Ludwig-Maximilians-Universität München
MSc Data Science
Master Thesis

Summer Semester 23

Domain Adaptation Techniques for blind super resolution applied to Thermal Remote Sensing

January 15, 2023

Student:
Massaro, Patricio
<p.massaro@campus.lmu.de>

Acknowledgments

Contents

1 Introduction

Remote sensing technology is an important source of Earth observation from different platforms and sensors, and it offers work on a large scale with cheap, fairly accurate, and faster results compared to the conventional methods.

Land Surface Temperature (LST) images at a high temporal resolution are of prime importance to efficiently monitor physical processes related to climate change such as water stress, evapotranspiration, wildfires or urban heat islands [1]. Additionally, LST has been approved as one of the high-priority parameters for the International Geosphere and Biosphere Program (IGBP) [2]. LST is retrieved from remote sensing images in the Thermal InfraRed (TIR) spectral domain. Most of the application mentioned before have very exigent requirements for the TIR data products, needing as many detailed images as possible (spatial resolution), but also having many images available per day to monitor such dynamic processes (temporal resolution). Unfortunately these requirements are not met by the current available products and research may be hindered by the lack of data. Bigger missions have high spatial resolution but have a revisit time of several days. Smaller satellite providers are deploying constellations of smaller satellites that provide very high temporal resolution, but their smaller payload result in low spatial resolution. Their sensors do not generally retrieve TIR data at a satisfactory spatial resolution for local scale applications and fine scale analysis, especially for highly heterogeneous environments like urban areas, diverse agricultural plots or sparse forests.

Super resolution (SR) is a post-processing technique that aims to increase the spatial resolution of images while preserving the physical consistency of the scenes. Using AI for SR has many applications in heterogeneus fields like medical imaging, computer vision and also remote sensing. While several deep learning architectures have been proposed with promising results in synthetic datasets, it remains a challenge to apply them to real data. This is because most of the assumptions made to generate the datasets needed for training do not represent the real world accurately. Throughout this thesis, super resolution of real world TIR data products coming from OroraTech mission FOREST-2 will be explored, so that better LST products can be developed and improvements in research may be achieved. Additionally, the impact of the degradation models used for dataset generation will be studied, and a framework that allows SR from LR TIR data products coming from any mission will be proposed.

In Chapter 2, a brief introduction to remote sensing and LST retrieval is presented, as long as the dimensions of quality of their products. In Chapter 3, the motivation of this work is presented, diving deeper into applications of TIR data and their requirements. The trade-off between spatial and temporal resolution that the current available products is also shown, making it the main driver super-resolution. In Chapter 4, the main techniques of super-resolution are presented, as well as the main challenges of the datasets used in the literature. The domain gap problem is also introduced, as well as the concept of blind super-resolution. In Chapter 5, the methodology of this work is presented, including the models architecture and the rationale behind the choices made, the degradation models used for dataset generation and the metrics used for evaluation. In Chapter 6, the data gathering process is introduced and the datasets that will be used in the experimentation are presented. In Chapter 7, the experimentation setup is

presented, including the training and parameters. In Chapter 8, the results of the experimentation are presented and discussed. Chapter 9 presents conclusions of this work, as well as the future work that can be done to improve the results obtained.

2 Thermal Remote Sensing

In general terms, remote sensing is the science and practice of acquiring information about an object without actually coming into contact with it. Remote sensing can also be defined as a technology for sampling reflected and emitted electromagnetic (EM) radiation from the Earth's terrestrial and aquatic ecosystems and atmosphere. This is typically done by recording images from airplanes and satellites to help identify or better understand features on the Earth's surface. A simple example of a remote-sensing instrument is a photographic or digital camera. A camera records energy in the form of light that is reflected from a surface to form an image. Most photographic cameras record visible light so that when we look at the photograph the image resembles the feature that was photographed. More sophisticated remote-sensing instruments are able to record energy outside of the range of visible light. Data from remote-sensing instruments can be recorded as images or, in the case of lidar, a series of point data.

2.1 Electromagnetic spectrum

The electromagnetic spectrum (EMS) includes wavelengths of electromagnetic radiation ranging from short wavelength (high frequency) gamma rays to long-wavelength (low frequency) radio waves. Most applications are focused on region of the spectrum starting in the ultraviolet and continuing through the microwave wavelengths. Optical sensors are used to measure ultraviolet, visible, and infrared wavelengths and microwave sensors are used for the microwave portion of the EMS. A fundamental physical principal that remote sensing relies on is that different features on the Earth's surface interact with specific wavelengths of the EMS in different ways. When working with optical sensors the most important property used to identify features on the Earth's surface is spectral reflectance, the ratio of the intensity of light reflected from a surface divided by the intensity of incident light. Different features have different spectral reflectance properties and this information can be used to identify individual features. For example, white sand reflects most visible and near-infrared light whereas green vegetation absorbs most red wavelengths and reflects most near-infrared wavelengths.

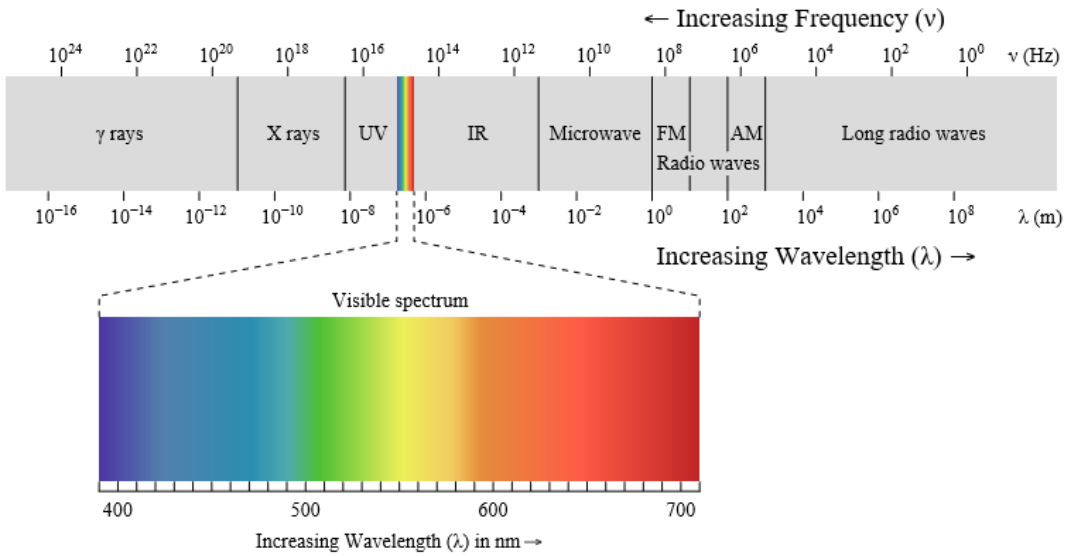

Figure 2.1: Electromagnetic spectrum

In the visual spectrum, usually reflected light is the measured. In the Infrared spectrum (IR), light emitted from a certain wavelength is measured. This is because all bodies have a temperature above absolute zero, and this temperature is detectable as radiation. IR spectrum goes from 1400 nm wavelenght to 1 mm and are further subdivided due to the width of the spectrum:

- Short-wave infrared (SWIR): 1,4 to 3,0 µm
- Medium-wave infrared (MWIR): 3,0 µm to 8 µm
- Long-wave infrared (LWIR): 8 to 15 µm
- Far-infrared (FIR): 15 µm to 1 mm

Like in the visual spectrum, TIR information is obtained in a purely passive way. Thermal infrared detectors called bolometers are mostly used on drones or ground stations due to their small volume, weight, and energy consumption. They offer a low accuracy in relative and absolute temperature measurement and are relatively inert, leading to a low ground resolution for moving systems (drones) but worthwhile for aircrafts with small payloads. More complex systems are generally cooled. Therefore, they offer the possibility of highly accurate relative and absolute temperature measurements. In contrast to bolometers, they are significantly larger and require more energy, making them unsuitable for use on drones. These types of detectors are used on all common TIR satellite systems.

2.2 Land surface temperature

Land Surface Temperature (LST) is the radiative temperature of the land—derived from previously mentioned thermal infrared radiation that the Earth's surface emits. LST is essentially a measure of how hot Earth's surface would feel to the touch in a particular

location. It is one of the key parameters that affect surface energy balance, regional climates, heat fluxes, and energy exchanges. When taking Earth's temperature, we're interested in thermal infrared (TIR) specifically, which has a wavelength of 3 to 14 µm (both MWIR and LWIR), corresponding to black body temperatures of between -60°C and 700 °C, according to plancks equation:

$$B_\lambda(T) = \frac{C_1}{\lambda^5 \left[\exp\left(\frac{C_2}{\lambda T}\right) - 1\right]}, \tag{1}$$

Where $B_\lambda(T)$ is the spectral radiance, C_1 and C_2 are physical constants, and λ is the wavelength of the radiation. For this reason TIR wavelengths are of particular interest for this work.

Accurate LST retrieval from TIR data depends on atmospheric effects, sensor parameters, i.e., spectral range and viewing angle, and surface parameters such as emissivity and geometry. Since emissivity and atmospheric effects are two fundamental factors to derive LST from thermal data, many researchers have proposed different approaches for LST retrieval considering these factors. These algorithms are named considering the number of TIR bands used. For instance, single-channel or mono-window algorithms use one TIR band. However, split window or multi-channel methods include more than one TIR band. Some examples of these methods are Radiative Transfer Equation (RTE) [3], Single Channel Algorithm (SCA) [4], and Mono Window Algorithm (MWA) [5]. As this work is focused on improving the resolution of the mapped radiances and not on the LST retrieval, the reader is referred to [6] for a more detailed review.

2.3 Quality dimensions of remote sensing data

Remote sensing data can be characterized by four quality dimensions, as stated in [7]:

- Spatial resolution: This is often simply referred to as resolution and is the size of a pixel in ground dimensions. In most cases an image's resolution is labeled with a single number, such as 30 m, which represents the length of a side of a square pixel if it were projected onto the Earth's surface. If the pixel were rectangular, then the length and width of the pixel would be provided.

- Spectral characteristics: This includes bandwidth, band placement, and the number of bands. Spectral bandwidth, or spectral resolution as it is often called, refers to the range of wavelengths that are detected in a particular image band. This is effectively a measure of how precisely an image band measures a portion of the EMS. Band placement defines the portion of the EMS that is used for a particular image band. For example, one band might detect blue wavelengths and another band might detect thermal wavelengths along the EMS. The properties of the features one is interested in sensing indicate which bands are important. The last spectral variable is the number of bands. The more bands that are available the more precisely spectral properties of a feature can be measured.

- Acquisition dynamics: This has two components. The first is the minimum time a particular feature can be recorded twice, often called the repeat frequency or

temporal resolution. Some sensors with a very wide field of view can acquire multiple images of the same area in the same day whereas some sensors have a repeat frequency of several weeks. The other component is the timing of the acquisitions. Dynamic features such as deciduous forests and events such as flooding often have an optimum time for which they should be imaged. For example, the identification of deciduous vegetation is aided by acquiring imagery during leaf-on and during leaf-off periods.

- Sensitivity of the sensor: This is defined by the dynamic range of the sensor as well as the range of digital numbers that can be used to represent the pixel values. Sensors have lower limits below which a signal is not registered and upper limits above which the sensor saturates and is unable to measure increases in radiance. The detail that can be measured between these extremes is determined by the range between the minimum and maximum digital numbers permitted for a particular data type. This potential range of values is often referred to as quantization or radiometric resolution.

3 Motivation

Land Surface Temperature (LST) images at a high temporal resolution are of prime importance to efficiently

Advancements in satellite technology, particularly in thermal infrared sensing, have emerged as a promising solution to monitor physical processes related to climate change such as water stress, evapotranspiration, wildfires or urban. Unlike ground-based methods, satellites can continuously monitor vast tracts of land regardless of smoke or geographical barriers, providing critical real-time data that can significantly enhance early detection and response efforts. However, while current satellite systems offer extensive spatial coverage and consistent data collection, they are not without limitations. In particular, in the quality dimension of spatial and temporal resolution

Throughout this section, the main applications that OroraTech is working on are presented, as well as the thermal data products requirements that the literature has identified as vital for their development. The trade-off between spatial and temporal resolution that the current available products will be further discussed, which is the main driver for super resolution.

3.1 Wildfire Monitoring

Forest fires can be natural or manmade phenomena that occurred in natural ecosystems and usually, they spread uncontrollably. The have increased steadily worldwide over the past decade, and according to the UN Environment Programme (UNEP), this trend will continue, with a potential 50% increase in forest fires by the end of the century [8]. The escalating frequency and intensity of wildfires across the globe have prompted a reassessment of our current monitoring systems and methodologies. Traditional ground and aerial surveillance methods are increasingly proving inadequate in the face of rapidly spreading, unpredictable fires, particularly those obscured by smoke and difficult terrain. This inadequacy hinders effective firefighting efforts and exacerbates the environmental,

economic, and human toll of these disasters. Luckily, in most cases, a layer of fume is not an obstacle for a satellite to detect a fire, as they rely on thermal infrared sensors that can measure radiance through the smoke.

Prevention is the most effective way to fight wildfires, and early detection is key to achieving this goal. With timely access to thermal data from space, we gain the tools to identify potential wildfires before they spread, minimizing damage and saving lives. Although satellite-based imagery is used by emergency response agencies to monitor large-scale wildfires that burn over extensive periods, the wait interval for a satellite overpass induces a considerable time delay, which prevents its application in time-sensitive fire detection scenarios, such as emergency evacuations, early detection or search-and-rescue operations [9]. OroraTech's looks to circumvent the overpass wait interval by launching a swarm of small satellite that have been especially helpful for detecting newly born fires that started as a result of a bigger fire spreading, burning the material around its vicinity. The team develops its own sensors capable of providing up-to-date thermal data of the entire Earth every 30 minutes, starting from 2026.

However, another parameter vital to measure fire risk, detect burn areas, and monitor vegetation recovery is the spatial resolution. The coarse spatial resolution that the smaller payloads these satellites provide does not enable the reliable detection of smaller fires. Additionally, higher spatial resolution is needed to detect better estimate the fire front, which is vital for the emergency response teams to plan their actions. Moreover, High resolution data has been used to validate characteristics of false positives in fire detection algorithms applied to low resolution scenes[10].

3.2 Urban heat

Productivity losses due to heat currently cost an estimated $100 billion annually only in the U.S alone. The U.K. experienced unprecedented temperatures too, temporarily knocking out services for giants like Google and Oracle and further affecting their clients. These are only a few examples of how businesses are affected by extreme temperatures in urban areas [11].

Public interest and concern about heat waves are steadily rising, especially in moderate climate areas such as North Europe. Heat waves are a major cause, specially in vulnerable population such as elderly people who are more prone to heat stress, pregnant individuals with difficulty adjusting to heat changes, outdoor workers who are exposed to extreme heat for prolonged periods and low-income neighborhoods with poor-quality buildings, between others [12].

Urban heat refers to the phenomenon where urban areas experience higher temperatures than their more rural surroundings. This effect can be attributed to various factors such as building geometry, thermal properties of building materials, radiation properties of surfaces, and anthropogenic heat release from sources such as traffic and industry [13]. In particular, the Urban Heat Island (UHI) phenomenon refers to the air temperature differences between urban areas and their surrounding rural regions, which is typically most pronounced during the evening and night hours.

Monitoring the UHI effect traditionally involves recording measurements from meteorological stations located in urban and rural areas throughout the region. However,

an alternative method is to use thermal remote sensing, which allows for the monitoring of large areas without the need for multiple physical sensors placed throughout the city. Research suggests that a Ground Sampling Distance (GSD) of 50m to 100m is the most suitable spatial resolution for urban thermal environment studies [14, 15, 16]. This spatial resolution requirement can be met by missions such as Landsat [17] and Terra [18], which provide a high spatial resolution in the thermal infrared band. However, their insufficient revisit time severely limits the analysis of dynamic processes with a temporal resolution in the order of hours, such as the UHI.

Existing studies on the UHI effect have been limited by the low spatial resolution of the images used and the lack of satellite images available at different times of the day [19, 20]. In particular, researchers explicitly state their limitations [20] due to the low frequency of revisits while studying the Park Cool Island (PCI) [21] phenomenon. The influence of vegetation cover on the urban heat island (UHI) phenomenon has been recognized as the foremost determinant [13], where parks have been found to exert a cooling impact. Figure 3.1 illustrates an example that demonstrates this influence. Converting the low-resolution images coming from private constellations into high-resolution images through post-processing techniques could enable new frontiers in the study of urban heat.

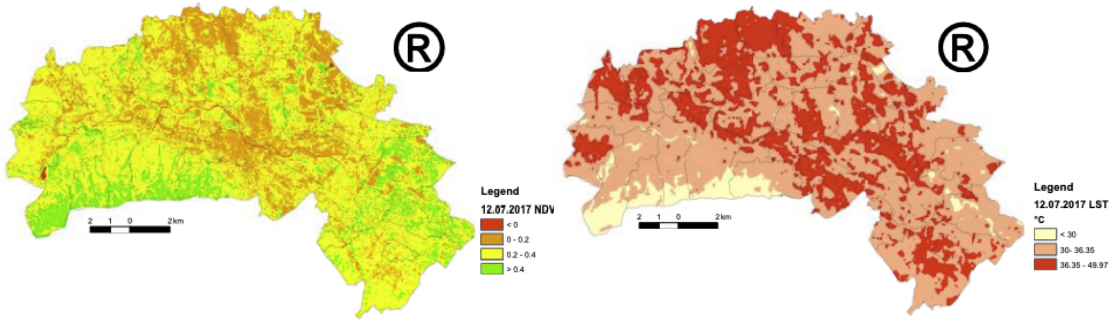

Figure 3.1: Normalized Difference Vegetation Index [22] and LST measurements for the zone of Skopje, North Macedonia. Urban areas with a lower vegetation index tend to have a higher temperature than their rural counterparts. Source : [23].

3.3 The spatio-temporal trade off

Sensors typically trade spatial resolution for temporal resolution and, it has been difficult historically to maximize both. Sensors that have a high spatial resolution often cover a smaller area than a sensor with lower spatial resolution. With a smaller field of view, it takes longer to cover the same area, thus as spatial resolution increases, temporal resolution decreases.

Essentially, it can be said that currently, the TIR data products that are used for developing LST data has either:

- high temporal resolution (sub-daily images) but very low spatial resolution (in the km range).

- High spatial resolution (¡ 100 m) but low temporal resolution (several days up to weeks).

Unfortunately, the applications mentioned before have requirements in both spatial and temporal resolution. The zone of interest, composed of sub-daily frequency and a GSD smaller than 100m, is not covered by any of the existing systems. Private missions, using smaller satellites, can leverage on constellations to provide a higher temporal resolution. However, the spatial resolution is still limited by the payload mass and energy consumption constraints. This trade-off can be described by displaying the resolution of some of the LST/TIR data products available, as in Fig. 3.2.

