POLITECNICO DI TORINO

Master's Degree in ICT for Smart Societies

Master's Degree Thesis

Methods for Blind Super-Resolution of satellite images

Supervisors

Prof. Enrico MAGLI

Prof. Diego VALSESIA

Candidate

Matteo IMPIERI

December 2023

Summary

Recently, deep neural networks have demonstrated remarkable efficiency at improving the resolution of low-resolution (LR) images. This is called the image super-resolution (SR) task. Numerous researchers have come up with network architectures to deal with the problem of multi-image super-resolution (MISR) by using supervised learning. This entails the availability of ground-truth high-resolution (HR) pictures for the purpose of training. However, collecting LR and HR images from the same device to avoid introducing any additional mismatch between them could be problematic, especially for satellite images, which involve cameras hundreds of kilometres from the subject.

The goal of this thesis is to provide a method to deal with MISR in a blind or unsupervised setting. This was made by leveraging other well-known networks to make blur kernels from LR images and do super-resolution. Specifically, these kernels were then applied to LR pictures in order to create a new set made by coarser-resolution (CR) images. The LR and CR sets were then used to train a supervised MISR network. The main dataset used throughout the research is the Proba-V challenge dataset, which contains collections of HR-LR satellite images.

The backbone of the researched methods is the MISR network PIUnet, proposed by Valsesia and Magli. At the same time, for the kernel estimation task, this thesis explores two different solutions: MANet, introduced by Liang et al., and the architecture DIP-FKP, also by Liang et al. It is important to note that MANet produces spatially-variant kernels, which are known from the literature to be better at representing the degradation of satellite images, while FKP instead generates spatially-invariant kernels.

In the first solution, MANet was trained with different datasets (Proba-V, Flickr2K, WorldStrat) and applied to Proba-V pictures in order to obtain blur kernels. After evaluating the aforementioned CR set, a modified version of PIUnet was then trained and validated; this variant also included a MANet module to evaluate and concatenate kernels in the SR process.

The DIP-FKP architecture, since it did not need to be trained, was directly applied to Proba-V LR images. The obtained kernels were then used to derive CR images and train PIUnet. For this method, it was used the plain version of PIUnet, therefore without modifying the SR process.

Additionally, this thesis suggests a technique for optimising the plain PIUnet network using the so-called PIUnet-FKP, a novel architecture that draws inspiration from DIP-FKP. Tests conducted on the Proba-V datasets demonstrated favourable outcomes for this method in relation to the corrected peak signal-to-noise ratio (cPSNR), indicating a progression towards the supervised case's performance. It was able to outperform both the proposed MANet and DIP-FKP solutions, as well as other techniques proposed in the literature for blind SR. Even if PIUnet-FKP achieved good results, it still remains far from the state-of-the-art supervised PIUnet.

This work encompasses several components, including a description of the devised methodologies and the outcomes achieved. Additionally, it provides an overview of the context of image and satellite image super-resolution, along with a literature review of the existing methodologies in the field. Furthermore, it presents a theoretical background on the instruments employed in the study.

Acknowledgements

I would like to express my sincere gratitude to the exceptional individuals who have played a crucial role in providing direction, encouragement, and help during the course of my academic endeavour, leading to the successful completion of my master's thesis.

Firstly, I would like to thank professors Enrico Magli and Diego Valsesia for giving me the chance to do research on a fascinating and challenging topic for my thesis while benefiting from their invaluable guidance and expertise.

I also want to thank my closest friends, Picci, Matte, Fede, Lollo, Sue, and Michi, for their unwavering support at both the best and worst times of my life, some from the start and some later on, and for helping me grow and become what I am. To me, you are like brothers.

I would like to express my gratitude to both of my parents, my mother and my father, for everything they have done for me throughout my life. Thank you, mom, for being there for me anytime I needed an ally or someone to rely on, and for your attempts to inculcate in me the virtues of patience and serenity. I am thankful to you, dad, for always having believed in me, supporting me, and teaching me the principles of never giving up and always aiming for the top. All of these things have helped me become the person I am today. I have always looked up to you and always will.

Lastly, I want to say that I am also grateful to the other people who have been a part of my life over the past two years and who I have not yet named. You have all, in various ways, helped me grow as a person.

Table of Contents

List of Tables

List of Figures

Acronyms

HR

high-resolution

LR

low-resolution

CR

coarse-resolution

SR

super-resolution

SISR

single-image super-resolution

MISR

multi-image super-resolution

NN

neural network

CNN

convolutional neural network

FNN

feedforward neural network

EO

earth observartions

PSNR

peak signal-to-noise ratio

MSE

mean square error

Chapter 1

Introduction

The primary objective of this thesis is to investigate a deep learning approach for addressing the multi-image super-resolution (MISR) problem, specifically within the blind or unsupervised framework while only utilising satellite imagery. While vital for many applications, such as environmental monitoring, urban mapping, disaster assessment, and others, satellite images are quite a challenging field for MISR since those pictures can suffer from different types of interference. These interferences include on-board sensor electronic noise, motion blur, but also adverse atmospheric conditions, such as occlusion due to clouds or changes in illumination between images of the same scene, and human activity. Another drawback of satellite images is that, with a few notable exceptions, such as the Proba-V challenge dataset [1], high-resolution (HR) and low-resolution (LR) images are typically obtained from separate platforms, resulting in additional disturbances related to the mismatched radiometric properties between satellites. This indicates that it may be advantageous to devise a super-resolution (SR) method that use only LR images, which are widely available and simpler to collect.

The *blind* application to super-resolution implies that no HR images are available for the training process. This means that the degradation that LR images have undergone is unknown, and therefore it must be carefully modelled. The degradation process can be modelled with a *spatially-variant*, anisotropic, or *spatially-invariant*, isotropic, degradation kernel $\mathbf{K}_t$, where t represents the time dependency. Having a spatially variant kernel associated to an image means that the degradation of each pixel of that image is modelled differently.

The degradation and spatial dimension reduction process for spatially variable kernels can be modelled as follows:

$$\mathbf{x}_{LR,t} = [\mathbf{K}_t \mathbf{x}_{HR}]\downarrow_s + \mathbf{n}_t \qquad t = 1, ..., T \tag{1.1}$$

where $\mathbf{x}_{LR,t}$ and $\mathbf{x}_{HR}$ represent the time-dependent LR image and HR images,

$\downarrow_s$ represents the spatial downsampling with scale factor s, and $\mathbf{n}_t$ is the additive noise. For spatially-invariant degradation kernels, the model is instead reduced to a convolution operation.

The orthorectification method, which is frequently employed for image fusion simplification, introduces an additional kind of deterioration that may be characterised as both spatially variable and time-dependent [2]. Consequently, it is reasonable to employ a time-dependent kernel and consider incorporating spatial variability to enhance its ability to represent the data.

As stated by Shocher et al. [3], the first step in unsupervised or blind SR should be to assume that the SR function is roughly scale-invariant. By making this assumption and knowing the degradation kernel, it is possible to artificially reduce the quality of the LR images to a coarser resolution (CR), using the degradation mechanism described in Equation 1.1; the scale factor between CR and LR and between LR and HR must remain the same. Using CR and LR images, a SR neural network can be trained in a supervised manner and expect it to generalise when tested on real images, therefore super-resolving LR images to HR images.

In this thesis, the state-of-the-art PIUnet [4] was employed to perform MISR; since this network was designed to perform non-blind SR, it was used the aforementioned method to obtain the training sets. While PIUnet's performance under plain supervised training is widely known, just a small number of handcrafted degradation kernels were tested, by Prette et al. [2], to determine its performance with CR and LR training sets.

It is well established that training on the incorrect kernel results in poor generalisation [5, 6], in fact, this thesis deeply analyze the degradation kernel estimation problem. Two different methods for kernel generation were explored: one anisotropic, MANet [7], and the other isotropic, DIP-FKP [8].

Following the second method based on isotropic kernels, this thesis also proposes a method to fine-tune PIUnet by exploring a novel architecture inspired by DIP-FKP.

Before diving into the explanation of the methodologies, tests and results obtained for the aforementioned methods, this thesis explores the background and the state-of-the-art of image super-resolution and deep learning techniques for SR and kernel estimation. Few citations about known satellite image datasets are also made.

1.1 Super-resolution imaging

The extremely difficult process of estimating one or more high-resolution (HR) images from one or more low-resolution (LR) observations is known as super-resolution (SR). SR has a wide range of applications, such as satellite and aerial imaging, medical image processing, facial image improvement, text image improvement, compressed images, video enhancement, and others.

The SR techniques seek to deliver features finer than the sample grid of a certain imaging system; this is done by increasing the number of pixels per unit area in an image. Hardware-based approaches to the problem of increasing the number of pixels per unit area are usually expensive for large-scale imaging devices. Therefore, algorithmic-based approaches, i.e., SR algorithms, have gained significant interest from the computer vision research community and are usually preferred to hardware-based solutions.

These techniques can be split into two main categories: *Single Image Super-Resolution* (SISR) and *Multi Image Super-Resolution* (MISR). As the name suggests, single-image super-resolution is a process of enhancing the quality and resolution of a single LR image. MISR, on the other hand, uses several LR images of the same scene or object to make a single HR image. Having several photos of the same scene is particularly helpful because small geometric shifts allow the images to carry complementary information that, when properly combined using SR methods, can greatly improve the spatial resolution [4]. Multiple images of the same scene can be collected simultaneously by multiple sensors or captured in different time intervals by the same sensor.

Both SISR and MISR have been investigated and developed using a variety of SR methodologies, but it is significant to note that, in recent years, deep learning and Convolutional Neural Networks have become increasingly prevalent in both cases; some of these methodologies and techniques have been comprehensively reviewed by Park et al. [9] and Nasrollahi et al. [10].

