborne Prism EXperiment
ASC    Abundance Sum Constraint
ASTER  Advanced Spaceborne Thermal Emission and Reflection Radiometer
ATPRK  Area-To-Point Regression Kriging
ATWT   A Trous Wavelet Transform
AVHRR  Advanced Very High Resolution Radiometer
AVIRIS Airborne Visible/Infrared Imaging Spectrometer
AVMAX  Alternating Volume Maximisation
BCCB   Block-Circulant-Circulant-Block
BN     Batch Normalisation
BRDF   Bidirectional Reflectance Distribution Function
CASI   Compact Airborne Spectrographic Imager
CAVIS  Clouds, Aerosols, Vapors, Ice, and Snow
CHRIS  Compact High Resolution Imaging Spectrometer
CMOS   Complementary Metal-Oxide-Semiconductor
CNN    Convolutional Neural Network
CPU    Central Processing Unit
C-SALSA Constrained-Split Augmented Lagrangian Shrinkage Algorithm
DECA   Dependent Component Analysis
DEM    Digital Elevation Model
DEISIS DLR Earth Sensing Imaging Spectrometer
DLR    Deutsches Zentrum für Luft- und Raumfahrt
DNB    Day/Night Band
DR     Dimensionality Reduction
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSen2</td>
<td>Deep Sentinel-2</td>
</tr>
<tr>
<td>DWT</td>
<td>Discrete Wavelet Transform</td>
</tr>
<tr>
<td>EDSR</td>
<td>Enhanced Deep Super-Resolution</td>
</tr>
<tr>
<td>EM</td>
<td>Electromagnetic</td>
</tr>
<tr>
<td>EnMAP</td>
<td>Environmental Mapping and Analysis Programme</td>
</tr>
<tr>
<td>EO-1</td>
<td>Earth Observing-1</td>
</tr>
<tr>
<td>ERGAS</td>
<td>Erreur Relative Globale Adimensionnelle de Synthèse</td>
</tr>
<tr>
<td>ESA</td>
<td>European Space Agency</td>
</tr>
<tr>
<td>FWHM</td>
<td>Full Width at Half Maximum</td>
</tr>
<tr>
<td>GPU</td>
<td>Graphics Processing Unit</td>
</tr>
<tr>
<td>GS</td>
<td>GramSchmidt</td>
</tr>
<tr>
<td>GSD</td>
<td>Ground Sampling Distance</td>
</tr>
<tr>
<td>HIS</td>
<td>Intensity-Hue-Saturation</td>
</tr>
<tr>
<td>HISUI</td>
<td>Hyperspectral Imager Suite</td>
</tr>
<tr>
<td>HSI</td>
<td>Hyperspectral Image</td>
</tr>
<tr>
<td>HYDICE</td>
<td>Hyperspectral Digital Imagery Collection Experiment</td>
</tr>
<tr>
<td>HypIRI</td>
<td>Hyperspectral Infrared Imager</td>
</tr>
<tr>
<td>ICE</td>
<td>Iterative Constrained Endmember</td>
</tr>
<tr>
<td>ISS</td>
<td>International Space Station</td>
</tr>
<tr>
<td>LAI</td>
<td>Leaf Area Index</td>
</tr>
<tr>
<td>LASSO</td>
<td>Least Absolute Shrinkage and Selection Operator</td>
</tr>
<tr>
<td>LMM</td>
<td>Linear Mixing Model</td>
</tr>
<tr>
<td>LP</td>
<td>Laplacian Pyramid</td>
</tr>
<tr>
<td>MNF</td>
<td>Minimum Noise Fraction</td>
</tr>
<tr>
<td>MNIST</td>
<td>Modified National Institute of Standards and Technology</td>
</tr>
<tr>
<td>MODIS</td>
<td>Moderate Resolution Imaging Spectroradiometer</td>
</tr>
<tr>
<td>MS</td>
<td>Multispectral</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Squared Error</td>
</tr>
<tr>
<td>MSI</td>