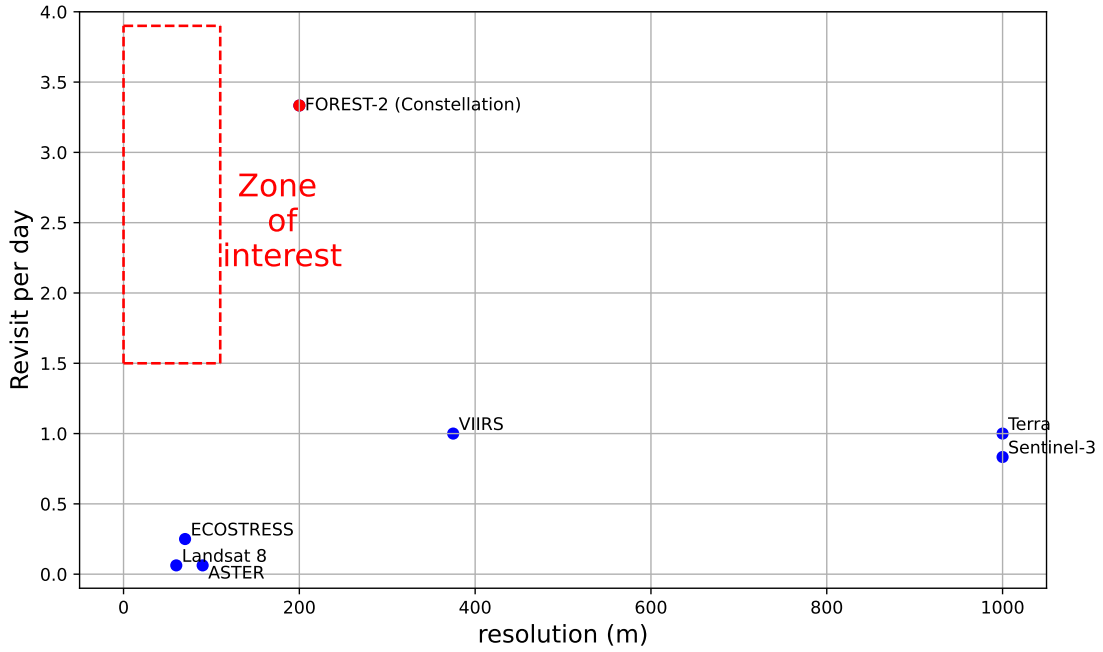

Figure 3.2: Scatter plot of the spatial and temporal resolution of some of the LST/TIR data products available. The trade-off is evident, no mission can provide products in the zone of interest. Constellations may help with temporal resolution, but the spatial resolution is still limited.

This opens the question of whether it is possible to increase the spatial resolution of the data products available using a post-processing technique such as super resolution, whithout compromising the physical consistency of the scenes. The main techniques of super resolution and their most difficult challenges to apply them to LST/TIR data are described in the next section.

4 Super resolution

Super resolution refers to an image processing technique looking to recover a corresponding high-resolution image from a low-resolution version of it, with applications that range from natural images [24], [25] to satellite [26] and medical imaging [27]. SR remains a challenging task in computer vision because it is considered an ill-posed problem: several HR images can generate exactly the same LR image.

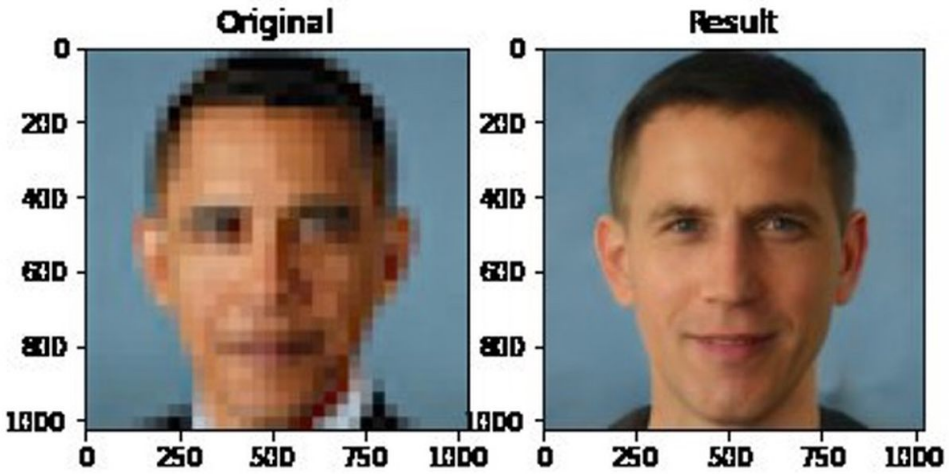

Figure 4.1: Example of super resolution as an ill posed problem. A blurry picture of Barack Obama can be generated from an HR image of another person.

Super resolution was first proposed in the 1960s, while the first use of multiple images dates of 1989. Traditional interpolation-based methods for upsampling images were the first type of algorithms used for super resolution. The most common techniques are nearest-neighbor, bilinear and bicubic interpolation. Nearest-neighbor interpolation is the most straightforward algorithm, as the interpolated value is based on its nearest pixel values. While this method requires almost no calculations, the results are usually blocky because there are no interpolated smooth transitions. Bilinear and bicubic interpolation produces smoother transitions using linear or cubic interpolation in both axes. Bilinear interpolation needs a receptive fields of 2x2 and is usually faster, bicubic needs a receptive field of 4x4. The latter is the most common baseline to quantify the improvement of any super resolution algorithm.

Machine learning was used for the first time in 2000. Deep learning appears as a branch of machine learning, emphasizing the use of multi-layer neural network cascade for feature exctraction and representation. The rise of the technology wave around 2010 changed the way of solving problems in different branches. Instead of piecing together individual feature extraction or functional modules to form a system, the focus is now to optimize parameters by global training after the whole system is designed, what is called end-to-end training.

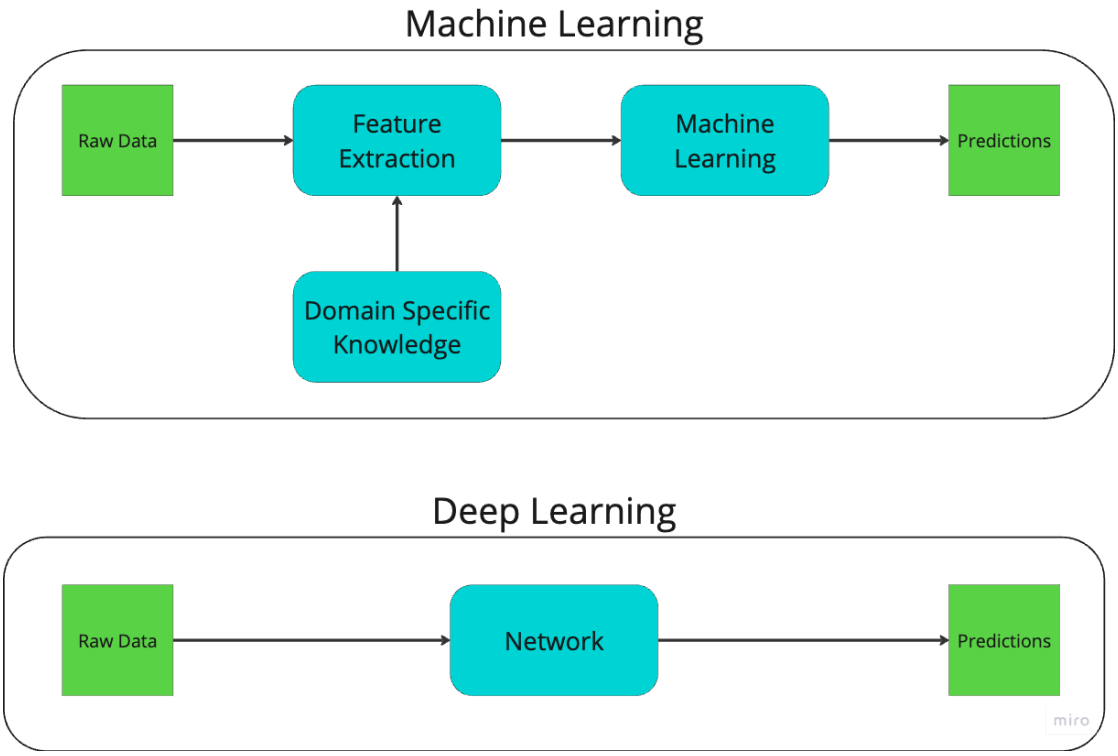

Figure 4.2: In traditional machine learning, the feature extraction step is crucial for performance, requiring a lot of domain knowledge. In deep learning, the feature extraction is learned from the data.

Super resolution using machine or deep learning is a supervised problem, meaning that the super resolved output must be compared to a HR ground truth image. The difference between the two images is used to calculate the objective loss function that the model seeks to minimize. In very few occassions, paired LR-HR images are available.For that reason, the most common approach is to generate the LR images from the HR ground truth using a known degradation model, such as bicubic downsampling + white noise. An example of this method is depicted in Fig. 4.3. The real degradation process is unknown, and is affected by numerous factors such as sensor-induces noise, lossy compresion, speckle noise, motion blur and optical limitations, between others. The disadvantages of using such a simplified degradation process to generate a dataset will be further discussed throughout this work.

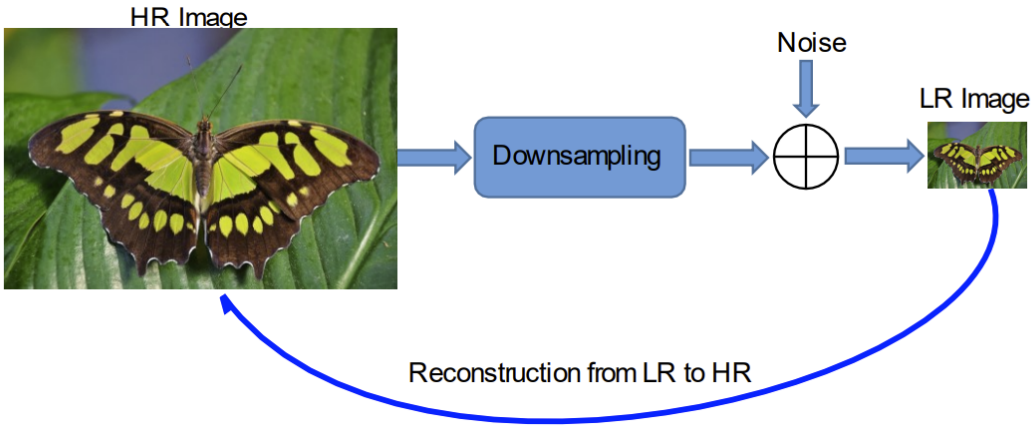

Figure 4.3: Example of generating a super resolution dataset, using a simplified known degradation model. Source: [27].

4.1 Single-Image Super Resolution

In a typical single-image super resolution (SISR) framework, the LR image I^{LR} is modeled as follows:

$$I^{LR} = D(I^{HR}, \Theta) = (I^{HR} * k) \downarrow_s + n \tag{2}$$

Where Θ are the function parameters, $I^{HR} * k$ is the convolution between a blurring kernel k and the unknown HR image I^{HR}, $\downarrow_s$ is the downsampling operator with scaling factor s and n is a noise term. The relationship between the LR and HR images D is known as the degradation model. SISR objective is to solve the inverse equation and obtain I^{HR} from I^{LR}, estimating D^{-1} in the process. As stated before, this an extremely ill-posed problem as D^{-1} is not injective, meaning there are infinite possibilities of I^{HR} for which the equation condition hold.

A variety of deep learning methods were developed over the years to solve the SR problem, all of them are trained using both low and high-resolution images (LR-HR pairs), most of them generated as in Fig. 4.3. The models can be classified based on the upsampling method chosen and it's location, the deep learning network and the loss used for learning.

4.1.1 Upsampling method

The upsampling is essential in deep learning-based SR methods. The most important feature of them is that, as opposed to traditional upsampling methods such as interpolation, they may add new information in the process.

Sub-pixel convolutional layers performs upsampling by generating several additional channels using convolution, and by reshaping these channels, the resolution of the output is upsampled. The layer has a respective wide field that helps learn more contextual information that results in more realistic details, at the cost of possible artifacts.

Deconvolution layers do the inverse of the convolution operation. That means predicting the probable HR image based on the feature maps from the LR image. The

process consists on inserting zeros between the pixels of the LR image, and then applying a convolution operation. The amount of zeros is determined by the scaling factor. This method is widely used in SR methods due to its compatibility with the normal convolution, but may cause uneven overlapping in the generated HR image, resulting in a non-realistic image with decreased performance.

The location of the upsampling layer plays an important role in the architecture. Pre-upsampling SR methods first upsample the LR image and then apply the convolutional layers. The convolutional network task is then to refine the already upsampled image. The biggest drawback is that the dimensions of the image is increzed at the beggining, resulting in higher computational and memory cost than other methods. Post-upsampling SR methods first apply the convolutional layers and then upsample the image. The convolutional network task is then to exctract features in a low-dimensional space. The computational cost is lower than pre-upsampling methods, but the extraction of the low level features for a good reconstruction may be more difficult than refining an already upsampled image. This frameworks may be combined by iteratively up and downsampling the image [28], or by performing progressive upsampling until the desired dimensions are reached [29].

4.1.2 Network design

In the last years, several deep learning network designs have been proposed to solve the SR problem. The ones that are most interesting for this work due to their wide use in the literature are residual learning and attention-based learning.

Residual learning aims to mitigate the vanishing gradient problem that commonly occurs in deep neural network. This is done by adding a skip connection between the input and the output of the network that usually consists in convolutional, batch normalization and non-linear activation layers. This allows the learning of the difference between the input and the output. Mathematically, the residual learning can be formulated as follows:

$$F(x) = H(x) - x \tag{3}$$

Where $H(x)$ is the mapping function of the network and x is the input. If the residual is local, the skip connection is made over a small block of layers. Global residual makes the input and the output of the whole network to be correlated, which is a very desirable property in SR, as the HR image should have significant correlation with the LR image. In this case, the network transform the LR image into an HR image by generating the missing high-frequency details. Attention learning is the idea where certain factors are given more preference.

In channel attention, a particular block is added in the model where global average pooling (GAP) squeezes the input channels; these constants are processed by two fully connected layers to generate channel-wise residuals that define how important is one pixel for each other. In SR, most of the models use local fields for the generation of SR pixels, while in a few cases, some textures or patches which are far apart are necessary for generating accurate local patches. This drives the development of attention blocks that extract non-local representations to add information of pixels that are far away from each other.

4.1.3 Loss functions

As in any supervised learning problem, the selection of the loss function is critical. In SR, they are used for measuring the error in the reconstruction of the HR image.

Initial research employed the loss at the fundamental block of an image, the pixel. The most common loss function is the Mean Squared Error (MSE), which is the average of the squared differences between the predicted and the ground truth images. The MSE is the most common loss function in image processing, but it is not the best choice for SR. This is because the MSE is very sensitive to outliers and tends to generate overly smooth results, as it converges to the mean of the distribution. Thus, researchers often have used L1 or MAE loss. These pixel losses focus on reconstruction fidelity and do not cater for the perceptual quality or textures of the image, resulting in less high-frequency details and overly smooth results. Other were designed to overcome this problem and will be discussed.

If The perceptual quality is an important objective of the SR task, the differences between the generated and groud truth images could be assessed using an image classification network. The distance between the high-level data representation on a determined layer of the network for both images can be calculated in the following way:

$$\mathcal{L}(I^{\mathrm{HR}}, I^{\mathrm{SR}}; N) = \frac{1}{H * W * C} \sum_{i,j,k} (r^l_{i,j,k}(I^{\mathrm{HR}}) - r^l_{i,j,k}(I^{\mathrm{SR}}))^2 \tag{4}$$

Where r^l is the output of the k-th channel of the l-th layer of the pre-trained classification network N when the input is I^{HR} or I^{SR}. H, W and C are the dimensions of the layer output (height, width and channels). Commonly used classification networks are VGG [30] or ResNet [31]. The purpose of the content los is to compare the information about image features from the network. This ensures the visual similarity between the original and generated image by comparing content and not individual pixels. Thus, content loss functions helps producing visually perceptible and more realistic loooking images and are widely used in SR [32, 33]. On the other hand, this type of loss may not focus on the physical consistency of the image, resulting in possible artifacts that may look realistic but are non-existant. This is one of the main reasons why the content loss is not used in remote sensing applications.

The adversarial loss is based on the generative adversarial network (GAN) [34]. The GAN is composed of two networks, a generator and a discriminator. The generator is trained to generate SR images that are indistinguishable from the real HR images, while the discriminator is trained to distinguish between the generated and real images. Training is performed in sequencial steps, where the generator is adjusted for better results that may fool the discriminator, and then the discriminator is adjusted to better distinguish between the generated and real images. When the generator is able to create outputs that conform to the distribution of the actual data, the discriminator is no longer able to distinguish between the generated and real images. In many cases, the mean squared error is used due to improved results:

$$\begin{aligned} \mathcal{L}_{GANg}(I^{SR}; D) &= (D(I^{SR}) - 1)^2 \\ \mathcal{L}_{GANd}(I^{HR}, I^{SR}; D) &= (D(I^{SR}))^2 + (D(I^{HR}) - 1)^2 \end{aligned} \tag{5}$$

Where D is the discriminator network. Results show that although the adversarial loss yielded lower physical consistency metrics, content and perceptual metrics were improved. The use of the discriminator was able to regenerate intrincate patterns that were very difficult to learn using ordinary deep learning methods. This is because the pixel-loss-based solutions perform a pixel-wise aggregation of the posible solutions in the pixel space, while adversarial loss drives the reconstruction towards the natural image manifold, producing more perceptually convincing solutions. The main drawbacks of the adversarial loss are the inherent instability in the training of GANs and the probable degradation in physical consistency metrics. The latter is the main reason why this type of loss will not be used throughout this work.