While SISR is a mathematically ill-posed issue in which missing pixel information is partially imputed to provide a reasonable and pleasant visual result, MISR requires some level of aliasing in the LR frames to make up for the missing HR frequencies. Simply put, MISR is feasible if at least one of the imaging model's contributing parameters varies from one LR image to the next; the parameters could be optical, atmospheric, motion blur, zoom, multiple images from different sensors, and different colour channels. Before the actual HR reconstruction it is therefore required to compensate for these changes. Blur/filter estimation and geometric registration, which addresses the geometric misalignment between LR images, are two common compensation techniques. [10]

One of the first techniques developed for MISR was the *Iterative back projection* method proposed by Irani et al. [11]; the algorithm starts with an initial guess of the target HR image, usually obtained after registering and averaging the LR images over an HR grid, and then refines it. The *Bayesian* methods, instead, entail combining statistical data from various images and modelling the uncertainty in the super-resolution process; one solution of this kind was proposed by Zontak et al. [12]. As for SISR, deep learning and CNNs have also been instrumental in advancing MISR techniques since they can capture intricate patterns and relationships across multiple images. Some works that use CNNs for MISR are DeepSUM [13], which won the Proba-V SR challenge [1], and PIUnet by Valsesia and Magli, which suggest a method that is completely invariant to temporal permutation and also measures the uncertainty of the super-resolved image.

Apart from SISR and MISR, two approaches may be distinguished within super-resolution: blind and non-blind. Blind super-resolution refers to the task of enhancing the resolution of an image without having full knowledge of the underlying degradation process that caused the loss of resolution, therefore not having the original HR version. It is called *blind* because the exact degradation model is unknown or uncertain. Since the algorithm or model must deduce the degradation process from the observable data, blind super-resolution is quite challenging. In many cases, multiple degradation processes could have led to the observed low-resolution data. In *non-blind* super-resolution, as the negative prefix suggests, the degradation process that led to the loss reduction in the image is known. In this case, information about the HR image and how the LR image was generated from it is available; this information can include details about the blur kernel, downsampling method, noise, etc. An example of non-blind SR architecture is the above-mentioned PIUnet [4], which in fact uses both LR and HR images to train the neural network.

1.2 Satellite image super-resolution

Mapping the Earth with high-resolution images is vital for a variety of applications such as environmental monitoring, weather forecasting, urban mapping, disaster assessment, military surveillance, and many others. While being extremely useful, satellite instruments must contend with limitations imposed by factors like payload sizes, downlink bandwidth, etc. that may reduce the spatial resolution of the pictures they capture or the timing of the availability of HR products. Spatial resolution is usually referred to as metres over pixel ($m/pixel$); this means that HR images will show fewer metres for each pixel, resulting in a bigger HR representation with respect to the LR one. The quality of the pictures taken from Earth Observations (EO) satellites is also influenced by several effects related to the high-speed motion of the satellites, many of which are already taken into account by HR sensors but not by LR ones. Satellite image super-resolution addresses the above-mentioned problems utilizing deep learning techniques and by estimating HR images from one or more LR pictures. SR allows for wider use of LR sensors, which require less bandwidth, usually weigh less, and cost less with respect to HR sensors.

MISR, in particular, is well suited for satellite image SR since EO satellites have fixed orbits and therefore can easily acquire multiple LR images of the same scene during several orbits. It's worth noting that dealing with multiple images acquired in different time instants, while being beneficial from the data availability point of view, can be very challenging, as the scene's content may change due to various reasons,

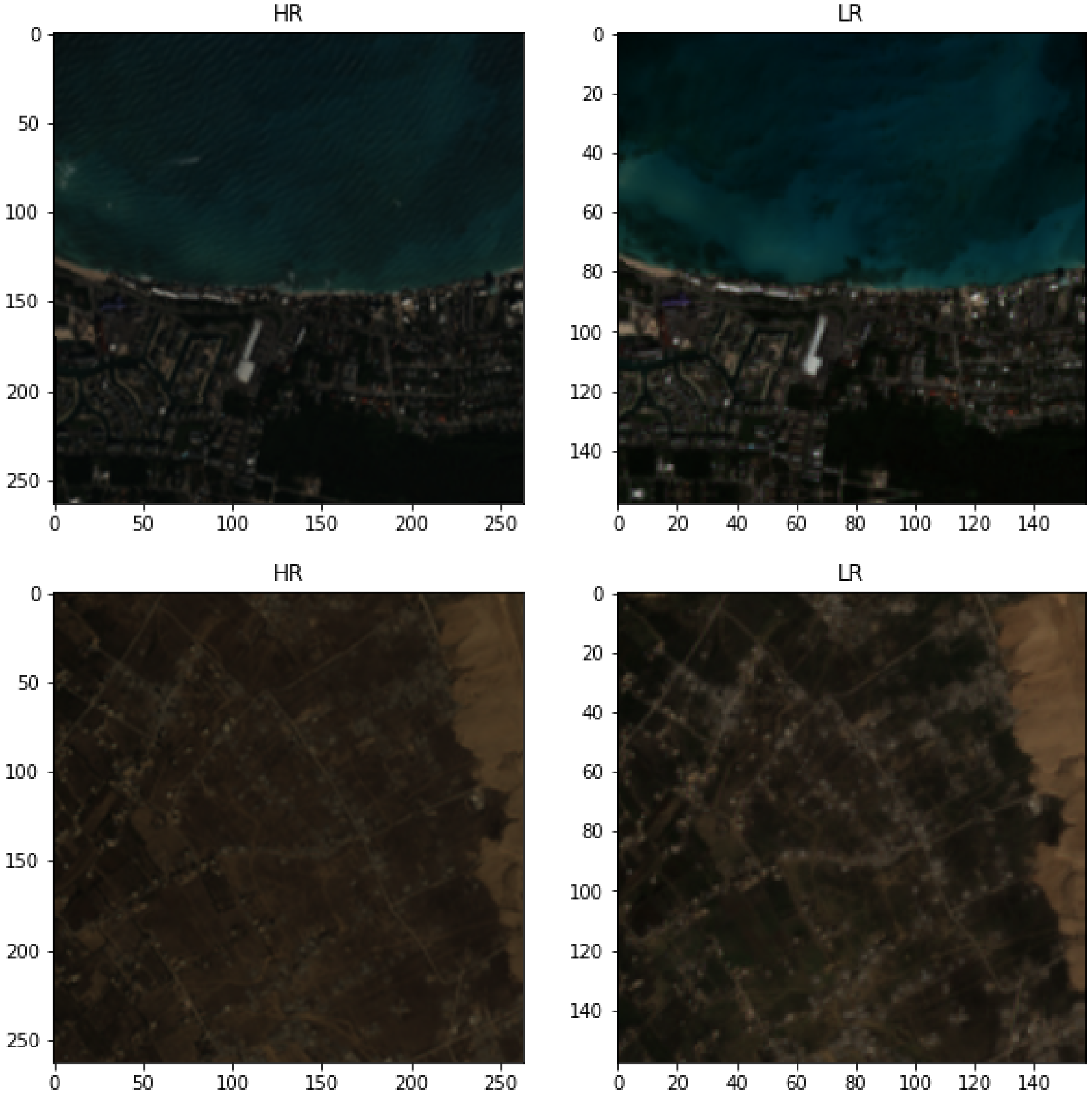

Figure 1.1: Examples of High Resolution and Low Resolution images taken from the WorldStrat dataset [14] and obtained from the same scene.

such as changes in illumination, occlusions due to clouds, human activity, etc.

Significant development has been made on the satellite image SR task, also thanks to the previously mentioned Proba-V challenge [1] by the European Space Agency. The winner network of the challenge, DeepSUM, was further improved with

DeepSUM++ [15], which introduces non-local operations. The current state-of-the-art is instead represented by the RAMS [16] model, which exploits feature attention at multiple stages (details on feature attention later 2.2.1).

More recently, the WorldStrat dataset [14], a publicly available dataset of high-resolution satellite imagery, was created to enable a worldwide analysis of terrain. With almost 10,000 sq km of distinct locations, this dataset is the biggest and most varied of its kind, ensuring a stratified representation of all land-use forms worldwide. WorldStrat is curated from high-resolution images captured by Airbus SPOT 6/7 satellites with a resolution of up to 1.5 m/pixel. Each high-resolution image is temporally matched with multiple low-resolution images from the freely accessible lower-resolution Sentinel-2 satellites at 10 m/pixel; Figure 1.1 shows some examples of HR and LR couples.

Chapter 2

Technical background

2.1 Proba-V dataset

The Proba-V super-resolution dataset collection, released by the European Space Agency, offers both high-resolution and low-resolution images captured by the same satellite. It can acquire HR photos at 100m resolution with a five-day revisit period and LR images at 300m resolution every day. Because of this, it is a great case study for supervised learning because it stops model performance from being messed up by degrading HR data to get LR pictures. However, when cloud cover and temporal variations are taken into account, only a restricted number of photos may be made accessible due to the longer revisit time for HR data. The dataset comprises single-band Level 2A images that fall into two categories: visible (RED) and near-infrared (NIR). For every HR scene, there are a minimum of 9 LR images that were obtained during a 30-day period. The available data for training are 415 RED and 393 NIR scenes, whereas the available data for validation with known ground truth are 176 RED and 170 NIR scenes. The spatial dimensions, in pixels, of each image are 384 × 384 for HR images, and 128 × 128 for LR images.

The datasets utilised in this thesis are taken from Proba-V and are characterised by their dimensions, denoted as $[N, X, Y, T]$. Here, N represents the number of scenes, while X and Y correspond to the spatial dimensions. T, on the other hand, represents the temporal dimension, indicating the number of images taken at different time instants and available for each scene.

2.2 Deep learning techniques

Deep learning techniques and neural networks, other than determining the decision rule as done in machine learning, learn the model of the data they want to describe.