<td>Multispectral Image</td>
</tr>
<tr>
<td>MTF</td>
<td>Modulation Transfer Function</td>
</tr>
<tr>
<td>MUSES</td>
<td>Multiple User System for Earth Sensing</td>
</tr>
<tr>
<td>MVC-MNF</td>
<td>Minimum Volume Transform Non-negative Matrix Factorisation</td>
</tr>
<tr>
<td>MVES</td>
<td>Minimum Volume Enclosing Simplex</td>
</tr>
<tr>
<td>NCC</td>
<td>Normalised Cross-Correlation</td>
</tr>
<tr>
<td>NDWI</td>
<td>Normalized Difference Water Index</td>
</tr>
<tr>
<td>NIR</td>
<td>Near-Infrared</td>
</tr>
<tr>
<td>NMF</td>
<td>Non-negative Matrix Factorization</td>
</tr>
<tr>
<td>NNLS</td>
<td>Non-Negative Least Squares</td>
</tr>
<tr>
<td>NOAA-20</td>
<td>National Oceanic and Atmospheric Administration-20</td>
</tr>
<tr>
<td>NTIRE</td>
<td>New Trends in Image Restoration and Enhancement</td>
</tr>
<tr>
<td>OLI</td>
<td>Operational Land Imager</td>
</tr>
<tr>
<td>OMP</td>
<td>Orthogonal Matching Pursuit</td>
</tr>
</tbody>
</table>
PAN  Panchromatic
PCA  Principal Component Analysis
PRISMA  PRRecursore IperSpeettrale della Missione Applicativa
PROBA-1  Project for On-Board Autonomy-1
PSF  Point Spread Function
QNR  Quality with No Reference
ReLU  Rectified Linear Unit
RGB  Red, Green and Blue
RMSE  Root-Mean-Squared Error
RNN  Recurrent Neural Network
ROSIS  Reflective Optics System Imaging Spectrometer
S2  Sentinel-2
SAM  Spectral Angle Mapper
SAR  Synthetic-Aperture Radar
SGD  Stochastic Gradient Descent
SISAL  Simplex Identification via variable Splitting and Augmented Lagrangian
SNR  Signal-to-Noise Ratio
SPICE  Sparsity Promoting Iterative Constrained Endmember
SPOT  Satellite Pour l'Oberservation de la Terre
SRE  Signal-to-Reconstruction Error
SSD  Sum of Squared Differences
SuNSAL  Sparse UNmixing by variable Splitting and Augmented Lagrangian
Suomi NPP  Suomi National Polar-orbiting Partnership
SupReME  SUPer-REsolution for multispectral Multiresolution Estimation
SupResPALM  Super-Resolution with Proximal Alternating Linearised Minimisation
SVD  Singular Value Decomposition
SVMAX  Successive Volume Maximisation
SWIR  Short-Wavelength Infrared
TIR  Thermal Infrared
UAV  Unmanned Aerial Vehicle
UQI  Universal Image Quality
USGS  United States Geological Survey
VCA  Vertex Component Analysis
VDSen2  Very Deep Sentinel-2
VIIRS  Visible Infrared Imaging Radiometer Suite
VIS  Visible
VNIR  Visible and Near-Infrared
WV-3  WorldView-3
Appendix C

List of Publications


Appendix D

Curriculum Vitae

Personal data

Name Charis (or Harry) Alexandros Lanaras
Date of birth 07 November 1990
Nationality Greek, British

Education

Oct 2013 - Jul 2018 ETH Zurich, Switzerland
PhD Student

Oct 2008 - Jul 2013 Aristotle University of Thessaloniki, Greece
Dipl. Engineer, Rural and Surveying Engineering

Professional Experience

Aug 2013 - Apr 2018 ETH Zurich, Switzerland
Research and Teaching Assistant