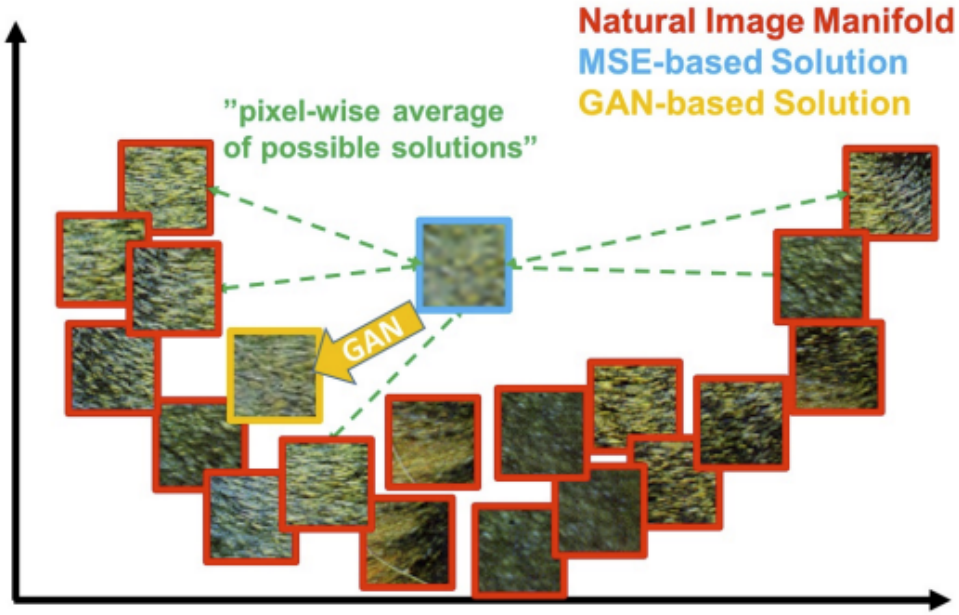

Figure 4.4: Illustration of patches from the natural image manifold and results coming from MSE pixel-loss (red) and GANs (orange). source:[32]

4.2 Multi-Image Super Resolution

Multi-Image Super-Resolution (MISR) is the task of yielding HR images by fusing multiple LR observations of the same scene, which allows the achievement of higher reconstruction accuracy than relying on only one image. The development of this approach progressed at a slower pace due to the extesive pre-processing requirements imposed on the input, as this algorithms have a high sensibility to the input variability and their proper co-registration.

When the input images are of the same nature, but taken at different points in the temporal dimension, the problem is often called multi-image super resolution. On the other hand, when the images are taken at the same time but they come from different sensors and show different spectral bands, it is called multi-spectral super resolution, which will be further discussed.

The main problem in MISR is the difficulty to generate a dataset with multiple images of the same scene, and it is the main reason why SISR is more popular. In 2019, the European Space Agency (ESA) organized an SR challenge [35] based on real-world scenes acquired by the PROBA-V satellite, each of which contains an HR image (100m GSD) coupled with at least nine LR images that are not perfectly co-registered and they may be taken months apart. This challenge, with a not-synthetically generated HR-LR image pairs, fostered a new generation of model architectures that are able to fuse the multiple LR images to create better reconstructions [36, 37]. Both of the cited networks were tested in synthetically generated datasets throuhout this work and showed better performance than SISR networks, but they were discarded because of the impossibility to have a multi-image dataset using real FOREST-2 images.

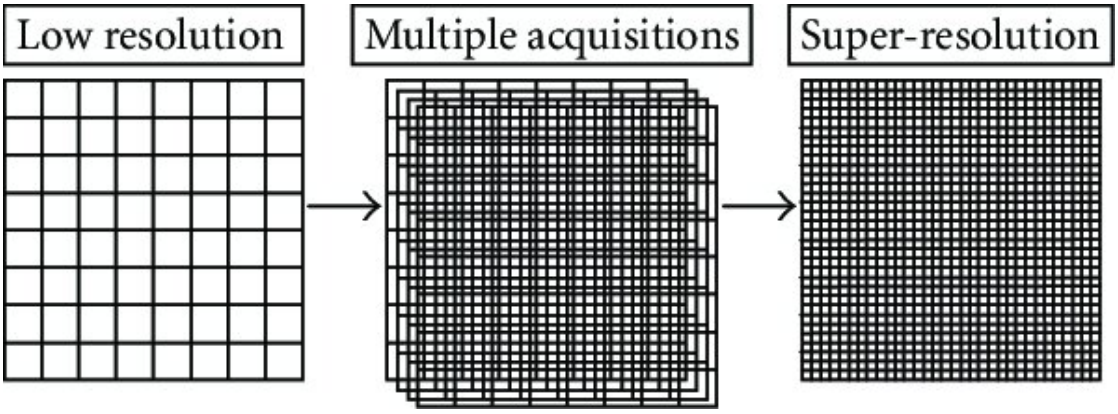

Figure 4.5: Multi-image super resolution algorithms combine multiple low-resolution image acquisitions into a high-resolution image. Source: [38]

4.2.1 Multi-spectral super resolution

Also Referred to as "hyper-spectral super resolution" in the literature, The term "Multi-Spectral" emphasizes the use of multiple spectral bands, in contrast with the multi-image approach detailed previously. While the concept bears similarities to MISR, the key distinction lies in MSSR's use of a single scene captured with different spectral bands, as opposed to multiple images, to reconstruct a superior, super-resolved image.

In the context of MSSR, each spectral band, corresponding to a specific wavelenght range, provides unique information about the observed scene. Some of the spectral bands yield better resolution because of their physical properties and the costs related to their sensors. Using this higher resolution bands to increase the detail in the lower bands seems like a resaonable approach.

Traditional pan-sharpening algorithms could be considered as deterministic MSSR algorithms. They are usually used to increase the resolution of a multi-spectral RGB image using the panchromatic band. The overlap between the wavelengths of the bands makes this algorithm straightforward and useful. However, it is ill-suited for Thermal Infrared (TIR) data due to the disjointed spectral domains of the visible and TIR bands. The result of pansharpening TIR data is shown in Fig. 4.6 While the general resolution of the image is improved, several TIR hotspots are darkened and highlights from the visible bands are translated to the super-resolved image. This is particularly problematic

for clouds, which have an inverse spectral response in the TIR and RGB bands.

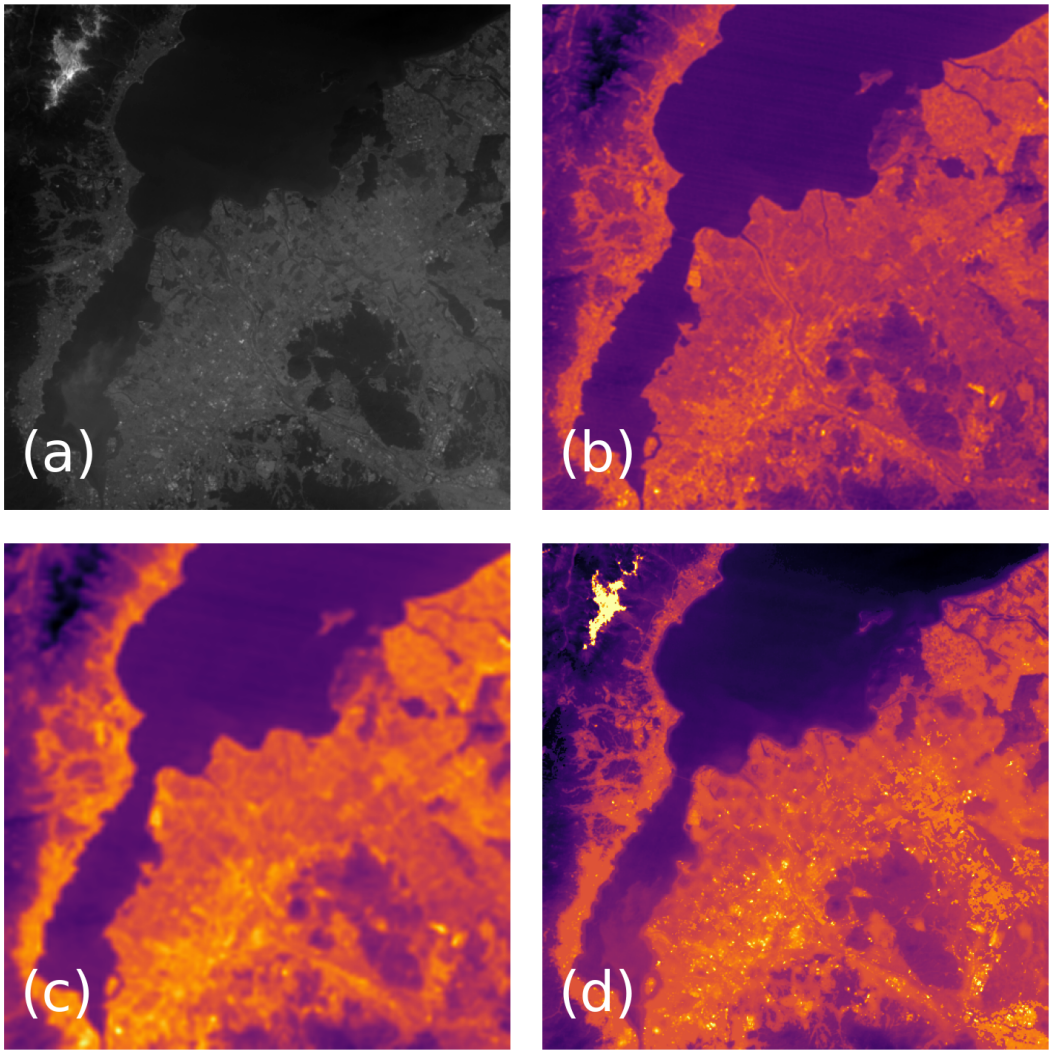

Figure 4.6: Example of Pan-sharpening on TIR data using a panchromatic band. (a) Panchromatic band, (b) HR TIR image, (c) Downsampled version of the TIR image , (d) Pansharpened image. The pan-sharpened image is less blurry than the LR, but a lot of artifacts are produced, specially in clouds. Source: [39]

In [39], A deep learning MSSR network is trained assuming the presence of common information between low-resolution LWIR images and their higher resolution RGB counterparts, with the objective of creating a super-resolved product in the LWIR band by an effective fusion. This improved image retains the essential thermal information, while simultaneously incorporates enhanced spatial resolution details captured from the visible bands. MSSR remains a more promising alternative than MISR because it doesn't have the pre-processing burden that the latter has, as the images are well co-registered in the spatial and temporal domain. Additionally, most satellites have multi-spectral sensors, making the dataset generation much easier.

4.3 The domain gap problem

SR is a supervised problem, the super resolved image is compared to the HR ground truth and the differences (pixel-by-pixel or perceptual) drive the gradients of the neural network to minimize the loss, in a fully supervised manner. The objective of this work is to increase the resolution of FOREST-2 images, but a high resolution version of FOREST-2 is not available. The only alternative is to use scenes from other missions that have a higher resolution.

Most of the research in the field of SR is conducted by artificially producing HR-LR pairs by downscaling the HR images with known kernels, as in Fig. 4.3. However, this is rarely the case when using "non-ideal", real world images. In spite of their success on synthetic datasets, the poor generalization capacity of the trained SR networks limits their application in real scenarios, leading to blurry images and strange artifacts in the SR results [40].

The domain gap problem occurs when there are systematic discrepancies between data use for training and the real-world data. This is described in Fig. 4.7, where the HR image is processed through different known degradations. If an SR model is trained using the left-most degradation, it will produce undesirable results if LR images generated by the other degradations are used as input. In this example, the left-most degradation seems to have better resolution and less noise than the rest. This will lead to noisier and blurrier results when using the other degradations as input.

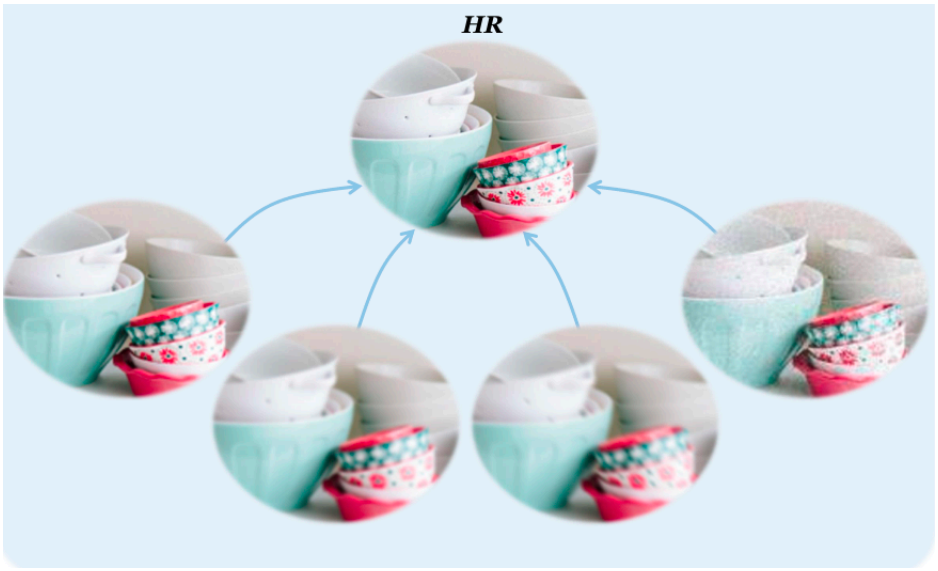

Figure 4.7: Effects of different degradation models on one HR image. Source: [41]

4.4 Blind image Super Resolution

The problem of SR with an unknown degradation process is known as blind SR. Growing attention has been paid to blind SR in recent years, towards filling the domain gap presented in 4.3. A schematic diagram of the problem is shown in Fig. 4.8. Non-blind SR methods assume that the degradation process is known, and maps the bicubic downsampled LR image to the natural HR image space. However, an arbitrary LR

input image, as a scene captured by a satellite, is usually degraded by an unknown process, which is difficult to be modelled explicitly. The arbitrary LR input is not in the same domain as the bicubic downsamopled LR image, and thus the non-blind SR methods are not successful in the super resolution process. There will be a large domain gap between the SR output and the desired image samples from the target natural HR domain, leading to a poor-quality result. Blind SR methods, on the other hand, aim to learn the degradation process from the training data, and map the arbitrary LR input image to the natural HR image space.

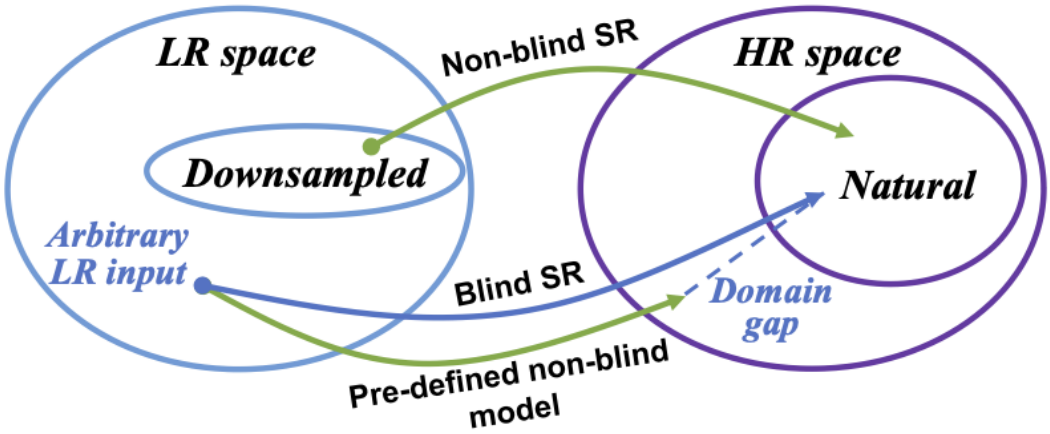

Figure 4.8: Domain interpretation of differences between non-blind and blind SR. Source: [41]

In the literature, two main approaches exist to bridge the gap: Explicit modelling based on an extension of eq. 2 and implicit modelling thought distribution learning of the degradation process. Explicit modelling can be further classified into two sub-categories according to whether they employ external datasets or rely on a single input image to solve the SR problem.

4.4.1 Explicit modelling with external dataset

This kind of methods use an external dataset to train an SR model well adapted to variant blurring kernels and noises. Typically, a traditional SISR is employed and an estimation of the kernel and the noise is used as a conditional input along with the LR image. After the training process, the model will be able to produce good results only in the now bigger pool of degradation types covered in the training dataset. According to whether the degradation is estimated or given, this approach can be further classified into two sub-categories.

Explicit modelling without kernel estimation aims to directly concatenate a pre-defined degradation map to the LR input, as depicted in Fig 4.9. This allows feature adaptation according to the specific degradation model and helps to cover multiple degradation types during training. The PCA technique used to project the degradation map can be replaced with a shallow neural network that may learn a kernel mapping that better fits the specific SR model used.

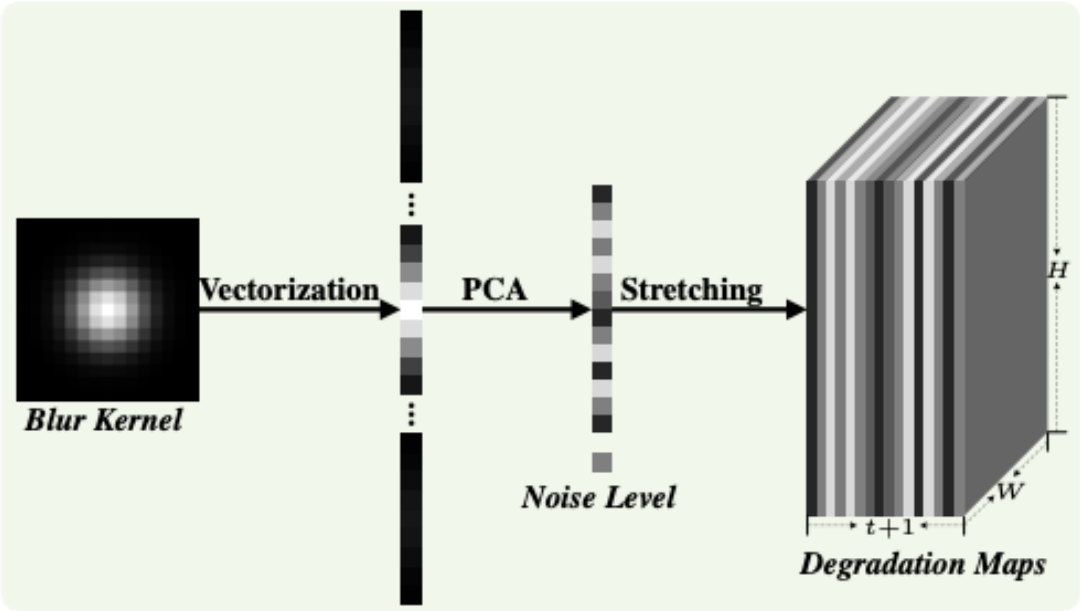

Figure 4.9: Dimensionality stretching strategy to concatenate the degradation map to the LR input. The vectorized kernel is projected onto a space of a lower dimensionality, and then stretched to generate t feature maps with the same shape of the input image. The noise level is also concatenated. Source: [42]

The biggest drawback of this approach is that it relies on an additional input of degradation estimation, specially the kernel. Howerver, estimating the correct kernel from an arbitrary LR image is not easy and kernel mismatch will result in a dramatic loss in SR performance. This method remains feasible only when a way of obtaining a reliable degradation estimation is available. Otherwise, a manual process to find the best input for better result is needed.