Artificial neural networks consist of interconnected nodes or neurons that process

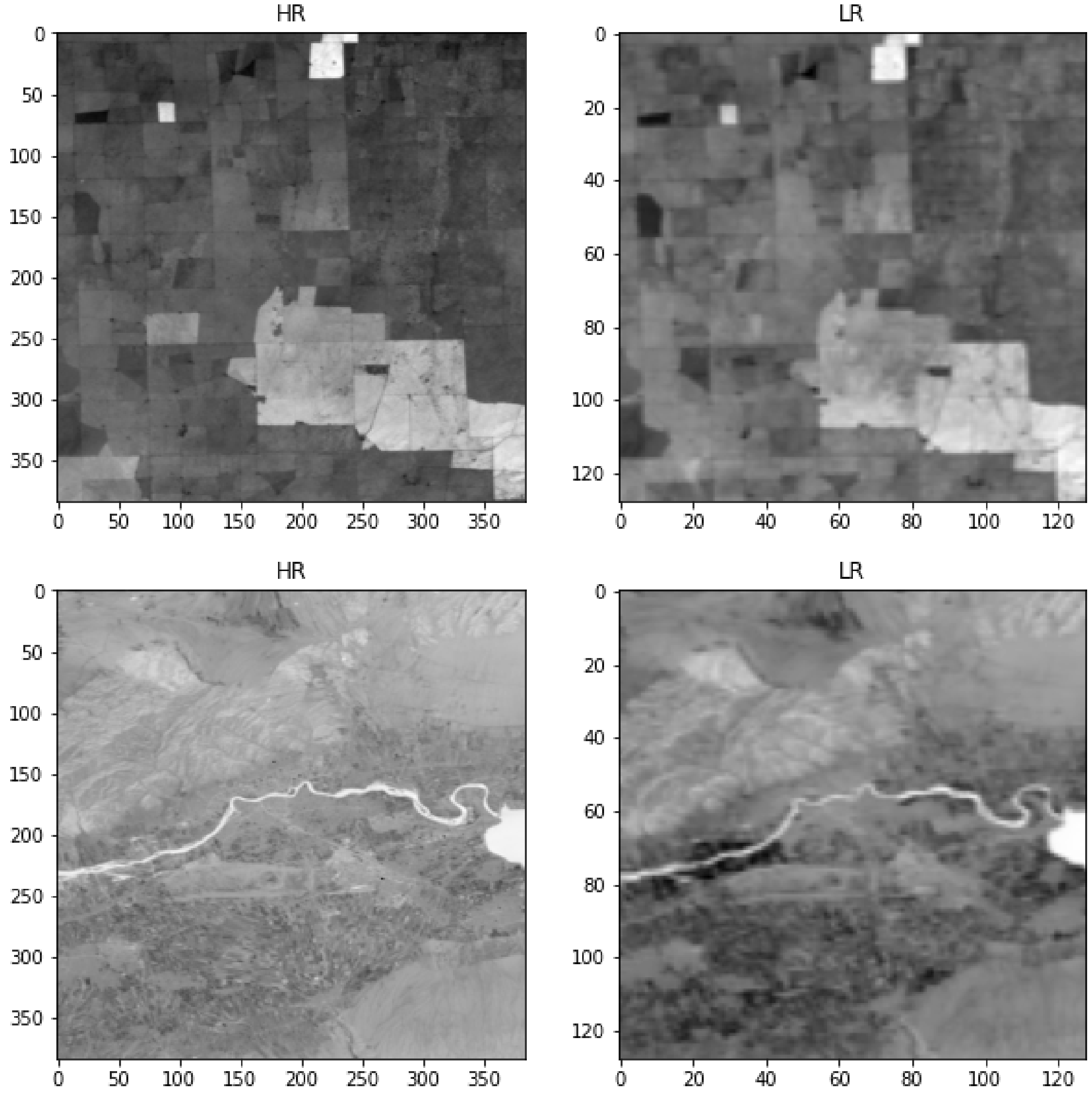

Figure 2.1: Examples of Proba-V near-infrared (NIR) LR and HR images taken from the same scene.

information and, when applied in the deep learning field, present many layers of neurons to learn from data. Each of these neurons receives input from other neurons, performs a computation, and produces an output that is passed on to other neurons; the connections between them have weights that determine the strength of the signal that is transmitted. Each layer of neurons performs a different type of computation:

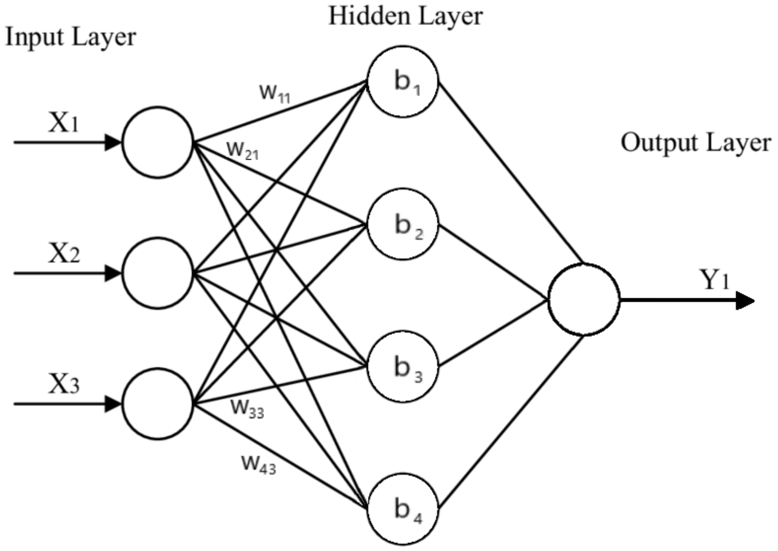

Figure 2.2: Simple feedforward neural network with one hidden fully connected layer. Each neuron has its correspondent bias value, e.g. b_1 for neuron 1, while each connection has a weight value, e.g. w_{21} for the connection between input neuron 1 and hidden layer neuron 2.

the input layer, which is the initial layer, receives the raw data; the network's ultimate output, which may involve a classification, regression, or generation job, is produced by the output layer; the intermediate layers in between are referred to as hidden layers, and they carry out calculations to change the input into a format that is better suited for the output job.

Figure 2.2 shows a simple example of *Feedforward Neural Networks* (FNNs), which present no loops between layers and have all the neurons of a layer connected with all the subsequent ones; it can be seen that the learnable parameters of a single layer are the **weights** $\mathrm{w_{j,i}}$ between connections and the neurons' **bias** $\mathrm{b_j}$. Another fundamental component of these networks is the activation function, a mathematical function that determines the output of a neuron. It is crucial to how neural networks work because it adds non-linearity to the model and makes it possible for neural networks to learn and reflect intricate, non-linear relationships in data; without non-linearity, a neural network would be equivalent to a linear regression model.

Convolutional Neural Networks (CNNs) are a type of deep learning neural network that is particularly well-suited for image recognition and analysis tasks. The key idea behind CNNs is to learn a hierarchy of characteristics that get more intricate and abstract as they go up, starting with low-level features like edges and corners and moving up to high-level features like shapes and objects.

In a convolutional layer, every input is not connected to every neuron; in fact, receptive fields refer to specific regions within the input data to which individual neurons are connected. The receptive fields are local since, in images, the content is usually limited to small regions and there are few long-term correlations among pixels that are very distant. This local connectivity is in contrast to the fully connected layers described before, where each neuron is connected to all neurons in the previous layer. The connections from the local receptive fields to each neuron have shared weights. Weight sharing means that the same set of learnable parameters, weights and biases, is used for all neurons within a *convolutional layer*; the shared weights define in this way a filter, the so-called convolutional kernel.

A complete convolutional layer consists of several different filters, each of which detects a different feature or pattern around the image; for example, a filter might capture the presence of edges at a specific orientation or the presence of textures in an image. As the network goes deeper into subsequent layers, neurons have access to larger and more abstract receptive fields, allowing them to capture more complex patterns and relationships. Each convolutional layer that presents multiple filters generates a multidimensional output map, or feature map, with a channel number equal to the number of filters applied to the input. To handle multiple input channels, which can be present by default in the image or obtained by the previous layers, convolutional layers use a different kernel for every channel and produce only one feature map by making a channel-wise summation of the convolutions' output (Figure 2.3).

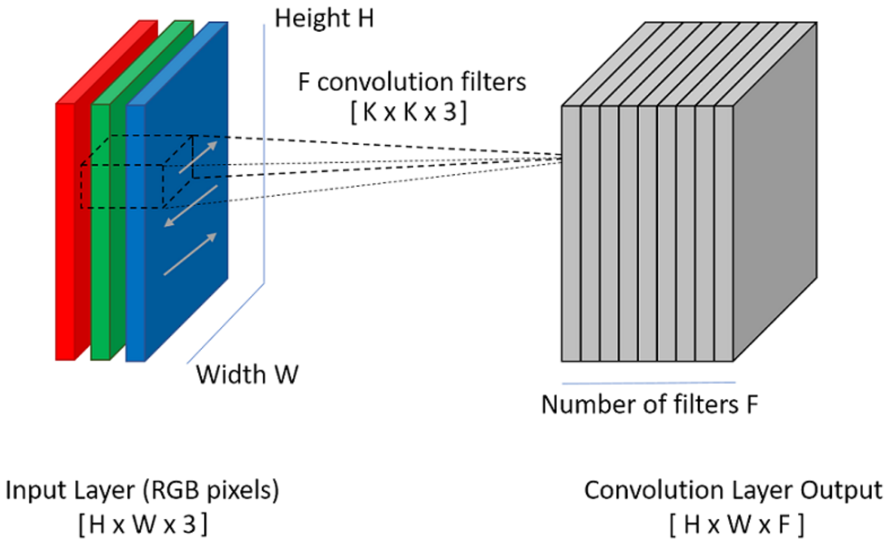

Figure 2.3: Example of convolution applied to a multi-channel image (RGB); K corresponds to the dimension of the convolutional kernel.

Convolutional neural networks built for image super-resolution tasks aim at increasing the resolution of an image, and therefore also its spatial dimensions while enhancing its visual quality. To increase the resolution of the images, CNNs use upsampling layers; these layers interpolate or expand the spatial dimensions of the feature maps, constructing a bigger representation of the input data. Some of the most common upsampling methods are bilinear upsampling, bicubic upsampling, and the PixelShuffle. Another fundamental component of super-resolution CNNs are the residual blocks; they were introduced by Kaiming He et al. [17], and contain skip connections that bypass one or more layers, allowing the output of one layer to be added directly to the output of a deeper layer. The skip connections facilitate the flow of gradients during backpropagation, which allows for the training of very deep networks, with a greater representational power.