Explicit modelling with kernel estimation aims to estimate the kernel from the LR input image in an iterative way until a good enough result is obtained [43]. The main idea is to take advantage of intermediate SR results because some of the artifacts caused by kernel mismatch show regular patterns that a corrector network can use to perfrom kernel correction. Methods like [44], enhance the approach by unifying the kernel correction and SR network into an end-to-end trainable network. However, the iterative nature of this method leads to higher inference time. Additionally, the optimal number of iterations is not known and must be determined empirically.

Other approaches propose to learn a bling SR model by merely covering more degradations with more realistic kernels in the training dataset, creating a large pool. Kernels from this pool are used to synthetize the training pairs in a non-blind setting. The more general training dataset enables the SR model to adapt to real input images. However, it is very hard to cover all the possible degradation types in the real world, and the model will fail when facing a new degradation type.

4.4.2 Explicit modelling with single image

SR modelling with a single image is based on internal statistics of natural images: patches of a single image tend to recur within and across different scales of the image [45]. This characteristic is very powerful, since it is image-specific and unsupervised. It was used first in 2009 in a method that does not used deep learning [46], and gained traction with KernelGAN [47]. KernelGAN interprets the maximization of patch recurrence as a data distribution learning problem, assuming that the downsampled version of an LR image generated by the optimal kernel should share the same patch distribution with the original LR input. Using a GAN framework, a deep linear network is used as a generator to parametrize the underlying SR kernel, and a discriminator distinguishes generated patches from those of the original LR image. Once training finishes, the output of the generator is an estimation of the blurring kernel of the input image.

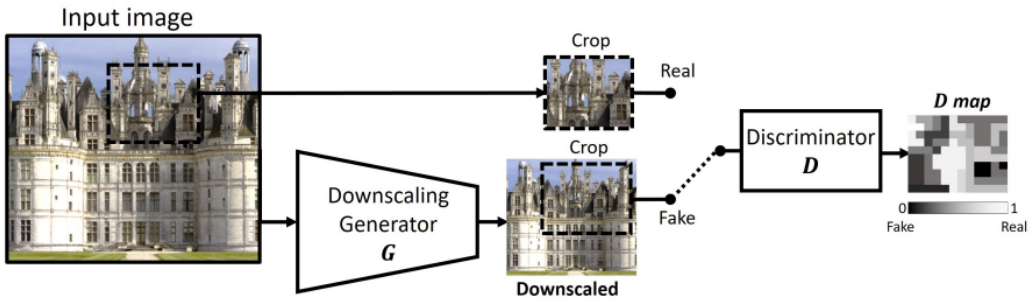

Figure 4.10: KernelGAN schematic diagram. The discriminator tries to distinguish between the generated patches and the original LR image patches. G learns perform 2x downscaling while fooling the discriminator by maintaining the same distribution of patches. Source [47]

This idea of self-supervision based on patch recurrence is also applied to perform SR without pre-training, as in zero-shot super resolution (ZSSR) [48]. In this case, the training is conducted using HR-LR pairs generated from a single LR image. The original input is regarded as HR and downsampled versions of it as LR using a kernel. The network trained on these image pairs will be capable of inferring relationships across different scales which is then used to super-resolve the input. ZSSR is still not fully blind, it requires an estimated blur kernel as input. For that reason, a joint framework that combines ZSSR and KernelGAN yields very good results. For a given image, KernelGAN estimates the blurring kernel that is then used in ZSSR to perform super resolution.

While the idea of self-supervision is very flexible and efficient, its basic assumption may fail in certain cases. Hence, this approach can only produce favourable SR outputs for a limited set of images that have recurring contents across scales.

4.4.3 Implicit modelling

Implicit modelling aims to grasp the underlying degradation model through learning from an external dataset. On paired HR-LR images, the SR model is already enough. However, these datasets are rarely available in real-world scenarios. Usually the data

available is unpaired, meaning that HR images and LR images with realistic degradations are available, but there is no correspondence between them. Existing approaches exploit the data distribution learning ability of GANs, where discriminators are used to distinguish between the generated images from the real ones, pushing the generator towards an appropiate direction.

First attempts for implicit modelling were based on CycleGANs [49], that consists of two generators and two discriminators that move from domain A to B and viceversa. The cycle consistency loss is based on the principle that after a round-trip transformation, the original image should be recovered. In CinCGAN [50], the HR input is transformed using bicubic downsampling before doing SR with a pre-trained network and is regarded as the clean LR domain. Two CycleGAN structures are applied to transform the LR input to the clean LR domain and to the HR domain. This way, no paired data is necessary.

Another way of performing implicit modelling is using a single GAN to learn the degradation process from HR to LR, and generate a supervised paired dataset that can be used for training the SR network. The generator simulates the degradation from the HR domain to the LR domain and the discriminator distinguishes between the generated LR images and the real LR images. In these methods, such as [51, 52], usually the discriminator architecture is focused to distinguish the images using the high-frequency contents of them, due to the fact that degradations usually have a big overlap at lower frequencies. To further reduce the domain gap, several extensions of the method are proposed. In [53], both the generated and real LR images are used to train the SR model. The super resolved version of the generated LR images can be compared with the original HR input using a pixel-wise loss, and the super resolved real LR images can be used for training trough a discriminator that distinguishes between them and the HR images.

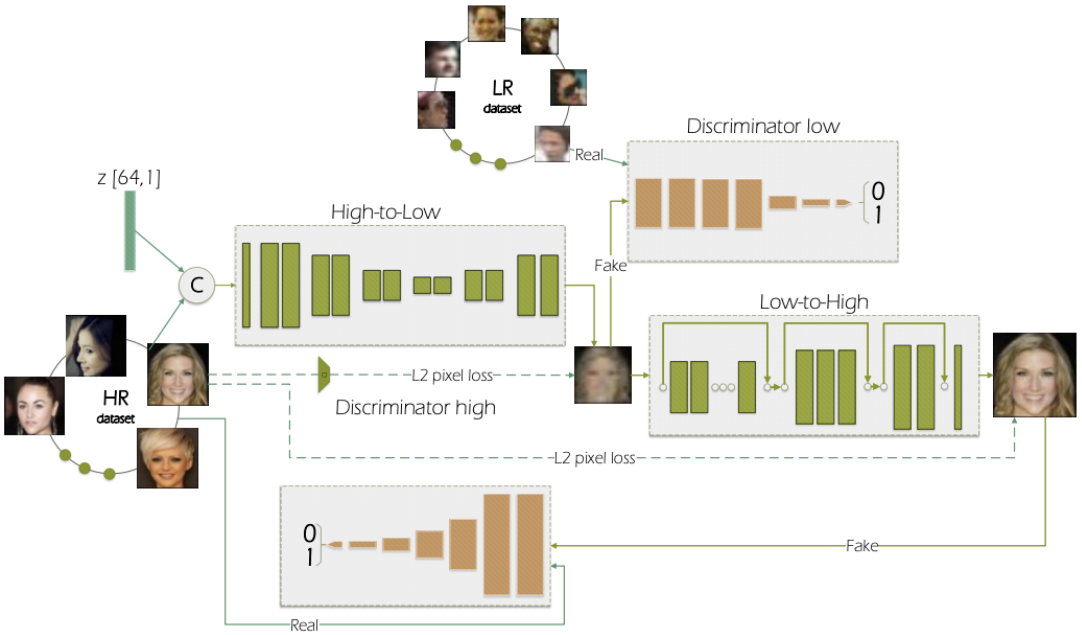

Figure 4.11: Degradation GAN schematic diagram. The architecture includes one LR generator, on SR network and two discriminators. Source: [52].

While very flexible, limitations of implicit modelling are the dependency on huge dataset that is just not possible in some applications. Additionally, several artifacts may be produced in the SR results due to the difficulty and instability of GANs training. The choice of the generator in the case of degradation learning GAN is also very important, if it is not well constrained enough, it will produce unrealistic results that will misguide the SR network and lead to poor results even after long training sessions.

5 Methodology

5.1 Models Architecture

5.1.1 Probabilistic degradation model

This architecture, proposed in [51], belongs to the implicit modelling blind super resolution methods. Most methods in this category try to adaptively learn the degradation process via a deterministic model in order to avoid the domain gap between synthetic and test image. However, some degradations in real scenarios are stochastic and cannot be determined by the content of the image. These approach may fail to model the random factors and content-independent parts of degradations, which will limit the performance of the following SR models. In this architecture, the degradation D is assumed as a random variable, and a network learns its distribution by modeling the mapping from a priori random variable z to D. Compared with previous deterministic degradation models, PDM could model more diverse degradations and generate HR-LR pairs that may better cover the various degradations of test images, and preventing the SR model from over-fitting to specific ones.

Other degradation-learning-based SR methods have an assumption: the degradation is completely dependent on the content of the image. However, this may not hold in most cases. Some degradations are content independent and stochastic, such as random noises. These random factors and content-independent parts of degradations could not be well modeled by these deterministic models. A better assumption is that the degradation is subject to a distribution, which may be better modeled by a probabilistic model.

The degradation is parametrized with two random variables, i.e., the blur kernel k and random noise $\mathbf{n}$, by formulating the degradation process as the linear function from Eq. 2. It can be divided into two linear steps [54]:

$$\begin{aligned} I_{\text{clean}}^{\text{LR}} &= (I^{\text{HR}} * k) \downarrow_s \\ \mathbf{I}^{\text{LR}} &= I_{\text{clean}}^{\text{LR}} + \mathbf{n} \end{aligned} \tag{6}$$

Usually,the two steps are mutually independent, as the blur kernels are mainly dependent on the properties of the camera lens while the noises are mainly related to the properties of sensors. Thus, the distribution of the degradation process can be represented as the product of the distribution of k and n, which can be modeled by learning the mapping from a priori random variable z to k and n.

$$p_D(D) = p_{k,n}(k, n) = p_k(k)p_n(n). \tag{7}$$

To model the distribution of the blur kernel k, we define a priori random variable z_k which is subject to multi-dimensional normal distribution. Then we use a generative module to learn the mapping from z_k to k:

$$k = \text{net}K(z_k), \quad z_k \sim \mathcal{N}(0, 1), \tag{8}$$

The spatially variant blur kernel is considered first. This implies that the blur kernel for each pixel of the image is different. In that case, we have

$$z_k \in \mathbb{R}^{f_k \times h \times w}, \quad k \in \mathbb{R}^{(k \times k) \times h \times w}, \tag{9}$$

where f_k is the dimension of the normal distribution z_k, k is the size of the blur kernel, h and w are the height and width of the image, respectively. Generally, the sizes of the convolutional weights are set as 3×3, which indicates that the learned blur kernels are spatially correlated. Otherwise, if the spatial size of all convolutional weights is set as 1×1, the blur kernel could be approximated by a spatially invariant one, which is a special case of the spatially variant blur kernel with h = w = 1. This approximation simplifies the dimensions of the problem drastically and is an appropiate assumption if the crops used for training the model are small enough. A Softmax layer is added at the end of the network to guarantee that all elements of k sum to one.

To model the distribution of the noise n, a vanilla generative module can also be used:

$$k = \mathrm{net}N(z_n), \quad z_n \sim \mathcal{N}(0, 1), \tag{10}$$

$$z_n \in \mathbb{R}^{f_n \times h \times w}, \quad n \in \mathbb{R}^{h \times w \times c}, \tag{11}$$

Where the height, width and number of channels of the image is noted as h, w and c respectively. In this work, c is always set to 1.

In other methods [55], the noise is modeled as a combination of shot and read noise. It can be approximated as a heteroscedastic Gaussian distribution, which is dependent on the content of the image.

$$n \sim \mathcal{N}(0, \sigma_{\mathrm{read}} + \sigma_{\mathrm{shot}} \cdot I^{\mathrm{LR}}_{\mathrm{clean}}), \tag{12}$$

This indicates that the noise is also related to the image content and the distribution of n should be expressed as:

$$k = \mathrm{net}N(z_n, I^{\mathrm{LR}}_{\mathrm{clean}}), \quad z_n \sim \mathcal{N}(0, 1), \tag{13}$$

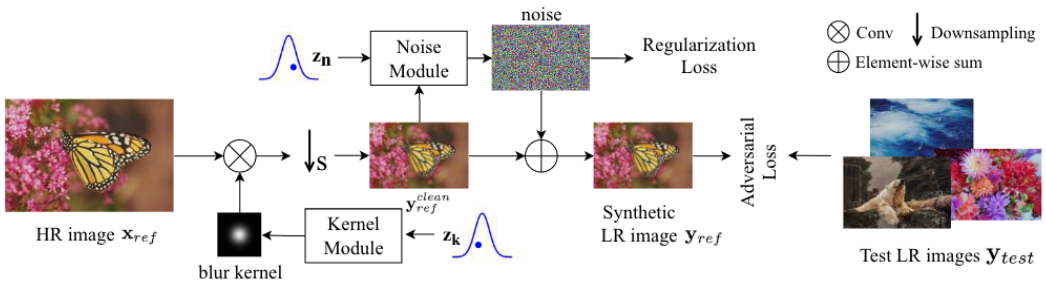

Figure 5.1: Schematic of the probabilistic degradation module. The discriminator is left out for a more intuitive description.

The probabilistic degradation model is optimized via adversarial training, which encourages the output of the generator to be similar with the test images [52]. To avoid overly noisy images, a constraint to the noise level is added to the loss function via a regularization term. A multiplication constant is also added to balance the magnitude of the two terms.

$$l_{\text{total}} = l_{\text{adversarial}} + 100\|\mathbf{n}\|_2^2. \tag{14}$$

This approach formulates the degradation process as a linear function, and the learned degradations can only impose a limited influence on the image content. In this way, it better decouples the degradations with image content and allows to focus on learning the degradations. This limitation eliminates the need of a guidance using a bicubically downscaled version of the HR image, as opposed to [53] or [52]. This guidance may be inappropriate, especially when the test images are heavily blurred.

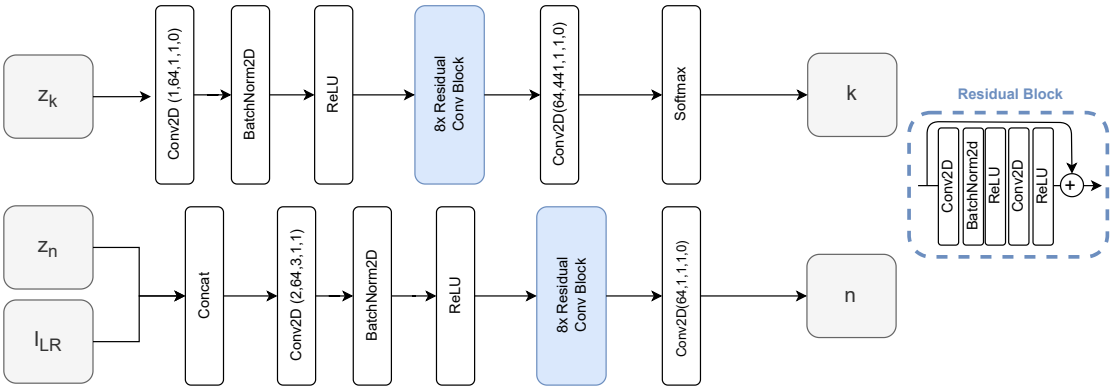

Figure 5.2: Schematic of the generative networks used in the kernel and noise module of the probabilistic degradation model. The parameters of the convolutional layers represent input channels, output channels, kernel size, stride and padding, respectively. The residual blocks use the same kernel size as the convolutional layers of each module. In the noise module, the random vector z_n is concatenated with $I_{\text{clean}}^{\text{LR}}$ before the first convolutional layer.

To discriminate the generated images from the test images, a PatchGAN discriminator is used [56]. This architecture assesses the structure of local image patches, allowing it to focus on high-frequency details of the image. This is particularly useful in the context of this work, where the generated images are expected to share a lot of information in the low frequencies of the domain and the differences with the test images are expected to be in the high frequencies. The architecture of the discriminator is shown in Fig. 5.3. The network tries to classify if each N $\times$ N patch in an image is real or fake. The size of the patches depend mostly on the number of convolutional layers with stride 2 that are employed. The discriminator outputs a matrix of values, where each value represents the probability that the corresponding patch is real. The final output is obtained by averaging the values in the matrix. The PatchGAN discriminator is trained to minimize a classification loss.

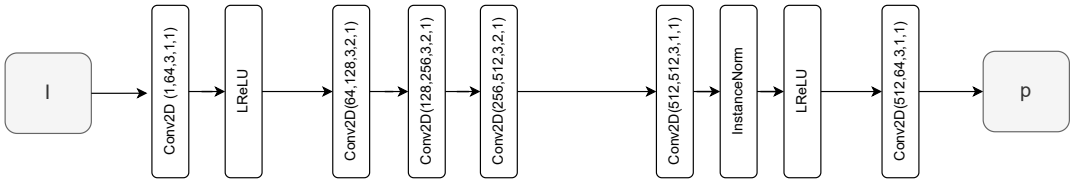

Figure 5.3: Diagram of the PatchGAN discriminator. The parameters of the convolutional layers represent input channels, output channels, kernel size, stride and padding, respectively.

The constrained nature of the probabilistic degradation model allows the possibility to train it simoultanously with a super resolution algorithm, as described in Fig. 5.4. In this way, PDM can be integrated with any SR model to form a unified framework for blind SR that can be trained in an end-to-end fashion, allowing for faster iterations. Other methods [57] [53] require that the training of the degradation model and the SR model in separate phases. They firstly train a degradation model and then use the trained degradation model to generate pairs and train the SR model. This two-step training method is time-consuming but necessary because the highly nonlinear degradation models used will produce undesirable results at the beginning of the training, which may mislead the optimization of the SR model.

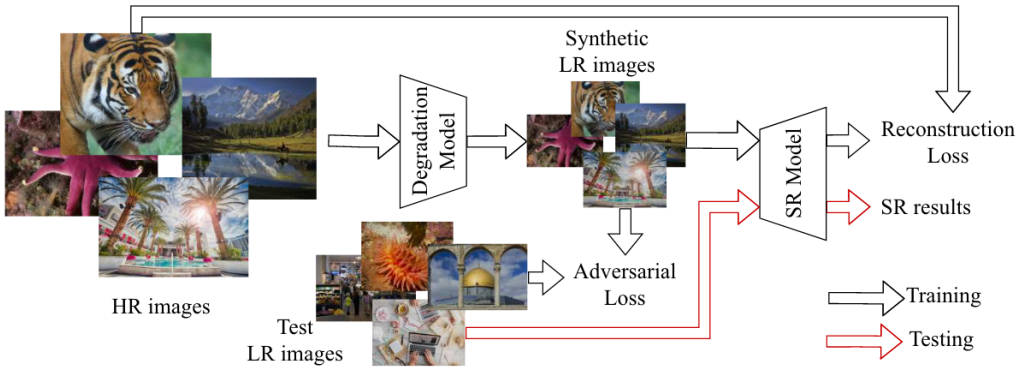

Figure 5.4: The probabilistic degradation model is used to encourage the degradation model to produce images in the same domain as the test LR images. After training, the SR model is directly used to super resolve the inputs. Source [51].