2.2.1 Self Attention

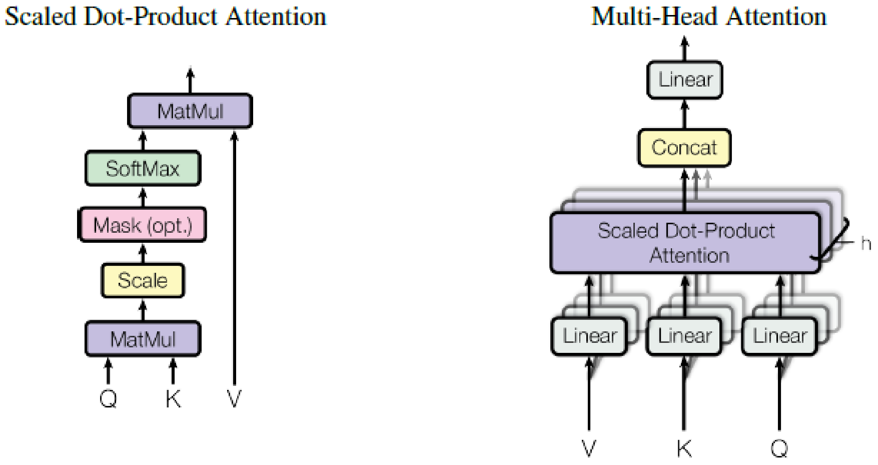

Figure 2.4: (left) Scaled Dot-Product Attention schema, from bottom to top. (right) Multi-Head Attention composed of h attention layers that operate concurrently, from bottom to top. From “Attention is all you need” [18].

Attention mechanism has significant importance in a range of deep learning models, especially within the domains of natural language processing and computer vision; it was first introduced by Vaswani et al. in the paper "Attention is all you need" [18]. This technique is derived from the notion of how individuals exhibit discerning attention towards specific aspects of information throughout the process of cognitive processing. This capability allows a model to assign varying degrees of significance to distinct components within the incoming data, utilising this knowledge to generate predictions or make informed decisions. Specifically, the self-attention mechanism is used to capture intricate linkages and interdependencies between different elements within a single sequence or grid, for example, the pixels of an image. The self-attention layers consist of the following fundamental elements: Query (Q), Key (K), Value (V), and the weighted output. Query, Key, and Value are linear transformations of the input pixels, which are derived from three distinct learned weight matrices, W_q, W_k, and W_v. The output is instead determined by calculating a weighted total of the values. The weight allocated to each value is determined by a compatibility function that evaluates the query's compatibility with the relevant key.

The functioning of the so-called *Scaled Dot-Product Attention*, as described in the paper "Attention is all you need" [18], is illustrated in Figure 2.4. This attention mechanism performs the attention computation on a collection of queries, represented as the matrix Q, by utilising packed key-value pairs denoted as K and V. The outputs matrix is evaluated as:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{2.1}$$

where $\frac{1}{\sqrt{d_k}}$ is the scale factor, d_k is the dimension of both queries and keys and softmax is a NN layer, typically used as the output layer in classification problems, that transforms raw numerical scores into a probability distribution across multiple classes.

In their paper, Vaswani et al. introduced also the concept of employing multiple linear projections to dimensions of queries, keys, and values, as opposed to utilising a single attention function. On each of these projections, it is then performed the attention function in parallel. The final values are obtained by concatenating and projecting the output of the multiple attentions. Figure 2.4 illustrates also this structure. The utilisation of multi-head attention enables the model to simultaneously focus on input from distinct representation subspaces at various places. Averaging prevents this when there is just one attention head.

In this thesis, as introduced by Valsesia et al. in "Permutation invariance and uncertainty in multitemporal image super-resolution" [4], self-attention is used over the temporal dimension; it is utilised as a permutation-equivariant mechanism that concurrently facilitates the efficient integration of information across multiple time instants.

2.2.2 RegNet

The RegNet architecture was first introduced by Bordone et al. in their paper titled "DeepSUM: Deep neural network for Super-resolution of Unregistered Multitemporal images" [13]. This architecture functions as a method for image registration within the convolutional neural network framework; it achieves this by dynamically calculating customised filters and subsequently applying them to higher-dimensional image representations.

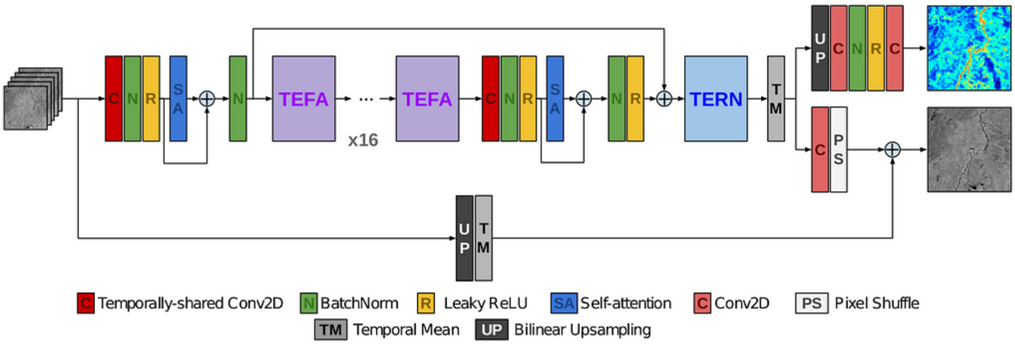

Figure 2.5: Architecture of PIUnet; it has two outputs, an uncertainty map and the SR image. From "Permutation invariance and uncertainty in multitemporal image super-resolution" [4].

RegNet operates with a set of N super-resolved images taken from a SISR block; all the images belong to the same scene. RegNet was conceived to align N-1 instances with respect to the first instance, which is considered as the reference, by only performing translational shifts and with an integer pixel precision. Consequently, the resulting outcome consists of a collection of N-1 2D filters that are to be spatially applied to every feature map of the N-1 input sources. The use of features by RegNet for the purpose of computing per-image optimum registration enables the exploitation of the network's feature space, hence enhancing the registration performance and ensuring its resilience to alterations in the scene.

2.2.3 PIUnet

PIUnet is a neural network designed to perform multitemporal image super-resolution, addressing also Permutation Invariance and Uncertainty estimation. It was proposed by Valsesia et al. [4] as a non-blind MISR network, while Prette et al. [2] already tested it in an unsupervised context.

Permutation invariance guarantees that any permutation in the temporal dimension of the LR inputs will not influence the output, making the model more robust. The SR image and the uncertainty of each pixel of that image are evaluated concurrently by using two parallel heads; Figure 2.5 shows an illustration of the PIUnet architecture.

This network accepts an indeterminate number of LR images as input, assuming that they have undergone preprocessing to achieve approximate registration with one another.

The fundamental mechanism in PIUnet is the self-attention: it is applied in the temporal dimension to exploit the correlation between multiple time instants and effectively combine their information. The backbone of the network is composed of repetitions of the TEFA block, Temporally-Equivariant Feature Attention, also proposed by Valsesia et al. [4]. This module uses 2D convolutions shared across the temporal dimension and temporal self-attention to extract spatial and temporal features; it then computes attention scores to weigh the feature channels. TERN, Temporally-Equivariant RegNet, is another module introduced in PIUnet, which instead is inspired from RegNet [13]. It adds the temporal-equivariant property to the RegNet architecture by using the self-attention to cross-correlate the features over the temporal dimension, instead of relying on an explicit image ordering (Section 2.2.2).

To perform SR and make the overall method invariant to temporal permutations, the output of the TERN module, which instead is equivariant to temporal permutations as the previous layers, is averaged over the temporal axis. The output is then upsampled and concatenated with the mean of the input LR images.

The second head of PIUnet evaluates an uncertainty map, pixel-by-pixel, in order to generate a quantifiable level of confidence for the super-resolution image. Its primary objective is to define the aleatoric uncertainty present in the SR image, for which the source is stochastic fluctuations in the input data.

2.2.4 MANet

MANet, *Mutually Affine Network*, proposed by Liang et al. [7], tackles the problem of estimating blur kernels directly from the input images rather than using a fixed set of them. Differently from other SR methods [5, 6, 8, 19], MANet estimates spatially variant kernels; spatially variant means that, for each pixel of the input image, the method evaluates a different kernel. While anisotropic kernels can increase the representational power, they are quite a challenging task due to the inherent locality of the degradation. MANet tackles this problem by using moderate receptive fields. It is also important to note that MANet takes as input just one image at a time.

MANet solves the problem of limited representational capacity caused by narrow receptive fields by adding a new layer called mutual affine convolution (MACon). The layers in question use dependency among various channels through mutual affine transformation as opposed to fully connecting all input and output channels, as seen in a standard convolutional layer. An affine transformation is a type of geometric transformation that preserves collinearity and ratios of distances between points. This particular design has the potential to enhance the capacity for feature representation and significantly decrease both the size of the model and the complexity of computations.

The MANet framework, shown in Figure 2.6, consists of two distinct modules, namely the feature extraction module and the kernel reconstruction module. The

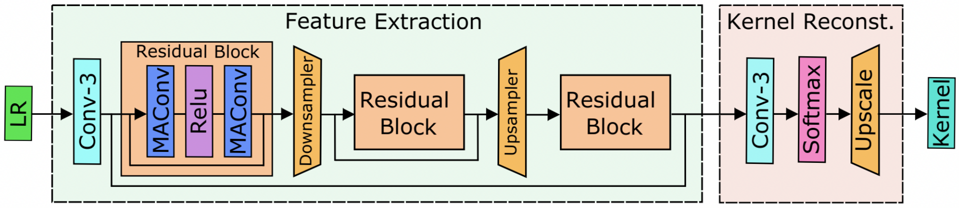

Figure 2.6: Architecture of the mutual affine network MANet. From "Mutual Affine Network for Spatially Variant Kernel Estimation in Blind Image Super-Resolution" [7].

feature extraction module uses a combination of convolution layers, residual blocks, a downsampler, and an upsampler to effectively extract features from a LR picture. The picture is further processed by a sequence of three residual blocks, each consisting of two mutual affine convolution layers with ReLU activation to enable the learning of non-linearity. The module further incorporates a convolutional layer and a transpose convolutional layer to perform downsampling and upsampling operations. In order to enhance the representation capabilities, two skip connections are incorporated. The kernel reconstruction module predicts kernels for each LR image pixel using a convolution layer and a softmax layer. The HR image's final kernel predictions are acquired through the utilisation of nearest neighbour interpolation. The mathematical representation of kernel prediction is expressed as $\mathbf{K} \in \mathbb{R}^{hw \times H \times W}$, where h, w, H, and W correspond to the dimensions of the kernel's height, width, HR image's height, and HR image's width, respectively. The architecture design of MANet guarantees that kernel estimation is not affected by picture patches that are more than 11 pixels apart, hence preserving locality in the representation.