This creates a very flexible framework for blind super resolution. The only data requirement is a big enough unpaired dataset of LR and HR images and an optional smaller paired dataset for validation and early stopping of the model training. The biggest limitation of this approach is the one that degradation-learning-based methods have: the HR images (source domain) and the LR images (target domain) must be well defined and the model will not generalize a domain that is not the LR images. Using this framework for general images is difficult, due to the variety of cameras and sensors, compression algorithms that play a part in the image distribution. However, this work is focused on 2 specific missions with a well defined degradation process, so this limitation is not a problem.

5.1.2 SRResNet

To perform the super resolution task, the SRResNet architecture is used. This network has been used extensively in the literature and has proven to be a good baseline for SR. It can also be easily extended for multi-spectral super resolution, as in [39]. It uses a residual network architecture,in a post-upsampling framework based on sub-pixel convolution layers and a pixel based loss function. Introduced in 2017 [32], SRResnet is the generator in the SRGAN architecture, which is a GAN-based super resolution method. The purpose of the GAN is to drive the reconstruction process towards the natural image manifold, producing more visually convincing solution. Additionally, a perceptual loss based on activation layers of a pre-trained VGG network [30] is incorporated to the training objective. As this work focuses on having super resolved images with high physical consistency and not on the perceptual superiority of the images, these two components are left out, and only the generator is going to be used. The architecture, with slight modifications, is detailed in Fig. 5.5.

First, features are extracted from the input using a convolutional layer with 64 filters and kernel size of 3, plus a ParametricReLU activation function. Before going through the core of the network, the feature map is reduced by half using a 1x1 convolutional layer. The core of the network is composed of 5 residual blocks, consisting of two convolutional layers, followed by batch-normalization layers and ParametricReLU activation functions. The convolutional layers have 3x3 kernels and 64 feature maps. To increase the resolution of the input image, two trained sub-pixel convolution layers are used.

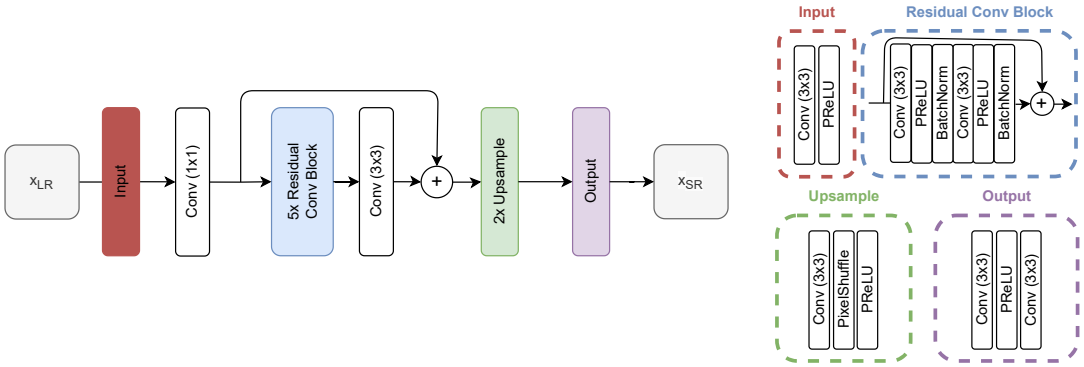

Figure 5.5: Modified SRResNet architecture. X_{LR} represents the low resolution input image, X_{SR} the super resolved image, which is then compared to the ground truth X_{HR}.

5.2 Baseline Degradation model

As stated in 4.3, early super-resolution methods commonly generated high-resolution (HR) to low-resolution (LR) samples using predefined degradation techniques, with bicubic downsampling being the most used setting [42]. This kind of synthetic data, while easy to obtain, often results in a domain gap problem, where the data used for training and assessing the model do not come from the same distribution as real data. This gap usually leads to performance drops when the models implemented in production environments. A possible solution is to synthesize samples with a stochastic degrada-

tion model, which includes a set of multiple blurring kernels and several random noises configurations that convert scenes from an HR mission into LR versions, as if they were taken using FOREST-2. The larger degradation space grant these models better generalization capabilities and let experts be part of the kernel definition process, based on prior knowledge of the degradation process. Unfortunately, the variety of predefined degradation's is still limited and still fail in most applications.

A degradation model like this one will be used as a baseline for this work.

5.2.1 Blurring Kernel

In the literature, the kernel of the degradation process is usually modelled as a fixed isotropic gaussian kernel, with a parameter σ that depends on the scale factor of the super resolution process. To provide more variability to it to each dataset pair, the parameters of the blurring kernel that determine it's width in both axis, σ_x and σ_y, are sampled from a normal distribution with a determined mean and standard deviation. Fig. 5.6 shows some examples of kernels generated using this method, in the upper row, both distributions have a similar mean and distribution, resulting fairly isotropic kernels. In the lower row, the mean of the x axis is much higher than the variance of the x axis, resulting in highly anisotropic kernels. The effects of these kernels on the HR-LR generation are shown in Fig. 5.7.

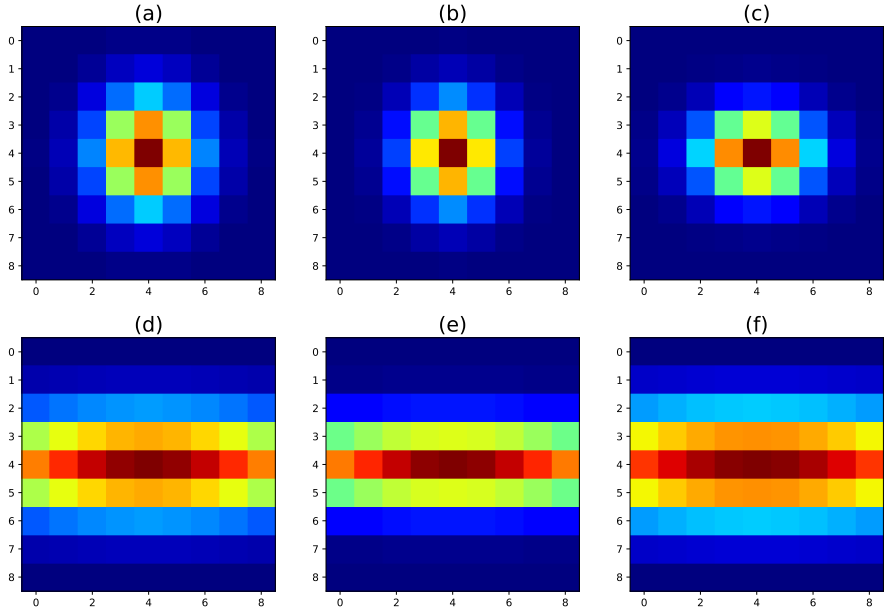

Figure 5.6: Example of kernels used in a stochastic degradation model. (a),(b) and (c) are generated using a symmetric variance on the x and y axis. (d) (e) and (f) are generated using an asymmetric variances, resulting in much more anisotropic kernels.

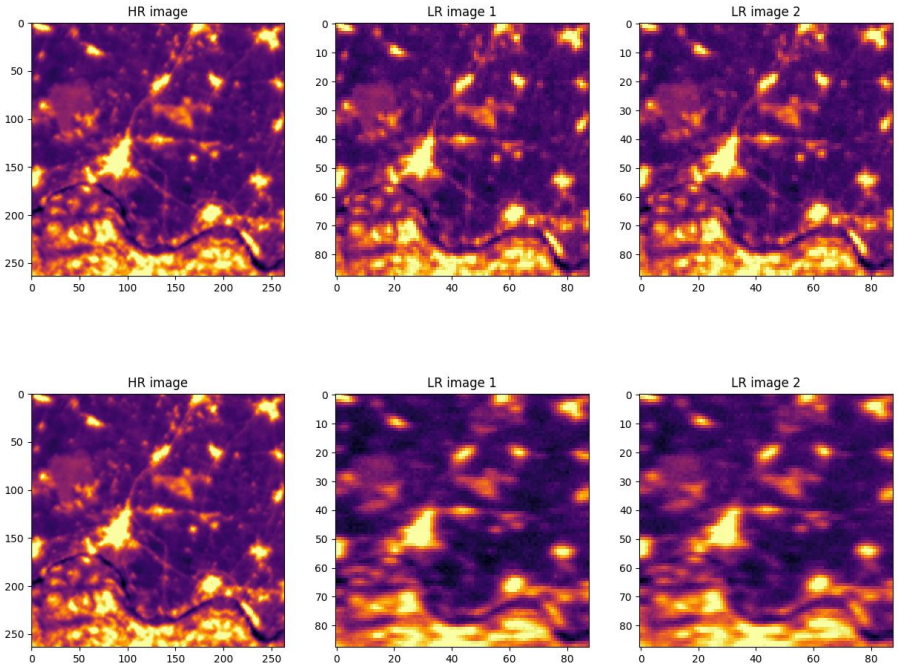

Figure 5.7: Effects of different blurring kernels on the HR-LR generation. The upper row contains images generated using blurring kernels with symmetric distributions. The lower rows contains images generated using asymmetric distributions for the variances, resulting in highly anisotropic kernels.

5.2.2 Radiometric error correction

FOREST-2 radiometric accuracy is 1K at 300K. Other missions report lower nominal radiometric accuracies, such as the case of ECOSTRESS instrument sheet [58] which is 0.5K at 300K. This difference in accuracy should be taken into account. To align these accuracies, we first calculate the additional error required using the following equation:

$$e_{\text{forest}} = \sqrt{e_{\text{eco}}^2 + e_{\text{extra}}^2} \tag{15}$$

where e_{eco} is the ECOSTRESS error, and e_{extra} is the additional error required for FOREST-2.

Using the above equation, we find that an additional radiometric error of approximately 0.8660K is needed. The next step involves converting this extra error into a radiance value. This requires calculating the derivative of the Planck equation at 300K, which is done numerically as follows:

$$\frac{\partial B}{\partial T} = \frac{B(\lambda, T + \delta T) - B(\lambda, T)}{\delta T} \tag{16}$$

By multiplying the results of equations 15 and 16, we can obtain the radiance error for both FOREST LWIR bands. The additional radiance errors for LWIR1 and LWIR2 bands are found to be 1.5472×10^{-1} W/sr/m^2/μm and 1.1444×10^{-1} W/sr/m^2/μm, respectively.

The difference in radiances will be split into two components. On one side, the cold Bias represents a systematic error in the measurement, this error acknowledges discrepancies that can be attributed to sensor calibration and atmospheric conditions. On the other side, the random noise accounts for unpredictable fluctuations in the measurement process. It could be due to a variety of sources like electronic noise in the sensor, random atmospheric disturbances, or other stochastic factors. As the extent of each component is not known and to give more variability to this basic degradation model, a random factor $\phi \in [0, 1]$ is introduced so that:

$$\begin{aligned} \varepsilon_{\text{final}} &= (1 - \phi) \times \varepsilon_{\text{radiance}} + \phi \times \eta \times \varepsilon_{\text{radiance}} \\ \eta &\sim \mathcal{N}(0, 1) \end{aligned} \tag{17}$$

The effects of the error correction is shown in Fig. 5.8. As the target radiometric error increases with respect to ECOSTRESS scenes, the loss of information is more noticeable.

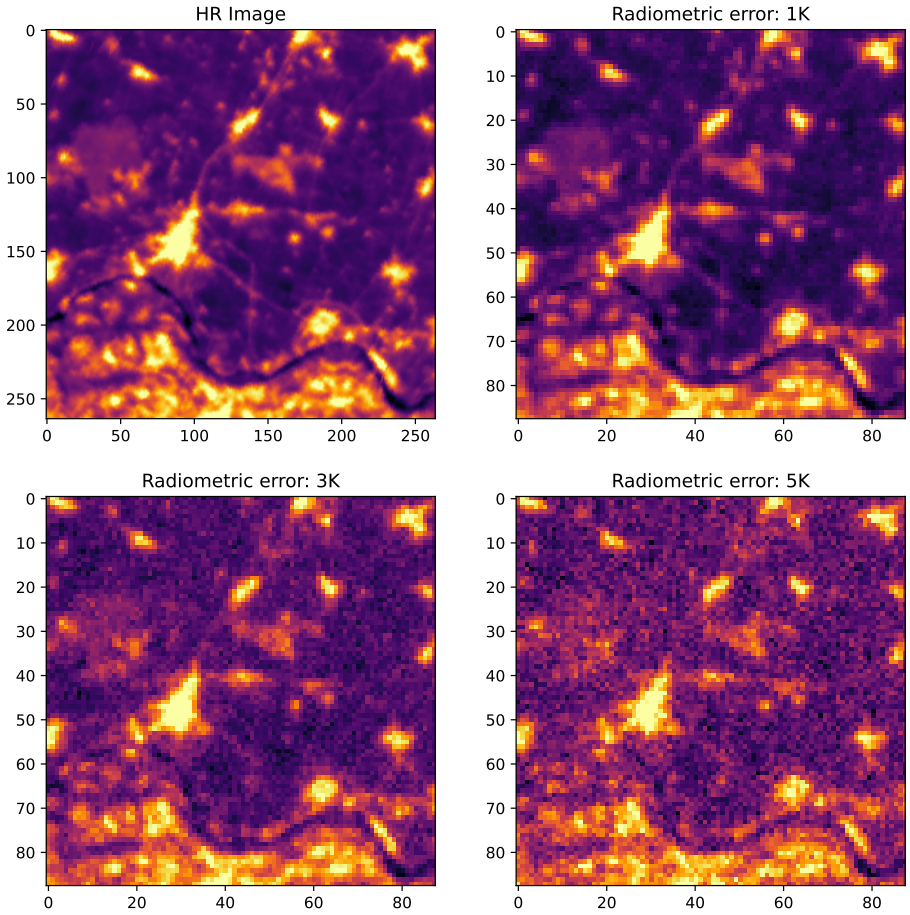

Figure 5.8: Effects of increasing radiometric error on the HR-LR generation.

5.3 Signal-to-Noise Ratio (SNR)

To quantify how much a signal is corrupted by noise, the Signal-to-Noise Ratio (SNR) is used used. It is defined as the ratio of signal power to the noise power and is usually expressed in decibels (dB). Mathematically, the SNR is often defined as:

$$\text{SNR} = 10 \log_{10} \left(\frac{P_{signal}}{P_{noise}} \right)$$

Where P_signal and P_noise represent the power of the signal and the noise tensors, calculated as the sum of their squared elements. A higher SNR indicates a clearer and more distinguishable signal in comparison to the noise. In the context of this work, it will be used to assess the power of the noise introduced by the probabilistic degradation model, compared to the clean image.

5.4 Referenced image quality metrics

When the ground truth high resolution image is available, the performance of a super-resolution algorithm can be evaluated using a variety of metrics. These metrics can be divided into two categories: pixel-based and perceptual-based. Pixel-based metrics are based on the pixel-wise comparison between the generated image and the ground truth. Perceptual-based metrics, on the other hand, are based on the perceptual similarity between the generated image and the ground truth. These metrics are build using a pre-trained deep neural network, which is usually trained on a large dataset of images. The following sections will describe the most commonly used metrics in the literature.

5.4.1 pixel-wise losses

The L_1 and L_2 losses are the most commonly used pixel-based metrics in the literature. Additionally, they are usually used as the loss function that drives the super resolution network gradients during training. In a general form, the L_1 and L_2 losses are defined as follows:

$$\mathcal{L}_{L_k} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|^k \tag{18}$$

Where y_i and $\hat{y}_i$ are the ground truth and the super resolved image, respectively, and k is the exponent of the loss function. The L_2 loss weights high-value differences higher than low-value differences due to the exponent of 2. This generates overly smooth for low values and a lot of variability in high values. For that reason, it is more common to see the L_1 loss being used in the literature and it will be employed in this experiment .

5.4.2 Peak Signal-to-Noise Ratio (PSNR)

Peak Signal-to-Noise Ratio (PSNR) measures the magnitude of the error, compared relatively with the reference image. It is usually used to quantify the amount of error or noise introduced during the image reconstruction process.

PSNR is calculated by first computing the Mean Squared Error (MSE) or $\mathcal{L}_2$ loss between an image and the reference, and then taking the logarithmic ratio of the maximum possible pixel value squared. The PSNR value is usually expressed in decibels (dB).

The formula for PSNR is:

$$PSNR = 10 \cdot \log_{10} \left(\frac{I_{MAX}^2}{\mathcal{L}_2} \right) \tag{19}$$

where I_{MAX} is the maximum possible pixel value of the reference image, and $\mathcal{L}_2$ is the mean squared error between the image and the reference.

A higher PSNR value indicates better quality of the super-resolved image, as it signifies a lower level of noise or error. However, it's worth noting that it may not always align with human perceptual evaluations of image quality, as it focuses on physical consistency.

5.4.3 Structural Similarity Index (SSIM)

Structural Similarity Index Measure (SSIM) takes into consideration changes in structural information, luminance, and contrast. By doing that, it manages to reflect better the perceived changes in noise level and contrast. The SSIM index is calculated by dividing the image into windows of a certain size, and then comparing corresponding windows in the reference and target images. The SSIM index for a pair of windows, say x and y, is given by:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{20}$$

where μ_x and μ_y are the average pixel values, σ_x^2 and σ_y^2 are the variances, σ_{xy} is the covariance of x and y, and c_1, c_2 are small constants to avoid division by zero. The final SSIM score for the images is calculated by averaging the SSIM indices of all windows. An SSIM score of 1 indicates a perfect structural match between the two images, whereas a score of 0 indicates no structural similarities.

5.4.4 Learned perceptual image patch similarity (LPIPS)

LPIPS is a perceptual metric that leverages deep learning to compute perceptual differences between images. Specifically, it uses the activations of a pre-trained convolutional neural network (in this case, VGG [59]) to extract perceptual features from the images. Then, it calculates the Euclidean distance between these feature vectors to measure the perceptual difference. This measure has gained popularity in SR tasks due to its high correlation with human judgments of visual similarity.

The LPIPS score is given by:

$$LPIPS(I, I') = \sqrt{\sum_{i=1}^{N} w_i \| f_i(I) - f_i(I') \|^2} \tag{21}$$

where I and I' are the images being compared, $f_i(I)$ denotes the i-th layer activation when image I is input to the pre-trained network, N is the number of layers considered, and w_i is the learned weight for the i-th layer.

A lower LPIPS score indicates a lower distance between the feature vectors, and thus a greater perceptual similarity between the two images. Due to the fact that in this work we are interested in the physical consistency of the super-resolved images, this metric will be shown but will not drive any decision during the training process.

5.4.5 Adjusting measures to a slight translations in the SR process.

In order to calculate the losses and performance metrics, the generated test images (SR) are compared against the ground truth high resolution images (HR). Additional changes should be introduced in a MISR environment [35]. First, minor shifts on the contents of the pixels are expected and the metrics should have some tolerance to small pixel-translations in the high-resolution space by evaluating on a sliding cropped image. That means, looking for a displacement of SR by at most d pixels in each direction that minimizes the error. An example of how this is applied in a loss that needs to be minimized can be found in Eq. 22

$$\mathcal{L}^*(I^{HR}, I^{LR}, d) = \min_{u,v \in [0,2d]} \mathcal{L}(I^{HR}_{u,v}, I^{SR}_{u,v}) \tag{22}$$

Additionally, commonly used metrics punish biases as much as noise in the reconstruction. For example, if $I^{SR} = I^{HR} + \epsilon$, where ϵ is a constant bias, a perfect reconstruction of I^{SR} is possible if ϵ is known. A quality metric should award a high score in super-resolutions with this characteristics in comparison to the introduction of noise and information loss. Metrics like L2/L1 losses and PSNR do the exact opposite and should have a bias compensation like the following:

$$\begin{aligned} \mathcal{L}^*(I^{HR}, I^{LR}, d) &= \min_{u,v \in [0,2d]} \mathcal{L}(I^{HR}_{u,v}, (I^{SR}_{u,v} + b)) \\ b &= \frac{1}{(W-d)(H-d)} \sum_{x,y} \left(I^{HR}_{u,v} - I^{SR}_{u,v}\right) \end{aligned} \tag{23}$$

where W and H represent the width and height of the image, respectively.

5.5 Non-referenced Image quality metrics

No-Reference Image Quality Assessment (NR-IQA) aims to develop methods to measure image quality in alignment with human perception without the need for a high-quality reference image. Most of them are based on two steps: feature extraction and quality prediction using a regression module. They rely on the assumption that natural images share certain statistical information and that any distorsion may alter these statistics [60]. The results from any image an arbitrary image is compared to a default model trained on a large dataset of natural scenes. The difference between them is used to predict the quality of the image. In the last years, researchers relied on deep learning to perform the two steps in a single model. The workflow of these models is shown in Fig. 5.9.

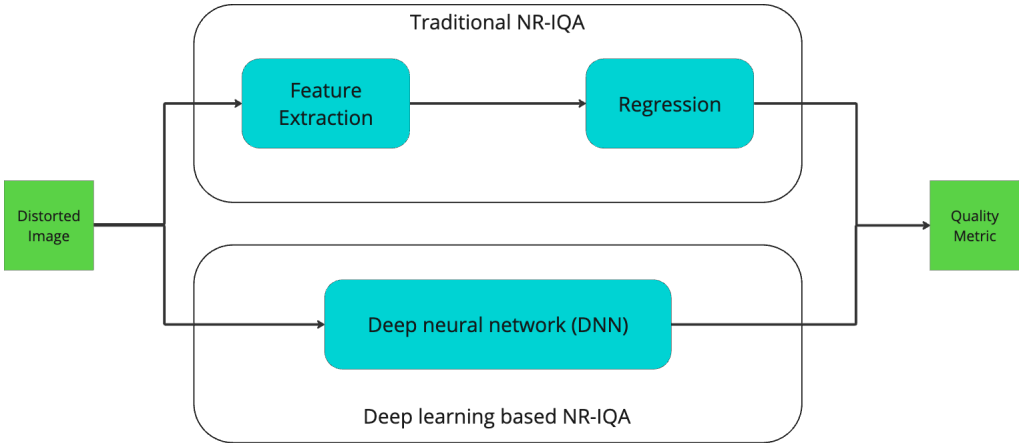

Figure 5.9: Workflow of a NR-IQA model.

5.5.1 Naturalness Image Quality Evaluator (NIQE)

The Naturalness Image Quality Evaluator (NIQE) [60] is a no-reference image quality assessment metric that quantifies the perceptual quality of images based on their naturalness. NIQE operates on the principle that pristine natural images exhibit specific statistical properties that can be quantified to establish a benchmark for quality assessment. NIQE employs a model based on a multivariate Gaussian distribution, characterized by a mean vector and covariance matrix, to represent the statistical attributes of a natural image's visual patterns. To assess the quality of an image, NIQE extracts a corresponding set of features and evaluates their deviation from this statistical model using the Mahalanobis distance. This distance measures the divergence of the image's features from those typical of high-quality natural images. A lower value suggests that the image closely resembles the statistical properties of natural images, indicating higher perceived quality.

However, NIQE provides a measure of image quality that aligns with the naturalness of human visual perception, and is not able to quantify the physical consistency of a generated image.

5.5.2 Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE)

Similar to NIQE, the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [61] operates on the premise that natural images possess certain statistical properties which are altered in the presence of distortions.

It operates by quantifying deviations from the statistical regularities observed in natural images, primarily using locally normalized luminance coefficients. Following this, a spatial domain model is utilized to calculate a set of features. These features, derived from the locally normalized luminance coefficients, are designed to capture the loss of "naturalness" due to the presence of distortions. These extracted features are fed into a Support Vector Regression (SVR) that was pretrained with a set of images with known quality ratings, to predict the quality score of the image.

A lower BRISQUE score is indicative of better quality, implying that the image has better quality score in the pre-trained model. The main difference between BRISQUE

and NIQE is that BRISQUE uses a Support Vector Regression (SVR) model and predicts directly the quality score, while NIQE calculates the distance between the multivariate Gaussian distribution model of the image and the one from the specific dataset where the evaluator was pre-trained.

5.5.3 Frequency Domain Analysis

The Fourier transform is widely used to analyze the frequency content in signals. It can be applied to multidimensional signals such as images, where the spatial variations of pixel-intensities have a unique representation in the frequency domain. Super-resolutions objective is to reconstruct missing high frequency components from a downscaled image. The expectation of a good SR algorithm is to amplify the high frequency components compared to a baseline like bicubic interpolation, while keeping noise at bay. The Fourier components provide global information about the image, as opposed to local information represented by pixel values in the spatial domain [62]. Using the Fast Fourier Transform (FFT), we convert the pixel intensity values of super-resolved images into a spectrum where each point represents a specific frequency contained in the spatial domain. The FFT is shifted so that the zero-frequency component is at the center of the spectrum. The resulting magnitude, after applying a logarithmic transformation, reveals the energy distribution across various frequencies. This is visualized in grayscale, where the intensity corresponds to the amplitude of the frequency components.

A radial profile of the FFT magnitude provides insights into how different spatial frequencies contribute to the image content in the vertical and horizontal direction. The radial profile is a function of the average intensity of frequencies at a given radius from the center of the Fourier transform. The average of the FFT magnitude is calculated for concentric circles of increasing radii, capturing a statistic of the frequency components in every direction. This metric serves as a benchmark for evaluating the performance of SR techniques against traditional interpolation methods such as bicubic interpolation.

Spatial frequency within an image context refers to the periodicity of the intensity variation over spatial dimensions, typically quantified in cycles per pixel. The central region of the frequency domain, after the shift operation, denotes the zero frequency. In contrast, the extremities of the domain delineate the highest frequencies, constrained by the image's discrete sampling rate. To quantitatively interpret these spatial frequencies, a radial-to-frequency mapping is necessary. This mapping accounts for the Nyquist frequency, which is delineated as half the sampling rate of the discrete imaging grid and acts as a threshold to prevent frequency aliasing. The conversion from a given radius in the FFT output to the corresponding spatial frequency is formalized as:

$$f(r) = \frac{r}{\frac{N}{2}} \cdot f_{\text{Nyquist}}, \tag{24}$$

where $f(r)$ signifies the spatial frequency associated with radius r, N represents the FFT image dimension, assuming a square configuration, and f_{Nyquist} the Nyquist frequency, which is 0.5 cycles per pixel in this case.

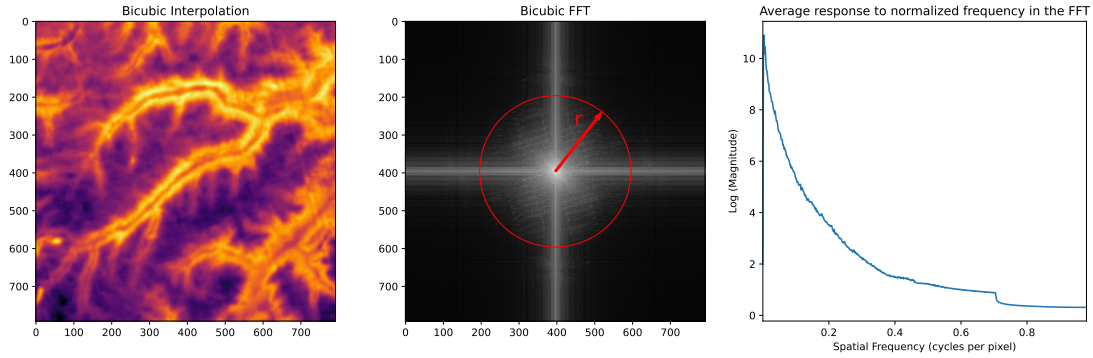

Figure 5.10: Steps of the frequency domain analysis. The Center image shows the log magnitude of the shifted FFT of a bicubic upsampled FOREST scene and an example of a radial profile, the average of all the points that have the same r is calculated. The right image displays the log magnitude obtained for every radial profile, translated into spatial frequency.

Through FFT a depiction of the amplification or attenuation of frequencies attributable to the SR techniques. Analyzing these profiles displays the ability of SR models for detail enhancement. However, it is important to note that this method does not account artifacts generated by the SR, and should be used in combination to other supervised metrics.

5.5.4 Gradient Distribution analysis

An alternative way of analyzing super-resolution results is by looking at the gradients of the images. HR images are sharper and thus each pixel, on average, has higher gradients magnitude with respect to both directions than their LR counterparts. A super-resolution algorithm should increase the sharpness of the edges, resulting in a gradient distribution that aligns more closely with that of the genuine HR image. An approximation of the gradients can be estimated by doing 2d convolutions between an image and the so called Sobel kernels displayed in Eq. 25 [63]. These kernels are designed to respond maximally to edges running vertically and horizontally relative to the pixel grid.

$$\hat{G}_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \qquad \hat{G}_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{25}$$

The kernels can be applied separately to the input image to produce the component of the gradient in each orientation G_x and G_y. The magnitude of the gradient is given by:

$$|G| = \sqrt{G_x^2 + G_y^2} \tag{26}$$

The gradient magnitude histograms of the results of different super-resolution algorithms will be assessed, there by quantifying the enhancement in edge sharpness. This histogram provide insights into the frequency and intensity of the edges within an image. A better SR model should demonstrate a histogram with higher frequencies of larger

gradient magnitudes, indicating sharper edges. However, it is important to note that this analysis is unsupervised and disregards the effect of noise and artifacts introduced during the super-resolution process and should be considered in combination with other supervised metrics like PSNR.

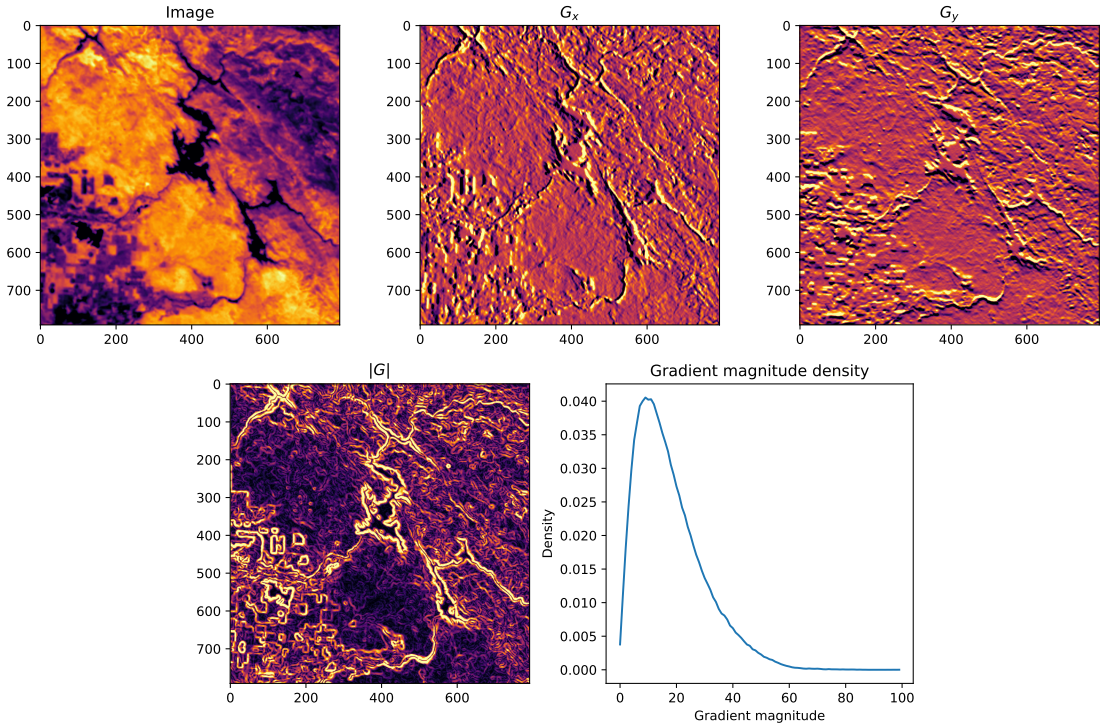

Figure 5.11: Steps to obtain a gradient magnitude density. Using the sobel operators, G_x and G_y are obtained from an image. The magnitude $|G|$ of each pixel is calculated using Eq. 26. The density can be estimated afterwards, using 100 bins in this case.

6 Datasets

6.1 Obtaining a high resolution dataset

Super-resolution is inherently a supervised learning task that needs the availability of high-resolution (HR) data. In scenarios where HR data from sources like FOREST-2 is unavailable, an alternative is to generate synthetic images from external missions, with similar characteristics as the FOREST-2 mission but with a superior resolution.

6.1.1 The ECOSTRESS mission

The NASA's ECOsystem Spaceborne Thermal Radiometer Experiment on Space Station (ECOSTRESS) mission is designed to provide new insights the effects of the Earth's climate dynamics [64], with focus on the following scientific objectives:

1. Identify the critical thresholds of water use and water stress in key climate-sensitive biomes, typically by observing the transition zones between biomes.
2. Identify when plants stop taking up water over the course of a day.
3. Improve the accuracy of drought estimates based on agricultural water use in the continental United States.

ECOSTRESS employs thermal infrared radiometers, specifically Prototype HyspIRI Thermal Infrared Radiometer [65] to measure the radiation emitted from the Earth's surface. It provides a spatial resolution of 69 meters with a temperature sensitivity of a few tenths of a degree [64]. The swath size is 400x400 km. The detector separates the energy from five different wavelengths using filters attached to the detector, producing five separate image layers for each scene. The pixels represent the intensity of thermal infrared radiation emitted by the Earth's surface at each wavelength. The mission has a 4-day diurnal repeat cycle.

In the spatial domain, ECOSTRESS constitutes an excellent candidate for generating synthetic HR images, as it's resolution constitutes approximately a x3 increase compared to FOREST-2.

In the spectral domain, it is important to confirm overlap between the missions bands. Given the narrower ECOSTRESS bands, the strategy will be averaging the radiances to align the spectral properties. Fig. 6.1 shows this spectral band comparison. In the case of the LWIR1 FOREST band, the overlap is significant with the first three ECOSTRESS bands. Althouth the overlap is less pronounced in the LWIR2 band, the radiation spectrum of black-bodies at prevalent surface temperatures suggest the feasibility of constructing a synthetic LWIR2 from the last two ECOSTRESS bands.

While FOREST's temporal resolution exceeds that of ECOSTRESS, allowing for the monitoring of new processes, this aspect is not the primary focus of the current study and will not be taken into account.

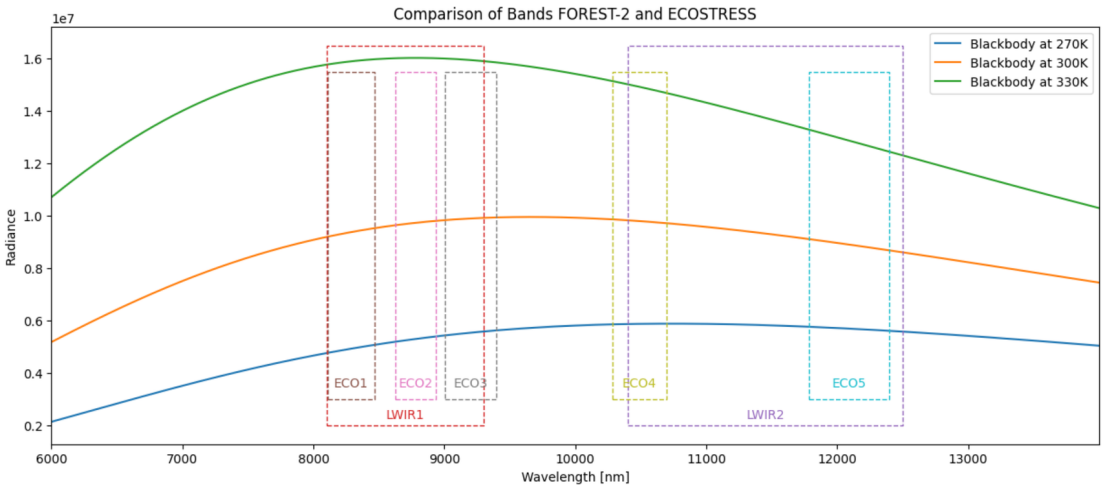

Figure 6.1: Wavelengths of the sensors in Ecostress and Forest satellites. The radiation spectrum of black-bodies at different temperatures are included for comparison.

6.1.2 Accessing ECOSTRESS Scenes

ECOSTRESS imagery is available via NASA's Application for Extracting and Exploring Analysis Ready Samples (AppEEARS) [66]. This tool allows the request of area samples via vector poligons. Using the product's API [67], Level 1 Mapped Radiance scenes of size 200x200 km with center on the locations provided in Fig. 6.1 were programmatically requested. Due to satellite hardware anomalies, certain spectral bands experienced acquisition gaps, needing a careful selection of date ranges to ensure the availability of all five bands [68].