2.2.5 FKP and DIP-FKP

FKP, which stands for *Flow-based Kernel Prior*, is a kernel estimation network created by Liang et al. [8]. It is based on normalizing flows and is meant to be added to existing blind SR models. Normalizing flows are generative models that use a series of invertible functions to change data from a complicated distribution to one that is simple and easy to work with. FKP is built, as represented in Figure 2.7, by stacking invertible flow layers, therefore becoming invertible itself.

The FKP framework employs many flow blocks to learn an invertible mapping between the kernel variable $\mathbf{k} \in K$ and the latent variable $\mathbf{z_k} \in Z$. Specifically,

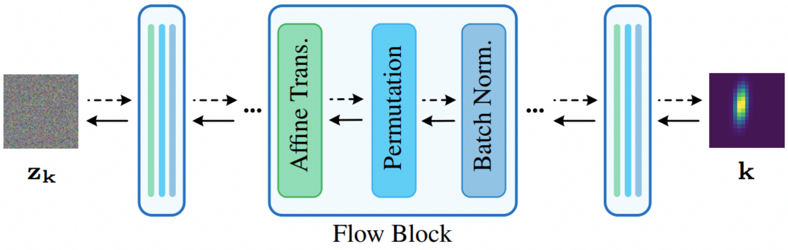

Figure 2.7: Architecture of the flow-based kernel prior (FKP) network. From "Flow-based Kernel Prior with Application to Blind Super-Resolution" [8].

each flow block consists of three sequential layers: a batch normalisation layer, a permutation layer and finally an affine transformation layer. Since $\mathbf{k}$ and $\mathbf{z_k}$ obey to two different probability distributions, Liang et al. define a bijection $f_\theta : K \to Z$. The bijection is parameterized by θ, which denotes the learnable parameters of the FKP network. By using this bijective function, the kernel $\mathbf{k}$ can be encoded as a latent variable $\mathbf{z_k} = f_\theta(\mathbf{k})$ in the latent space. Conversely, the value of $\mathbf{k}$ may be precisely reconstructed by the inverse mapping $\mathbf{k} = f_\theta^{-1}(\mathbf{k})$. The network parameters θ are optimized by giving a collection of training kernel samples and minimizing the negative log-likelihood (NLL) loss.

To perform kernel estimation, FKP is plugged into existing kernel estimation models as a kernel prior. This network picks a random sample of a latent variable $\mathbf{z_k}$, which corresponds to a random kernel. It then fixes the model parameters θ and updates $\mathbf{z_k}$ by gradient back-propagation under the guidance of kernel estimation loss. The model parameters θ are held constant while is updated $\mathbf{z_k}$ via gradient back-propagation under the direction of kernel estimation loss.

An example of blind SR architecture that incorporates FKP as kernel prior is DIP-FKP, also proposed by Liang et al. [8] and stemmed from the Double-DIP framework [20]. This architecture, Figure 2.8, is composed of two modules: DIP [21], a self-supervised and randomly initialised encoder-decoder network that works as image prior, and FKP for kernel estimation. During testing phase, FKP generates a kernel prediction to blur the SR image obtained by DIP; the resulting LR prediction is then compared with the LR input image by means of mean square error (MSE). Updating both DIP and FKP this architecture is able to produce simultaneously a SR image and a blur kernel from the LR input.

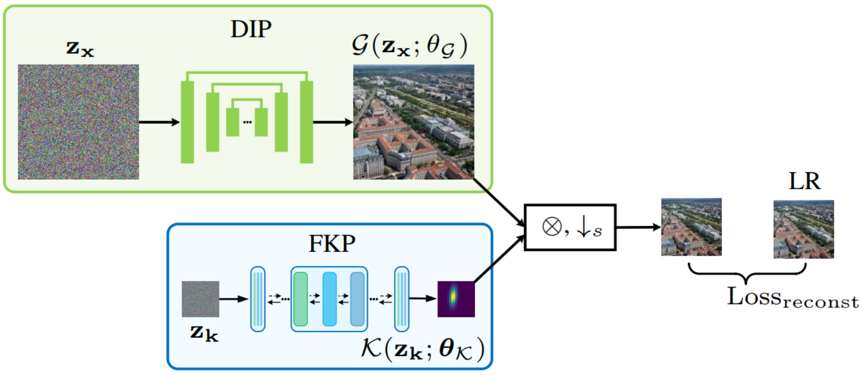
DIP
$\mathbf{z_x}$
$\mathcal{G}(\mathbf{z_x};\theta_{\mathcal{G}})$
FKP
$\mathbf{z_k}$
$\mathcal{K}(\mathbf{z_k};\boldsymbol{\theta}_{\mathcal{K}})$
$\otimes, \downarrow_s$
LR
$\text{Loss}_{\text{reconst}}$

Figure 2.8: Illustration of the DIP-FKP architecture. From "Flow-based Kernel Prior with Application to Blind Super-Resolution" [8].

Being DIP a self-supervised network, it has limited capacity in SR image reconstruction, in fact, this thesis explores a method that overcomes this problem by using a similar architecture but with the PIUnet framework.

Chapter 3

Proposed methods

This section examines the methodologies and the tests developed throughout the course of this thesis. As previously mentioned in the Introduction (Section 1), the PIUnet [4] architecture was employed as a baseline to perform MISR. In order to perform the majority of the tests it was used the Proba-V [1] challenge dataset, which contains 10 images for each LR scene, taken in different time instants, along with the correspondent HR version, which was used only for validation purposes. As explored by Prette et al. [2], PIUnet was stripped of its uncertainty head and used in a blind application. To utilise PIUnet in the reverse scenario, blind SR, it was necessary to create a new dataset of LR-HR image pairs and assuming scale-invariability of the SR function [3]. The dataset is evaluated as follows: for each scene, 9 LR images are degraded to a coarser resolution, the CR images, while 1 is preserved without any additional degradation. The CR and LR images are then used as the dataset for supervised training.

This thesis is thus focused on the evaluation of optimal degradation kernels to obtain CR images, which, with the LR ones, can be used to train in a supervised manner a modified version of PIUnet. To assess the performance of this network it was used the corrected PSNR, cPSNR [13]; this metric is different from the normal PSNR since it takes into account only pixels that are not hidden in both the target HR image and the reconstructed image.

The initial approach employed in the study utilises the MANet network [7]. In summary, the approach involves the supervised training of MANet in order to acquire a trained network that could be used for evaluating spatially-variant degradation kernels of any kind. Obtained the kernels and CR images, the modified version of PIUnet was tested with different losses.

The second method utilises the DIP-FKP architecture [8] to directly derive spatially-invariant degradation kernels for a predetermined collection of images. Subsequently, the LR pictures underwent degradation to CR, and the plain version of PIUnet was trained and tested.

Following an assessment of the efficacy of the second technique, a novel architecture inspired by DIP-FKP was employed to fine-tune PIUnet on the validation pictures, resulting in the creation of improved super-resolution versions.

3.1 Kernel estimation

3.1.1 MANet

This study proposes to use MANet as a plug-in network, exploiting its functionalities to evaluate CR images; it was also used inside the SR algorithm to reintroduce kernel information in the process, creating the so-called *Alternating PIUnet*.

In order to use MANet as desired, it must be trained beforehand. Liang et al. [7] have demonstrated its performances over the DIV2K dataset [22], which contains a collection of diverse real-world LR and HR images, but it was never publicly tested on satellite images, which, as explained before in Section 1.2, contain particular impairments.

The MANet training was performed with a pre-processed version of the Proba-V NIR train dataset; the scale factor between HR and LR images was $s = 3$, while the LR dimensions, in matrix representation $[N, X, Y, T]$, were [393,128,128,10]. Since MANet works with single LR images, the dataset was transformed to [3930,128,128], where the first dimension concatenates the number of images with the temporal dimension; moreover, as usual in the super-resolution context, the input data were split by creating multiple patches from a single image. The input dataset, after the aforementioned processes, turned out to be [35370,96,96], where 96 is the spatial dimension of the patches, and 35370 is the new number of input data, meaning that nine patches for each image were generated.

During MANet training, the image patches were blurred by random 9×9 anisotropic Gaussian kernels and then fed to the network. The spatial dimensions of the kernels, 9×9, were chosen to be small since it is expected that the influence of the degradation in satellite images is restricted to small pixel portions. The network was trained by means of L1 kernel loss, evaluating therefore the absolute difference between the real kernel and the estimated one, and using that error for backpropagation. Every 5000 iterations, the model parameters were collected; 5000 iterations correspond roughly to 30 epochs.

For the preliminary tests, the model from iteration 20000 was used. The trained model was used to derive the real kernels from the original Proba-V NIR train LR images. With the estimated kernels, the CR set was then created before training PIUnet. It is important to highlight that, even if the kernels were obtained from the same LR set, the CR images were evaluated by randomly applying the estimated kernels; randomness is used to boost the network generalisation power.

The plain PIUnet architecture was modified to perform these experiments; besides not using the uncertainty head, new layers were added to assess the effectiveness of re-introducing the degradation kernel information during SR. These layers, namely the DAN layers, concatenate in the temporal dimension the anisotropic kernels to the images and then apply a 3D convolution, normalisation, and multi-head self-attention. In order to use precise kernels, a pre-trained MANet module was added at the beginning of the network; this module went through each input patch and produced the relative degradation kernel. The kernel is then processed with the image as in the DAN layers. This new version of PIUnet was called *Alternating PIUnet*.

A L1 registered loss, which will be referred to as *SR loss* from now on, was used for the Alternating PIUnet training. This loss function measures the absolute difference between the LR picture (the SR target) and the reconstructed image, taking into account a pixel mask as well. It may be simplified as follows:

$$L_{SR} : |\, \hat{x}_{LR} - x_{LR}\,| \tag{3.1}$$

where $\hat{x}_{LR}$ is the reconstructed LR image and x_{LR} is the target LR image.