Area	200 x 200 km
Products	Mapped Radiance (5 bands) Quality (5 Bands)
Dates	2018/08/20 - 2019/03/04 2023/05/01 - 2023/08/15

Table 6.1: Requests configuration

6.1.3 Selecting the best scenes

The AppEEARS platform returns multiple scenes that correspond to the specified area sample within the requested timeframe. This includes 5 mapped radiance measurements alongside their corresponding Quality Assurance (QA) bands. Additionally, a CSV file is provided, detailing quality statistics for each scene. The interface returns any scene that overlaps with the requested area. For that reason, some GeoTIFFs may be significantly smaller than others, with variances up to 90%. Moreover, an important number of these GeoTIFFs may contain a high percentage of bad quality pixels, rendering them unsuitable for model training. Furthermore, as highlighted in the ECOSTRESS frequently asked questions [69], the accuracy of radiance measurements is highly dependent on clear sky conditions; cloudy scenes typically yield negligible radiance emissions.

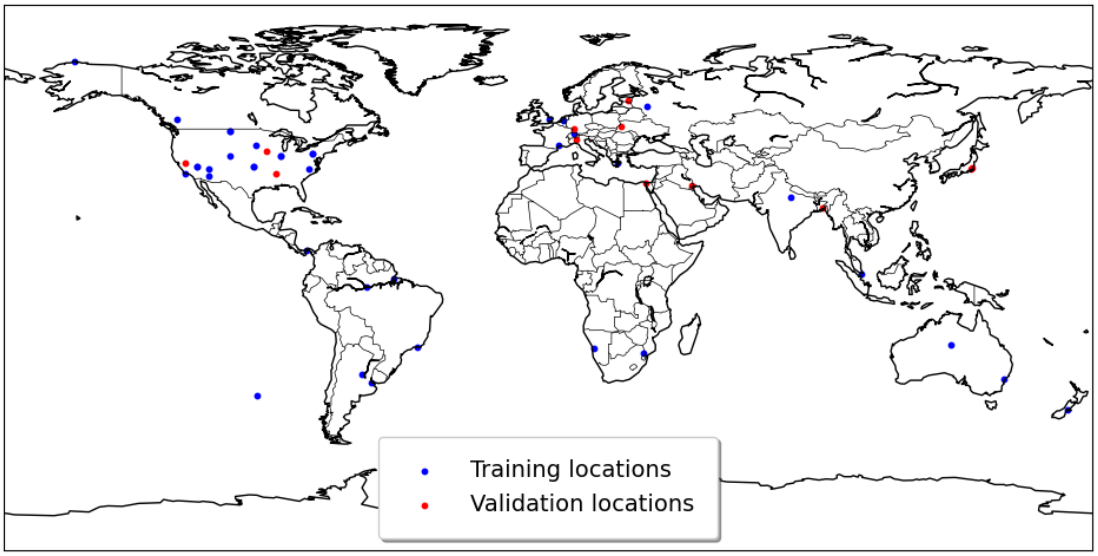

Figure 6.2: Location of the samples taken from ecostress.

The dataset includes several GeoTIFFs for each scene. Downloading the entirety of this dataset is impractical due to its huge size. From the 50 scenes, each one is potentially replicated over 20 times over the 10 months request window. Such a dataset, given its magnitude, cannot be used for model training with the available hardware resources. Therefore, a procedure is developed to identify and select the most appropriate scene for each month, based on a predefined set of criteria:

1. Scenes should have a low proportion of bad quality pixels.
2. Scenes should have a considerable size so that many crops can be taken from it.
3. As clouds imply low radiance values, clear sky scenes will have high radiance values.

The procedure to get the best scene for each month is detailed below:

Algorithm 1 Process applied to the scenes returned from one area request.

QA statistics:
Get the average proportion of good pixels p_{gp} for the 5 radiances of the scene.
Discard scenes where $p_{gp} < 60$.
Scene Statistics:
Get the biggest scene of each month.
Calculate the proportion between the size each scene and the biggest of the month.
Discard images which size proportion is smaller than 0.2.
Calculate the median of the radiance values of the scene.
Selecting the scene of the month:
Merge the QA statistics and the Scene statistics.
For each month, get the 3 scenes with the greatest p_{gp}.
Select the scene that has the greatest median radiance value.

Applying this procedure, a dataset comprised of 5031 scenes taken from 50 area requests is reduced to 379 scenes.

6.1.4 Data Processing

In order to be able to use the data in a super-resolution algorithm, a set of processing steps must be performed on it.

The diagram in Fig. 6.3 displays the processing pipeline. The input are the 5 Mapped radiance and their respective quality bands.

Mapped radiances 1,2 and 3 are averaged to form the LWIR1 synthetic FOREST, mapped radiances 4 and 5 are averaged to form the LWIR2 synthetic FOREST. If any of the bands are missing, the corresponding LWIR synthetic forest is discarded.

The fill values in the mapped radiances and the data quality classes are used to create a binary mask for each spectral band. If a pixel is considered problematic, it is marked as a 1 in the binary mask. The QA band for a synthetic FOREST LWIR band is built using an OR operation on the corresponding ECOSTRESS spectral involved in it's construction. After being constructed, both the synthetic LWIR and the corresponding QA band are reprojected to the best utm epsg code, based on the latitude and longitude of the scene.

Value	**Description**
	Fill Value Classes
-9997	Pixel not seen
-9998	Missing data due to striping (not filled in)
-9999	Missing/bad data
	Data Quality Classes
0	Good
1	Missing stripe data, filled in
2	Missing stripe data, not filled in
3	Missing/bad data
4	Not seen

Table 6.2: Fill Value and Data Quality Classes

The synthetic LWIR are not suitable for the super-resolution task yet. They are too big to be kept in memory, and not all their values are of good quality. For that reason, for each scene, a number of random crops of size 264x264 pixels are taken. The random crop processor pipeline is displayed in Fig. 6.4. It is an iterative process where at each stage, crops that do not comply with the quality considerations (all pixels are of good quality and no stripe noise was detected) are discarded until the target number of crops per scene is achieved. Additionally, the Affine Transformation is translated so that the images can be georeferenced.

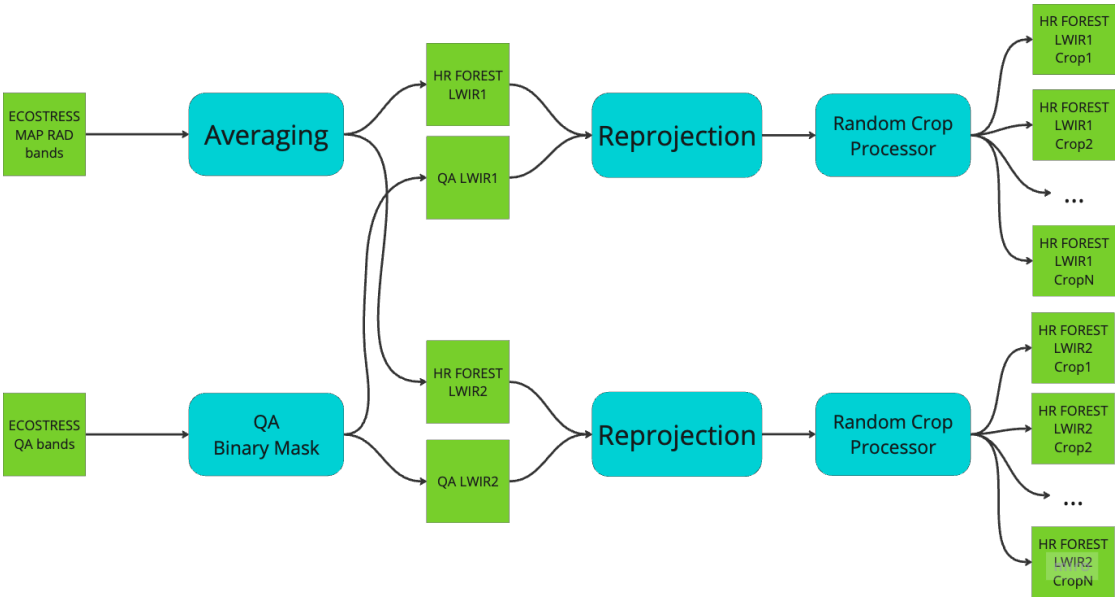

Figure 6.3: Data processing workflow

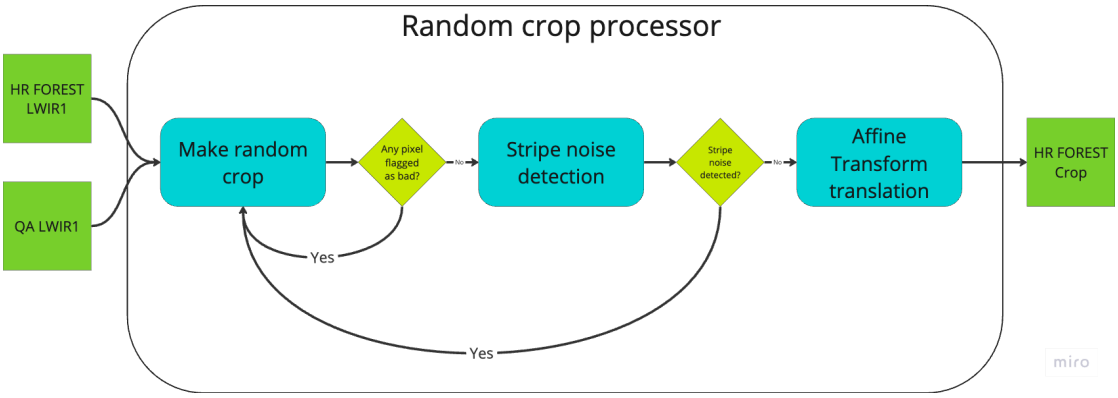

Figure 6.4: Random crop processor

6.2 Obtaining FOREST-2 data

To obtain a dataset of FOREST-2 image, the company provides an internal API that allows the download of the scenes captured by the satellite. The download of the scenes is done programmatically in the locations provided in Fig. 6.5.

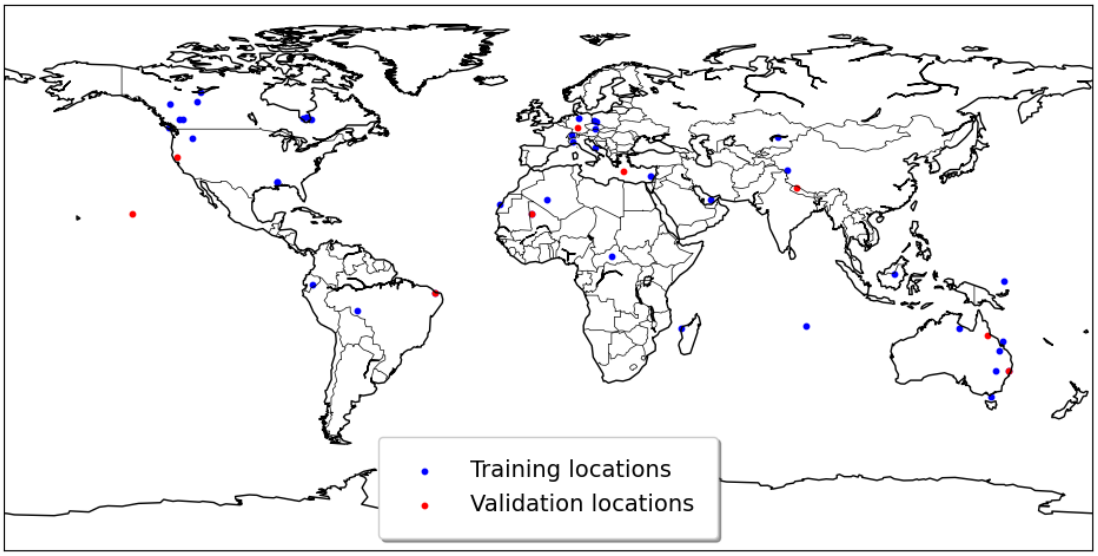

Figure 6.5: Location of the FOREST-2 scenes.

The scenes are downloaded in the NetCDF-4 format, containing the LWIR1, LWIR2 and MWIR bands, as long as the latitude and longitude information. An example of a scene is shown in Fig. 6.6. As in this work, the focus is on the long wave infrared, the MWIR band is discarded.

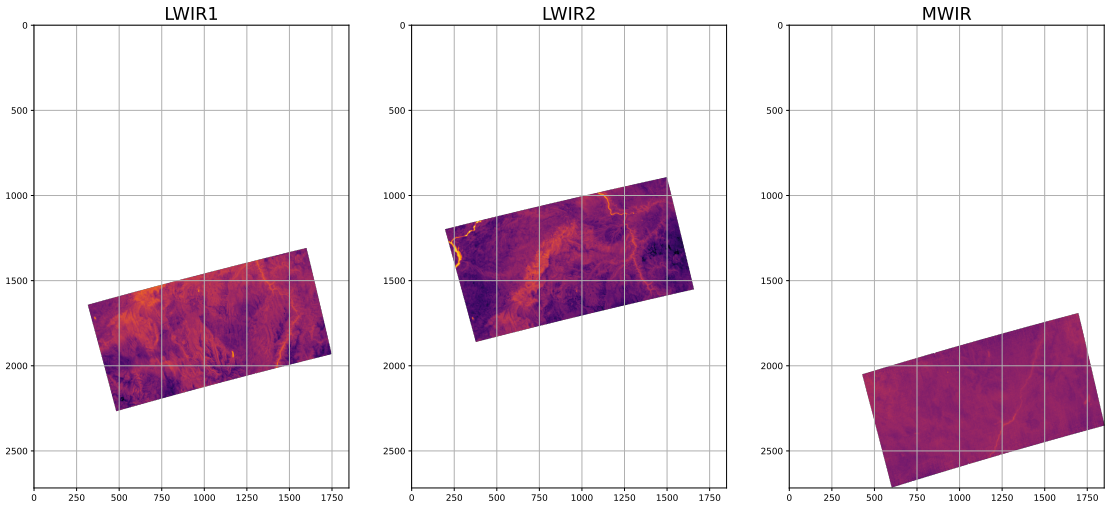

Figure 6.6: LWIR1, LWIR2 and MWIR bands of a FOREST-2 scene downloaded from the company's API.

The provided array have an enormous proportion of NA values and are not suitable for taking crops. For that, a bounding box is defined for each band, removing most of the NA values. The resulting scenes are shown in Fig. 6.7.

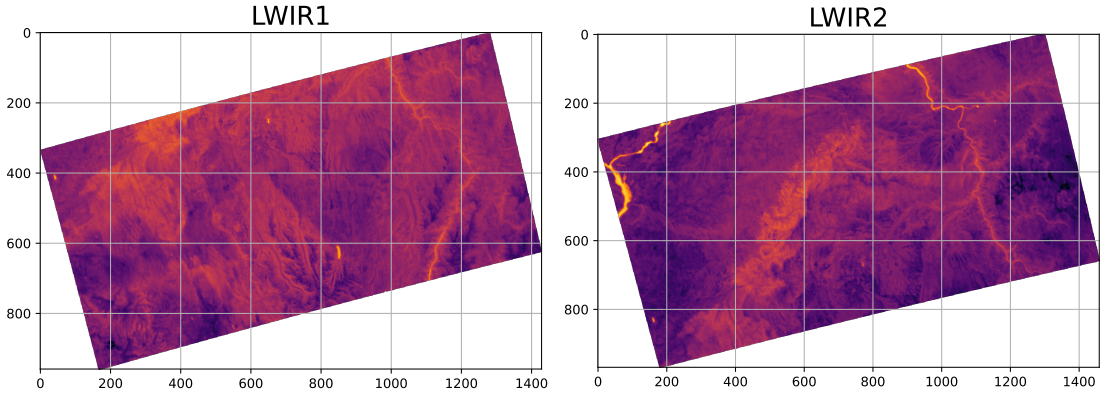

Figure 6.7: LWIR1 and LWIR2 of a FOREST-2 scene downloaded from the company's API, after cropping NA values.

A similar random crop processor as in Fig. 6.4 is employed afterwards. The objective is to obtain crops 3 times less than the original size, as the objective is to train a SR model with a scale factor of 3. If any crop has any NA value or strip noise is detected, it is discarded and the process restarts. Using the latitude and the longitude information, the affine transformation is calculated so that the crops can be georeferenced, if necessary.

6.3 Datasets

For a better understanding of how the proposed architecture works, several datasets combinations are used. The implemented pytorch dataset class loads and yields samples from two different file locations, one for the HR images (source domain) and one for the LR images (target domain). The source domain are the synthetic FOREST-2 images produced from ECOSTRESS, while the target domain is composed of LR images, coming from different souces. The samples are usually unpaired, meaning that the scenes are not compared on an image-to-image basis, but the implementation allows the use paired datasets in order to calculate supervised metrics. In case any of the domains has less samples than the other, the class will bootstrap it to match the size of the other.

	$\mathcal{D}_{\mathrm{SF-SF}}$				$\mathcal{D}_{\mathrm{SF-RF}}$				$\mathcal{D}^{\mathrm{Paired}}_{\mathrm{SF-RF}}$			
	Training		Validation		Training		Validation		Training		Validation	
	Source	Target	source	target	source	target	source	target	source	target	source	target
Image	Synth FOREST	Degraded synth FOREST	Synth FOREST	Degraded synth FOREST	Synth FOREST	Real FOREST-2	Synth FOREST	Real FOREST-2	Synth FOREST	Real FOREST-2	Synth FOREST	Real FOREST-2
n	13764	13764	2676	2676	13764	4000	2676	1200	13764	4000	??	??
scale ratio	x3	x3	x3	x3	x3	x3	x3	x3	x3	x3	x3	x3
crop size	264	88	264	88	264	88	792	264	264	88	??	??
Paired?	No		Yes		No		No		No		Yes	

Table 6.3

6.3.1 Synthetic FOREST - Degraded Synthetic FOREST

The dataset $\mathcal{D}_{\mathrm{SF-SF}}$ is built by taking the HR synthetic FOREST crops and applying the baseline degradation model proposed in 5.2. The 264x264 crops are reduced to 88x88.

The training set is used to train the SR Resnet model, while the validation set is used to monitor the training process and avoid overfitting. Even though in this case the HR and LR version of the same scene is available, the training dataset is unpaired by shuffling the samples. The validation set is not shuffled, and thus can be used to calculate supervised metrics like PSNR and SSIM.

The parameters used for the degradation model are described below:

Parameter	Value
Scale ratio	x3
Gaussian Kernel size	21
Gaussian kernel sigma in X axis	$\sim \mathcal{N}(1, 0.3)$
Gaussian kernel sigma in Y axis	$\sim \mathcal{N}(1, 0.3)$
target radiometric error	1.5K
white noise factor	0.5
constant noise factor	0.5

Table 6.4: Parameters used in the degradation model employed to generate the $\mathcal{D}_{\mathrm{SF-SF}}$ dataset.