Two more loss functions were tested to be used with the aforementioned one: the *reCONstruction loss*, L_{CON}, and the *Kernel Estimation loss*, L_{KE}. Both losses are meant to be used together with the SR loss weighted by a coefficient, α and β respectively. L_{CON} looks at the error in reconstructing the CR input images, while L_{FE} computes the error in estimating the degradation kernels by comparing them to the down-sampled versions of the ones that were used to get the CR images. The losses are expressed as:

$$L_{CON} : \Sigma_t \big| \hat{K}_t(\hat{x}_{LR}){\downarrow_s} - x_{CR,t} \big| \qquad L_{KE} : \Sigma_t \big| \hat{K}_t - (K_t){\downarrow_s} \big| \tag{3.2}$$

where t corresponds to the time instants, $\hat{K}_t$ represents the estimated kernels, $\downarrow_s$ is the spatial down-sampling operation with scale factor s, and K_t represents the kernels used for CR evaluation.

After trying out different combinations of loss functions and loss coefficients, the MANet degradation kernels were looked at and plotted. This was done because the PIUnet results on the Proba-V validation set turned out to be worse than expected (Table 4.1). Figure 3.1 shows some examples of the degradation kernels compared with more classical Gaussian kernels.

Given the premise that the degradation kernels of satellite pictures should show similarities to widely used and conventional kernels, such as the Gaussian kernels, it becomes apparent that MANet or its associated training configurations were exhibiting erroneous behaviour. After that, more tests were done to find more common kernels and see how well they worked in actually training PIUnet.

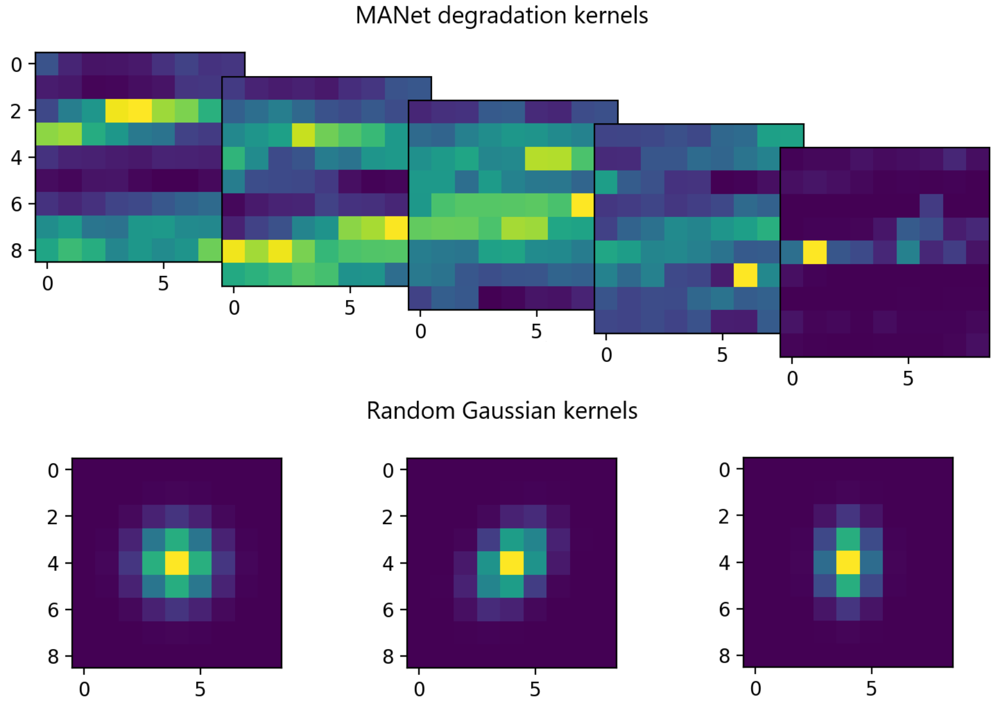

Figure 3.1: (top) Examples of single-pixel 9 × 9 degradation kernels obtained from MANet trained with Proba-V and applied to Proba-V NIR train LR images. (bottom) Examples of random 9 × 9 Gaussian kernels.

Before training MANet with datasets different from Proba-V, models taken from different iterations were tested. In particular, the model 5000 and the model 80000 were tested; both models, however, have given kernels similar to the ones in Figure 3.1.

MANet was also trained with the Flickr2K[22] and WorldStrat datasets [14]. For both datasets, during MANet training, were used their LR sets without patches, along with the settings described before for Proba-V. The kernels obtained showed a common behaviour inside each set and different behaviours between datasets; however, they were always conceptually different from a traditional degradation kernel. Figure 3.2 shows a few examples for both sets. The results obtained with PIUnet confirmed the claim: the cPSNR was always far away from an acceptable value, even worse with Flickr2K and WorldStrat.

Since MANet worked well with more common real-world images [7], it was thought that the issue might be with how it interacted with the satellite images and their particular degradation. Therefore, it was concluded that further studies are needed to use MANet in this application.

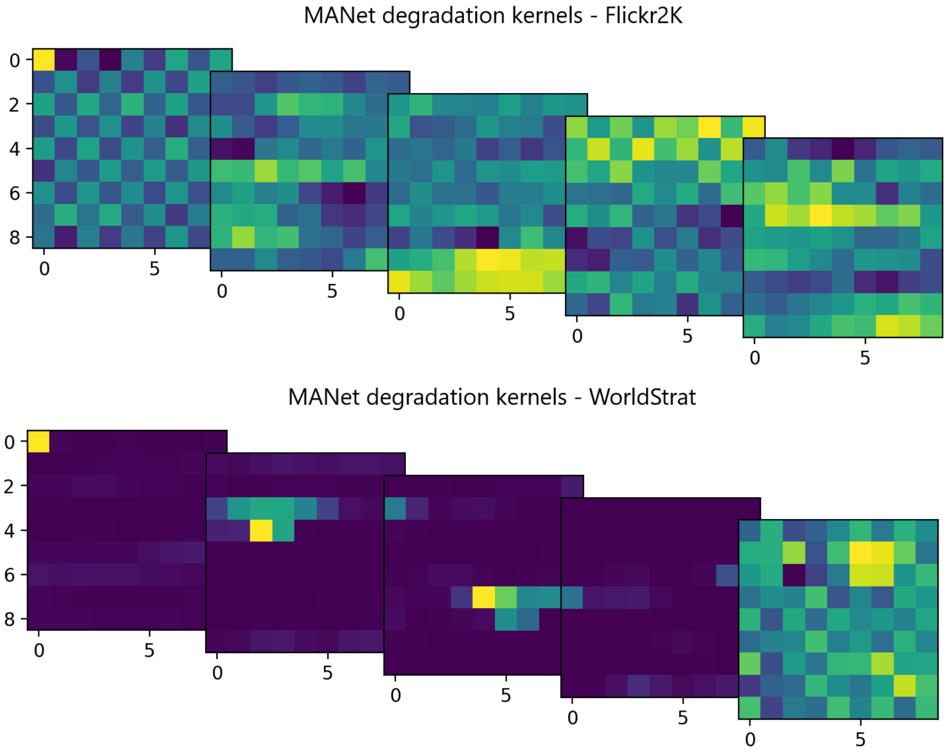

Figure 3.2: Examples of anisotropic 9 × 9 degradation kernels obtained through MANet applied to Proba-V NIR train LR images. On the top, MANet was trained with Flickr2K and on the bottom with WorldStrat.

3.1.2 DIP-FKP

After demonstrating the inability of MANet at generating realistic degradation kernels for satellite images, the architecture DIP-FKP [8] was experimented.

This architecture is composed of two modules: DIP, which is responsible for unsupervised SR and does not require any prior training, and FKP, which instead necessitates pre-training before its use. In this thesis, it was employed the pre-trained FKP model for scale factor $s = 3$ available on the FKP project GitHub [8]; the model was trained with random Gaussian kernels. The dimensions of the degradation kernels generated by the FKP module were also determined by the scale factor between LR and the desired HR image. Specifically, for a scale factor of 3, the dimensions of the kernels were 15 × 15.

The initial experiments with DIP-FKP were conducted utilising the Proba-V NIR train LR images. Since DIP-FKP must be optimised for each input image during the testing phase, it required a lot of computation time to complete the whole dataset. For this reason, it was initially employed a reduced version of the LR NIR dataset, which contains just one picture per scene. The total number of images used was, therefore, 393. It is noteworthy that Proba-V images have a single channel; however, DIP-FKP requires images with three channels representing colours. Consequently, it was imperative to preprocess the Proba-V images by converting them into a three-channel format, which was done by replicating the original channel three times.

By optimising the DIP-FKP architecture performing blind SR on each input image, the desired optimised set of isotropic degradation kernels was obtained. Figure 3.3 shows some examples of these kernels. The results show that DIP-FKP was able to make degradation kernels that are very similar to conventional ones, with the exception of few pixels that have non-zero values and are not in the core distribution. This peculiar behaviour may be attributed to the degradation caused by the satellite.

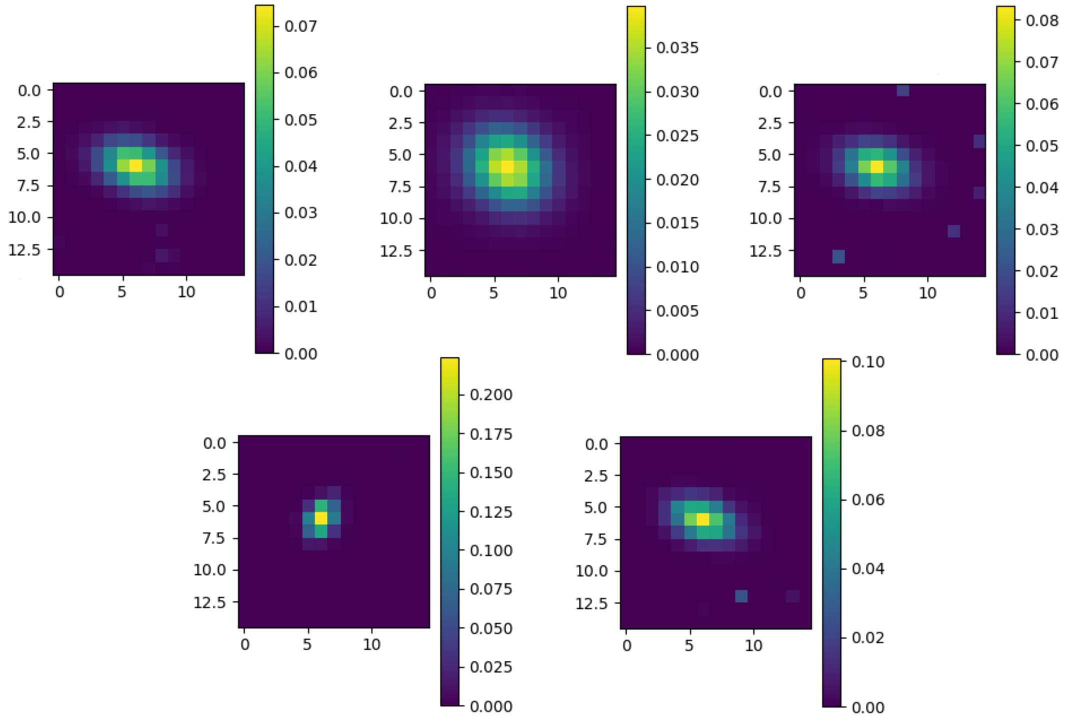

Figure 3.3: Examples of isotropic 15 × 15 degradation kernels evaluated by DIP-FKP while tested with Proba-V NIR train LR images.

Subsequently, the kernels were employed to generate the CR set and train the plain version of PIUnet, exluding the uncertainty head. The CR set was made by applying the degradation kernels randomly to the Proba-V NIR train LR images. PIUnet was trained using the traditional SR L1 loss, and its performance was assessed on the validation set using the cPSNR metric. The results obtained during the validation process suggested that using DIP-FKP to evaluate degradation kernels of satellite images is a viable option, as it yielded results that are closely aligned with those obtained using other manually designed kernels [2].

3.2 PIUnet fine-tuning

Finally, this thesis explores a novel method to enhance the performance of a pre-trained PIUnet model. The proposed method employs the DIP-FKP architecture, whereby the DIP module is substituted with PIUnet, resulting in the formation of the PIUnet-FKP framework. The training process of PIUnet was carried out according to the methodology outlined in the preceding section (Section 3.1.2). This involved utilising the DIP-FKP kernels and then gathering the network parameters from the best model.

The architecture PIUnet-FKP was designed with the intention of utilising the information acquired by PIUnet during the training phase on the whole dataset, and adapting it to single validation images at a time. This architecture was trained using the Mean Square Error (MSE) method. It involved calculating the discrepancy between the LR input picture and the degraded SR image generated by the PIUnet model. Indeed, FKP is used to generate kernels of increasing precision, resulting in a more accurate blurring of the SR image.

In order to demonstrate the capabilities of this architectural design, experimentation was conducted using the Proba-V NIR and Proba-V RED validation datasets. In the test, a total of 50 iterations were performed for each input picture. During each iteration, a SR image and a degradation kernel were generated using PIUnet and FKP, respectively. Subsequently, the SR image was blurred with the degradation kernel using a convolution operation and then down-sampled. The down-sampled image was then compared with a random LR input image, and the MSE was evaluated. By means of backpropagation, the latent variable of FKP was trained and PIUnet fine-tuned. The best model was chosen by evaluating the cPSNR at each iteration. The best cPSNR scores (Table 4.2), which were obtained with NIR and RED sets, demonstrate that the suggested strategy could be successful in enhancing the performance of PIUnet in the blind scenario.

Chapter 4

Results

The previous chapter provided a detailed explanation of the various approaches employed, tests carried out, and first results achieved in the endeavour to improve satellite picture resolution in a blind context. This chapter provides a thorough explanation of the significant findings derived from these methodologies, clarifying their effectiveness, constraints, and comparative evaluations.

The first methodology concerns the use of MANet [7] as a plug-in network for kernel estimation. MANet was trained by fixing some parameters and changing the training datasets. The fixed parameters were the scale factor $s = 3$, the kernel dimension, 9×9, and, for the generation of the random Gaussian kernels, they were $\sigma_1, \sigma_2 \sim \mathcal{U}(0.7,10)$ and rotation angle $\theta \sim \mathcal{U}(0, \pi)$. For what concerns the training dataset, Proba-V [1], Flickr2K [22] and WorldStrat [14] were used.

Figures 3.1 and 3.2 show some of the kernels that were obtained by using different trained models of MANet on Proba-V images. Figure 4.1 instead shows a comparison between high-resolution (HR) images, down-sampled and blurred with kernels, and the correspondent low-resolution (LR) picture. The blur kernels were obtained by applying MANet, trained with the three datasets, to Proba-V NIR train LR images. As can be seen, all three datasets led to wrong artificial LR images. These representations maintain a few of the principal characteristics; for example, in the first scene, it can always be identified the crossing wavy line, while in the second, the whiter spot on the top-left angle. However, these representations over-smooth the image, losing most of the details. The worst representations, as anticipated in Section 3.1.1, were obtained from Flickr2K and WorldStrat.

The Alternating PIUnet architecture (Alt PIUnet) was trained with CR and LR obtained from the Proba-V NIR train set and validated with the Proba-V NIR validation dataset. To obtain the CR set, only the MANet model trained with Proba-V was used, since the other models seemed to perform way worse. In particular, three loss functions were tested: SR loss (L_{SR}, equation 3.1), reconstruction loss (L_{CON}), and kernel estimation loss (L_{KE}, equation 3.2). L_{SR} was always used, while the

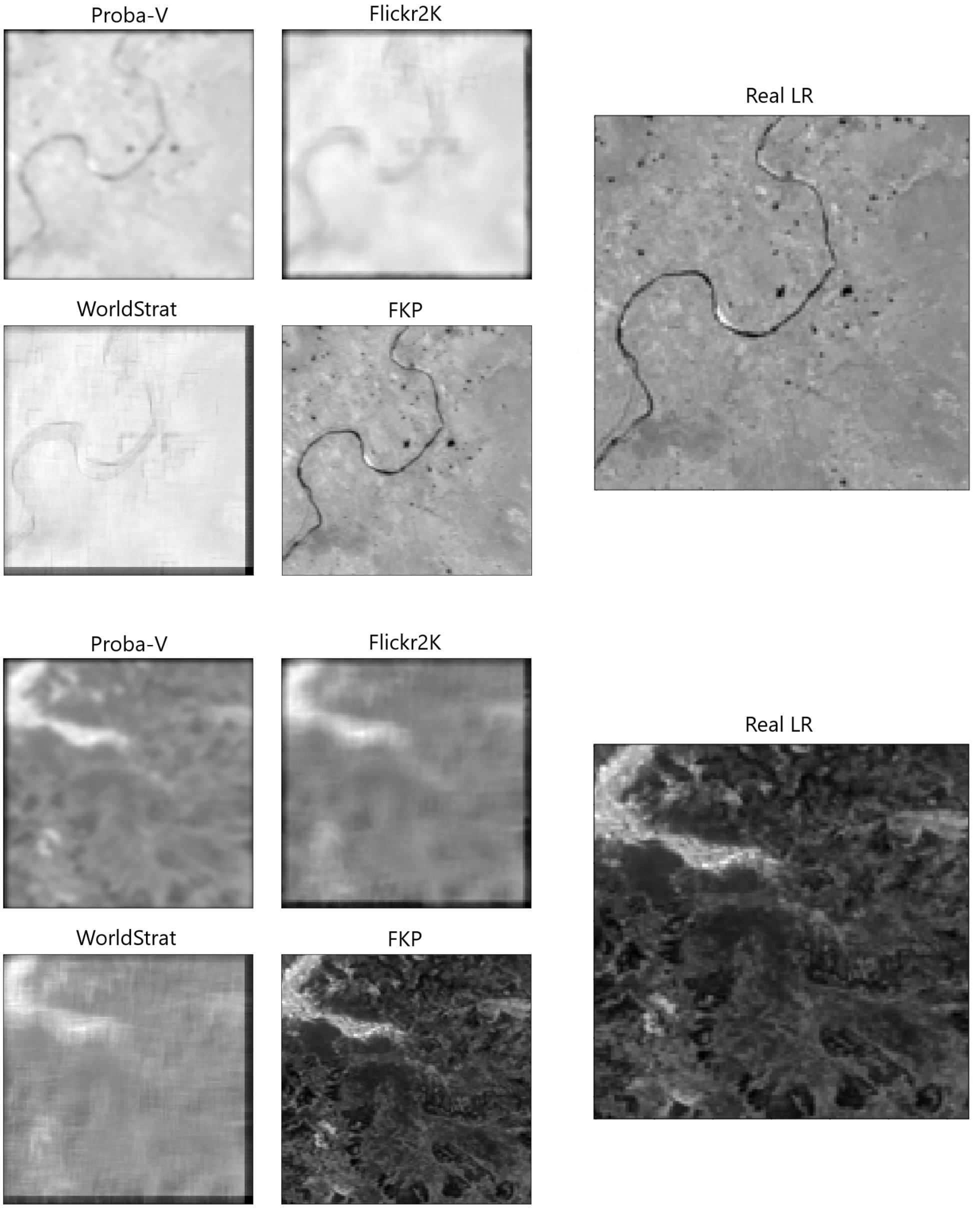

Figure 4.1: Real LR Proba-V images compared with the HR versions, which were down-sampled and blurred. The blur kernels were obtained using MANet trained with Proba-V, Flickr2K and WorldStrat, and with FKP.

other two were combined with the first loss using a coefficient, α for L_{CON} and β for L_{KE}. The quality of the SR images, compared to HR ones, was evaluated with the cPSNR metric [13]. Table 4.1 reports the mean cPSNR obtained for the validation set with different loss combinations, compared with classic SR methods and the state-of-the-art supervised PIUnet. The cPSNR achieved by the blind methods, which were trained on images that had been over-smoothed, was higher than that of the classic methods (Bicubic interpolation + mean and IBP [23]), but as expected, it was much lower than that of the plain supervised PIUnet. As will be shown shortly, blind PIUnet can get better results. In fact, it was found that no combination of losses could produce higher cPSNR values because the main issue was with the kernels and the process of estimating them.

	Classic		**Unsupervised deep SR**			
	Bicubic int. +Mean	IBP	Alt PIUnet with L_{CON}	Alt PIUnet with L_{KE}	Alt PIUnet w/ L_{CON}+L_{KE}	PIUnet
cPSNR	45.44	45.96	46.13	46.04	46.00	48.72

Table 4.1: Quantitative performances obtained on Proba-V NIR validation set by different SR architectures; measured with cPSNR (dB).

The second methodology for which results were collected regards DIP-FKP [8]. As MANet, this architecture was used to obtain degradation kernels and then create the CR set. The FKP module of the architecture was pre-trained using Gaussian kernels and a scale factor $s = 3$; when the module is plugged in the DIP-FKP architecture, the same scale factor produces 15 × 15 degradation kernels. Examples of these degradation kernels can be found in Figure 3.3. It is evident from Figure 4.1 that the FKP kernels were able, differently from MANet ones, to blur the HR images and obtain a picture very similar to the LR one, suggesting that they could be used to perform blind SR.

To test the FKP-generated kernels, it was employed the plain version of PIUnet with the super resolution loss (L_{SR}); as datasets instead, both the Proba-V NIR and RED sets were used. Table 4.2 shows the performance, in terms of cPSNR, of blind PIUnet with FKP; it also compares that network with the results obtained by Prette et al. [2] with the same network but using different kernels. The three models studied by Prette are based on Gaussian filters, which can be spatially invariant (SI), spatially variable and not correlated, or spatially variable and correlated.

	SI kernel	Pixel kernel (uncorr.)	Pixel kernel (corr.)	FKP kernel	Fine-tuning
cPSNR (NIR)	46.78	46.69	46.98	46.84	47.06
cPSNR (RED)	49.02	48.99	48.97	48.86	49.25

Table 4.2: Quantitative performances of Blind PIUnet with different degradation kernels [2] and after its fine-tuning with the PIUnet-FKP architecture; measured with cPSNR (dB).

When PIUnet with FKP kernels was used, it produced good results, but not as good as Prette's results with Gaussian spatially-variant and correlated kernels, especially for the RED dataset.

Finally, the PIUnet fine-tuning process (Chapter 3.2) was looked into, and the results are shown in Table 4.2. The fine-tuning process was iterated 50 times for each input validation image. The mean iteration at which the best model was found was 5 for the NIR set and about 4 for the RED set. Moreover, in the first dataset, 47 images out of 170 were not improved, while in the second, they were 44 out of 176. The new PIUnet-FKP architecture improved the cPSNR that was obtained using only PIUnet with FKP kernels, and it also outperformed all of Prette's models, getting closer to the performance of the PIUnet non-blind case.

Chapter 5

Conclusions

The work of this thesis dealt with multiple methods to perform blind satellite image super-resolution (SR) using pre-existing deep-learning architectures. The research yielded promising results regarding the use of FKP kernels in blind SR, but conversely, it yielded unfavourable outcomes when employing MANet.

As it is proposed, MANet seems unable to be trained as it should be, therefore using random Gaussian kernels, and then generalize the kernel estimation onto arbitrary satellite image kernels. It has not been investigated further enough to establish the causes with certainty, but it was supposed that satellite image kernels needed more powerful and specialised methods to be reconstructed. So, it is suggested that more research be done on how the degradation of satellite images affects kernel estimation and how it can be distinguished from the degradation of images not taken from satellites.

When FKP was plugged into the DIP-FKP architecture, on the other hand, it was able to optimise the kernel estimation on a single image and therefore produce realistic degradation kernels. Visually comparing LR images and blurred HR images, the result is satisfactory, and the performance obtained with PIUnet demonstrated it. While achieving good results, this method leaves room for further improvements since blurred images still lose fine details, which are helpful in SR, while being degraded. The kernels also occasionally present strange artefacts, highlighting the strange behaviour of the degradation of satellite images.

Finally, the novel PIUnet-FKP architecture demonstrated the effectiveness of utilising FKP for the purpose of fine-tuning a pre-trained, unsupervised version of PIUnet, confirming the potential of PIUnet in a blind context. However, it necessitates deeper studies on a precise method to identify the best iteration and obtain the best attainable cPSNR.

Bibliography

[1] https://kelvins.esa.int/proba-v-super-resolution (cit. on pp. 1, 4, 5, 18, 25).

[2] Nicola Prette, Diego Valsesia, Tiziano Bianchi, and Enrico Magli. «Towards Unsupervised Multi-Temporal Satellite Image Super-Resolution». In: *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*. 2023, pp. 5135–5138. DOI: 10.1109/IGARSS52108.2023.10281856 (cit. on pp. 2, 13, 18, 24, 27).

[3] Assaf Shocher, Nadav Cohen, and Michal Irani. «Zero-Shot Super-Resolution Using Deep Internal Learning». In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3118–3126. DOI: 10.1109/CVPR.2018.00329 (cit. on pp. 2, 18).

[4] Diego Valsesia and Enrico Magli. «Permutation Invariance and Uncertainty in Multitemporal Image Super-Resolution». In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–12. DOI: 10.1109/TGRS.2021.3130673 (cit. on pp. 2–4, 12–14, 18).

[5] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. «Blind Super-Resolution With Iterative Kernel Correction». In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 1604–1613. DOI: 10.1109/CVPR.2019.00170 (cit. on pp. 2, 14).

[6] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. «Blind Super-Resolution Kernel Estimation using an Internal-GAN». In: *CoRR* abs/1909.06581 (2019). arXiv: 1909.06581. URL: http://arxiv.org/abs/1909.06581 (cit. on pp. 2, 14).

[7] Jingyun Liang, Guolei Sun, Kai Zhang, Luc Gool, and Radu Timofte. «Mutual Affine Network for Spatially Variant Kernel Estimation in Blind Image Super-Resolution». In: Oct. 2021, pp. 4076–4085. DOI: 10.1109/ICCV48922.2021.00406 (cit. on pp. 2, 14, 15, 18, 19, 21, 25).

[8] Jingyun Liang, Kai Zhang, Shuhang Gu, Luc Van Gool, and Radu Timofte. «Flow-based Kernel Prior with Application to Blind Super-Resolution». In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2021, pp. 10596–10605. DOI: 10.1109/CVPR46437.2021.01046 (cit. on pp. 2, 14–18, 22, 27).

[9] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. «Super-resolution image reconstruction: a technical overview». In: *IEEE Signal Processing Magazine* 20.3 (2003), pp. 21–36. DOI: 10.1109/MSP.2003.1203207 (cit. on p. 3).

[10] Kamal Nasrollahi and Thomas Moeslund. «Super-resolution: a comprehensive survey». In: *Machine vision and applications* 25 (2014), pp. 1423–1468. DOI: 10.1007/s00138-014-0623-4 (cit. on p. 3).

[11] M. Irani and S. Peleg. «Super resolution from image sequences». In: *[1990] Proceedings. 10th International Conference on Pattern Recognition.* Vol. ii. 1990, 115–120 vol.2. DOI: 10.1109/ICPR.1990.119340 (cit. on p. 3).

[12] Maria Zontak and Michal Irani. «Internal statistics of a single natural image». In: *CVPR 2011.* 2011, pp. 977–984. DOI: 10.1109/CVPR.2011.5995401 (cit. on p. 3).

[13] Andrea Bordone Molini, Diego Valsesia, Giulia Fracastoro, and Enrico Magli. «DeepSUM: Deep Neural Network for Super-Resolution of Unregistered Multitemporal Images». In: *IEEE Transactions on Geoscience and Remote Sensing* 58.5 (2020), pp. 3644–3656. DOI: 10.1109/TGRS.2019.2959248 (cit. on pp. 3, 12, 14, 18, 26).

[14] Julien Cornebise, Ivan Orsolic, and Freddie Kalaitzis. «Open High-Resolution Satellite Imagery: The WorldStrat Dataset – With Application to Super-Resolution». In: *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track.* 2022. URL: https://openreview.net/forum?id=DEigo9L8xZA (cit. on pp. 5, 6, 21, 25).

[15] Andrea Bordone Molini, Diego Valsesia, Giulia Fracastoro, and Enrico Magli. «Deepsum++: Non-Local Deep Neural Network for Super-Resolution of Unregistered Multitemporal Images». In: *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium.* 2020, pp. 609–612. DOI: 10.1109/IGARSS39084.2020.9324418 (cit. on p. 6).

[16] Francesco Salvetti, Vittorio Mazzia, Aleem Khaliq, and Marcello Chiaberge. «Multi-Image Super Resolution of Remotely Sensed Images Using Residual Attention Deep Neural Networks». In: *Remote Sensing* 12 (July 2020). DOI: 10.3390/rs12142207 (cit. on p. 6).

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. «Deep Residual Learning for Image Recognition». In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: `10.1109/CVPR.2016.90` (cit. on p. 11).

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. «Attention is All You Need». In: 2017. URL: `https://arxiv.org/pdf/1706.03762.pdf` (cit. on pp. 11, 12).

[19] Kai Zhang, Wangmeng Zuo, and Lei Zhang. «Deep Plug-and-Play Super-Resolution for Arbitrary Blur Kernels». In: *CoRR* abs/1903.12529 (2019). arXiv: `1903.12529`. URL: `http://arxiv.org/abs/1903.12529` (cit. on p. 14).

[20] Yosef Gandelsman, Assaf Shocher, and Michal Irani. «"Double-DIP": Unsupervised Image Decomposition via Coupled Deep-Image-Priors». In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 11018–11027. DOI: `10.1109/CVPR.2019.01128` (cit. on p. 16).

[21] Victor Lempitsky, Andrea Vedaldi, and Dmitry Ulyanov. «Deep Image Prior». In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 9446–9454. DOI: `10.1109/CVPR.2018.00984` (cit. on p. 16).

[22] Radu Timofte et al. «NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results». In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2017), pp. 1110–1121. URL: `https://api.semanticscholar.org/CorpusID:484327` (cit. on pp. 19, 21, 25).

[23] Michal Irani and Shmuel Peleg. «Improving resolution by image registration». In: *CVGIP: Graphical Models and Image Processing* 53.3 (1991), pp. 231–239. ISSN: 1049-9652. DOI: `https://doi.org/10.1016/1049-9652(91)90045-L` (cit. on p. 27).