6.3.2 Synthetic FOREST - real FOREST (Unpaired)

The dataset $\mathcal{D}_{\mathrm{SF-RF}}$ is composed of the 264x264 HR synthetic FOREST-2 crops as the source domain and 88x88 real FOREST-2 crops as the target domain. Unfortunately, the validation dataset is not paired, as the HR and LR images are completely different scenes. Thus, supervised metrics are not available for the super resolved target domain images is not available. The metric used to determine the best model is the PSNR from the super resolution of the output of the GAN's generator.

6.3.3 Synthetic FOREST- real FOREST (Paired)

While the training dataset is the same as in the previous case, the validation dataset is composed of a limited amount paired scenes between ECOSTRESS and FOREST are available. This samples allow the calculation of supervised metrics like PSNR and SSIM. As supervised metrics to compare the real FOREST-2 SR with the ground truth is now available, the PSNR is used to determine the best model.

7 Experiment Setup

7.1 Training

Instance normalization

8 Results and discussion

For each dataset, the combination of the probabilistic degradation model and the SR model (from now on, a pipeline) was trained. Each pipeline has 3 main components:

- A generator, used to generate LR images similar to the target domain, from HR images coming from the source domain.
- A discriminator, used to distinguish between real and generated LR images.
- A SR model, used to super resolve the LR images generated by the generator or the real LR images coming from the target domain.

The pipeline trained on $\mathcal{D}_{\mathrm{SF-SF}}$, using unpaired HR-LR pairs generated by applying the baseline degradation model described in 5.1 to the synthetic FOREST-2 images, will be referred to as the baseline pipeline. While the employed degradation model is stochastic, it has known parameters. The objective is to observe how the GAN is able to imitate a known degradation model in order to produce LR images.

The pipeline trained on $\mathcal{D}_{\mathrm{SF-RF}}$, using unpaired HR-LR pairs of synthetic and real FOREST-2 images, will be referred to as the adapted pipeline. In this case, the degradation model is unknown and the objective of the GAN is to to estimate it, generating LR versions of the synthetic FOREST images that come from the same distribution as the real FOREST images.

8.1 Source domain

This subsection will analyze the results from the experiments performed on the source domain. The process consists of degrading the synthetic HR FOREST images using the generator trained using adversarial learning and then super resolving it using the corresponding SR model from the pipeline. This is the equivalent of the black arrows flow described in fig. 5.4. As in this case the ground truth is known, the performance of the super resolution can be evaluated using metrics like PSNR and SSIM.

Fig. 8.1 shows the results of the baseline and the adapted pipeline, when applied to one sample from the source domain (a synthetic HR FOREST-2 image). For comparison, a pipeline consisting of simple gaussian blurring + downscaling for degradation and bicubic upsampling for SR is also shown.

While the baseline kernel is very simple and the noise is more or less uniform across the image, the adapted kernel is more complex and the noise seems to be strongly correlated with the image intensity. It is important not to overinterpret this result, as the kernel and noise are estimated using overparametrized models, and multiple combinations of kernel and noise may produce similar results. However, it is interesting to see that the adapted pipeline is able to estimate a more complex degradation model, which is closer to the real degradation model used in the target domain.

The degraded LR images present considerable differences. While the baseline pipeline produces images very similar to gaussian blurring + downscaling, the adapted pipeline produces much more blurry images with more noise, suggesting that FOREST-2 produces less resolution than what was initially expected. This is also confirmed by calcu-

lating the PSNR between the LR image generated by each pipeline with the gaussian blurring + downscaling LR image, which yields worse results for the adapted pipeline.

The super resolved produces by both pipelines yield better performance than bicubic interpolation, and they are very similar between them. This suggests that the super resolution model is able to recover the details lost during a more complex degradation processes, but there seems to be a limit to the amount of detail that can be recovered. It is observed that even though the starting point is different (baseline LR is less blurry than adapted LR), the final result is very similar.

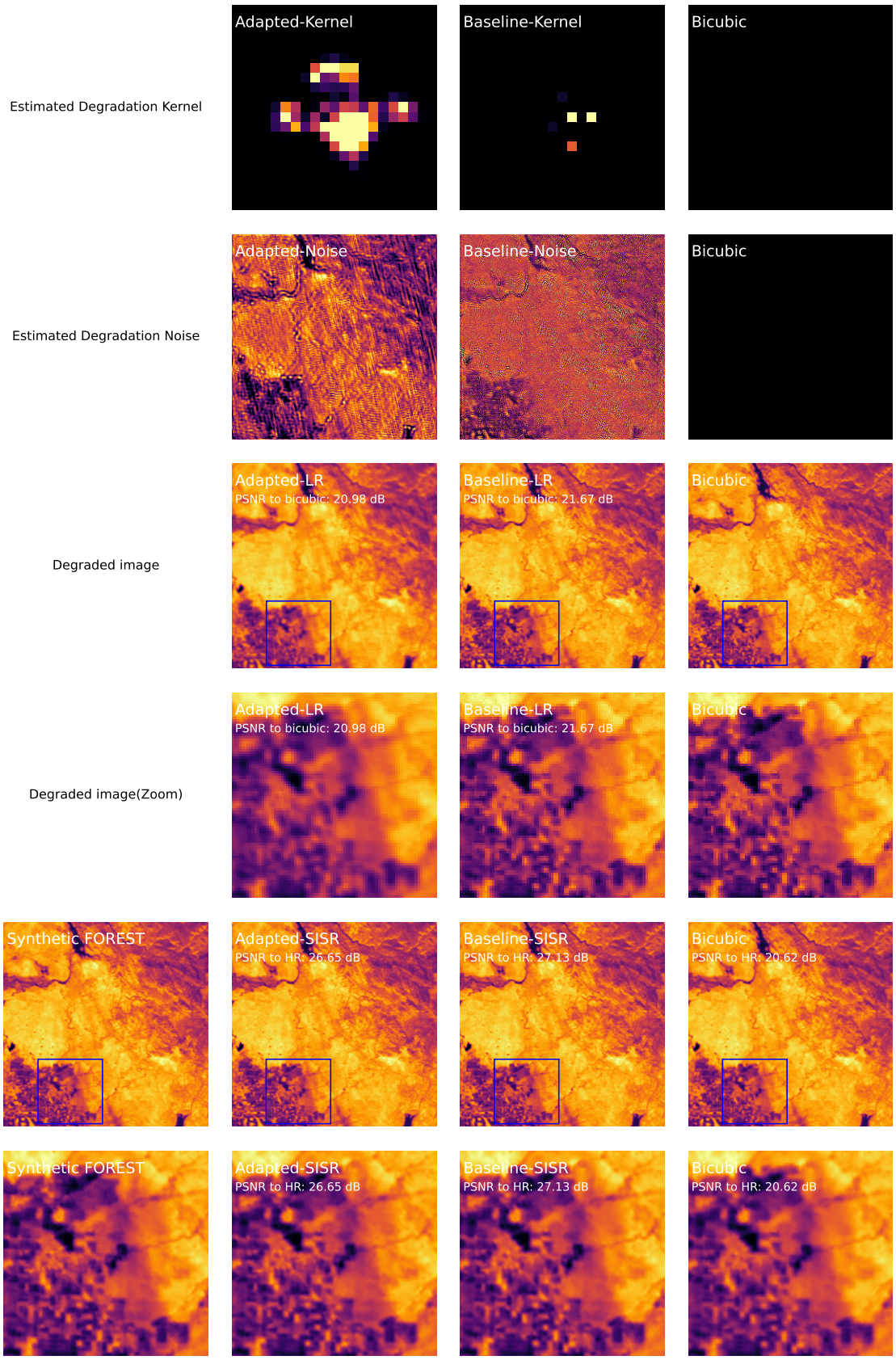

Figure 8.1: Applying different degradation models on an HR sample. The 2 most upper rows show the estimated degradation kernels and noise of each pipeline, the bicubic downsampling does not estimate a kernel or noise. The degraded LR images from each model and a zoom is displayed on the two subsequent rows. In this case, the PSNR is calculated against the gaussian blurring + bicubic downsampling LR. The synthetic FOREST-2 (ground truth) and the super resolved images, with a zoom, are displayed in the last 2 rows. The PSNR for each SR method is calculated against the HR synthetic FOREST-2.

In Figs 8.2 the frequency domain of the LR images is analyzed. By inspection of the FFTs, it is observed that the adapted-LR loses more information than the baseline-LR, as the log magnitude of the FFT get cut more close to the center. The baseline-LR FFT is very close to the gaussian blurring + bicubic upsampling FFT, suggesting that the baseline pipeline is able to mimic this known degradation model.

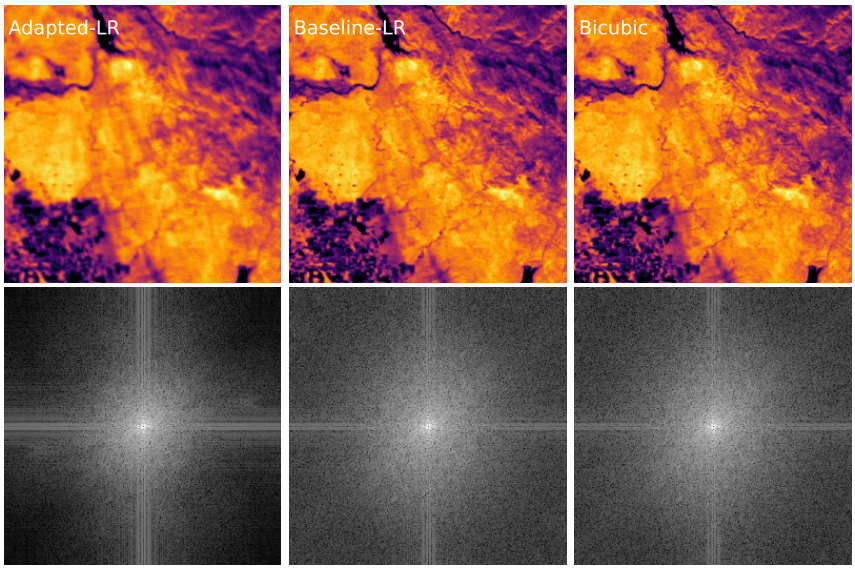

Figure 8.2: Log mangnitude of the FFT for the LR images obtained by the pipelines and the gaussian blurring + bicubic upsampling.

The radial profile of the log magnitude of the FFT for the LR images shown in Fig. 8.3 confirmed what was observed previously. The adapted-LR image diminishes the high frequency components much more than the baseline-LR image with amplifications of -6dB in frequencies starting at 0.1 cycles per pixel, with a stable effect of -6dB from 0.3 to 0.7 cycles per pixel. It is important to note that 0.1 cycles per pixel at a 210m GSD corresponds to a cycle frequency of 2100 m^{-1}, 0.3 cycles per pixel corresponds to 700 $m^{-}1$ and 0.7 cycles per pixel to 300 m^{-1}. This suggests that the degradation model from the real FOREST-2 images is more complex and loses more information than the baseline degradation model. An analysis for the whole validation dataset will be further discussed to verify that this behaviour is consistent across different scenes and conditions.

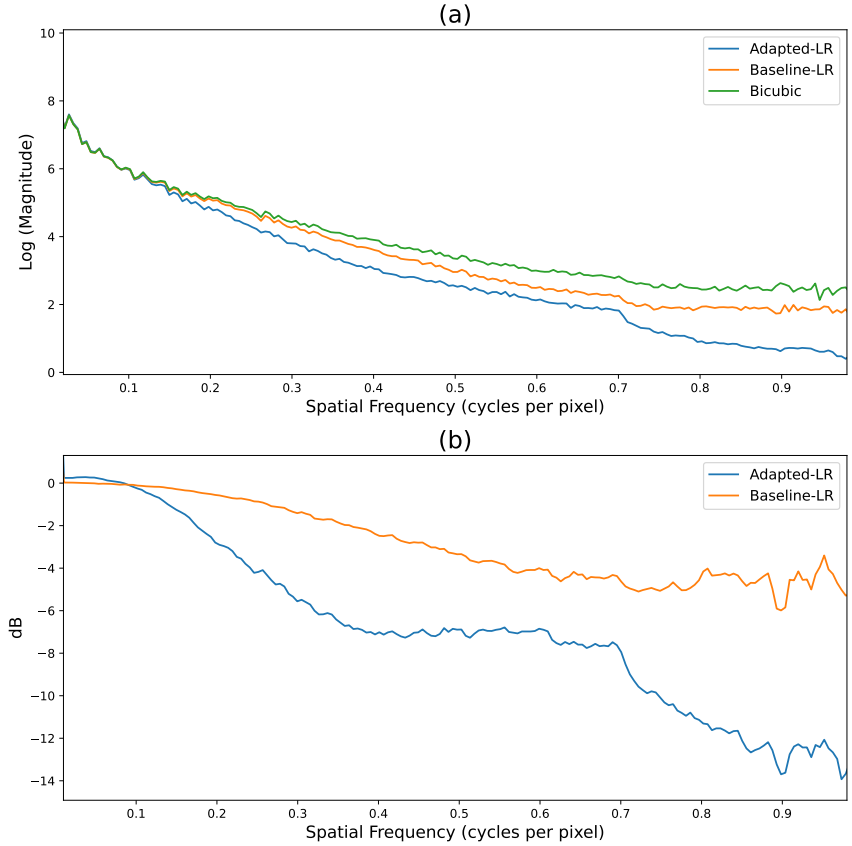

Figure 8.3: (a) Radial profile of the log magnitude across spatial frequency of the LR images obtained by the pipelines and the gaussian blurring + bicubic downsampling model. (b) Amplification in dB of each pipeline with respect to the gaussian blurring + bicubic downsampling.

When analyzing the super resolved images versus the ground truth in the frequency domain, a very similar frequency response is observed for both pipelines. Moreover, the SR images are able to stay above -3dB, a common threshsold used in the literature, up until 0.3 cycles per pixel, which correspond to $300\frac{1}{m}$ when each pixel equals 70m. This suggests that the SR model in the adapted pipeline is able to recover the lost information at those frequencies due its more complex degradation model. Starting at 0.3 cycles per pixel, a decrease in amplification is observed for both pipelines, but more steeply for the adapted pipeline. This may be related to the fact that the adapted degradation model diminishes cycles at higher frequencies even more than the baseline degradation model. A limit for the SR algorithm is also noted, even using an optimistic degradation model such as the baseline, the SR model is not able to recover higher frequencies with respect to the original, HR image. Even if it is slightly better than bicubic upsampling, the diminishing of the higher frequency components is dramatic.

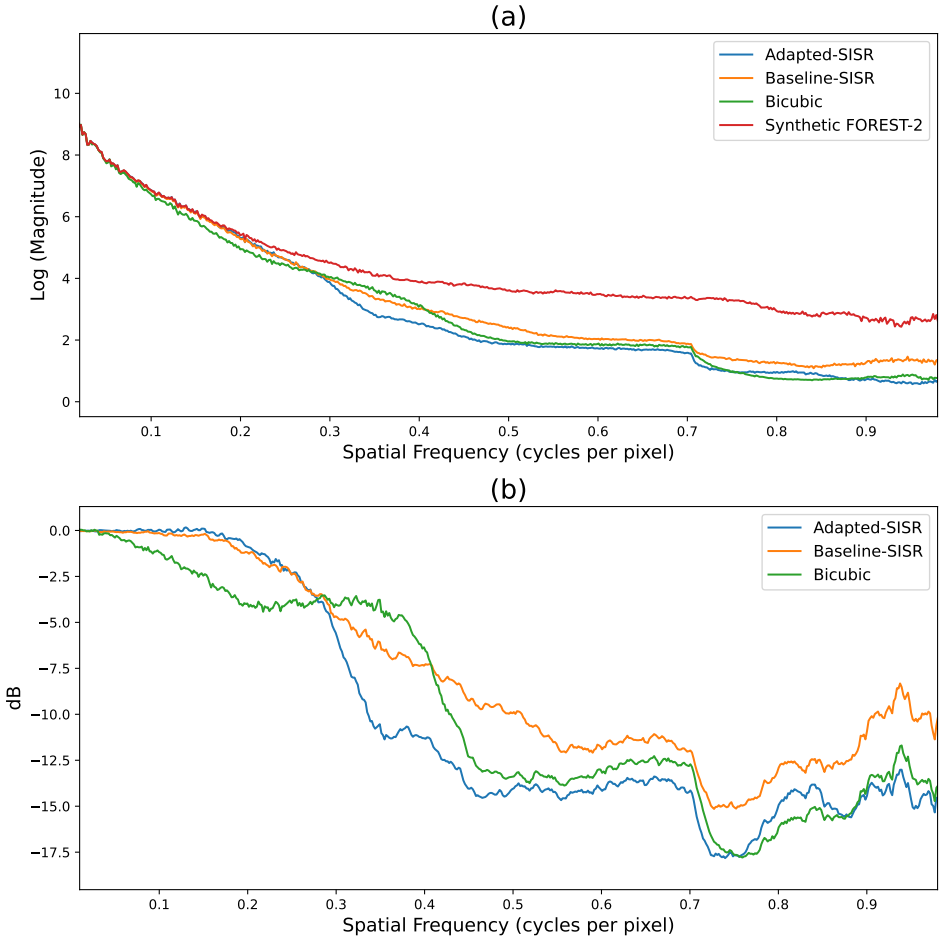

Figure 8.4: Frequency domain analysis of the SR images and the ground truth displayed in 8.1. In (a), the log of the magnitude of the FFT for the SR images and the ground truth is shown, while in (b), the amplification of each SR image with respect to the ground truth is shown.

8.1.1 Probabilistic degradation models comparison

In order to better understand the stochastic nature of the generator, a kernel was extracted 2000 times from the it using different realizations of the random variable z_k. The mean and standard deviation of the sampled kernels was then computed. It is important to note that the experiment configuration assumes that the kernel does not depend on the pixel content or position, resulting in one kernel per image. The results are shown in Fig. 8.5. In order to make the standard deviation of each pixel comparable, its value is normalized by the mean value of the corresponding pixel. This allows to express the standard deviation as a percentage of the mean corresponding pixel value.

While the baseline and the adapted kernel have the maximum mean and std in the same pixel, the adapted one is much more spread out. The baseline kernel is composed of a few pixels very close to each other. This suggests that the adapted kernel is more complex and spread out, while the baseline kernel is simpler and more concentrated. The result of this is that the adapted kernel produces more blurry images, as the kernel is more spread out, while the baseline kernel produces less blurry images, as the degradation is

more concentrated. The figure also displays the benefits of the probabilistic degradation model. Using only one HR image, the generator is able to produce a wide variety of LR pairs. Allowing the training of SR models able to generalize better to the real world.

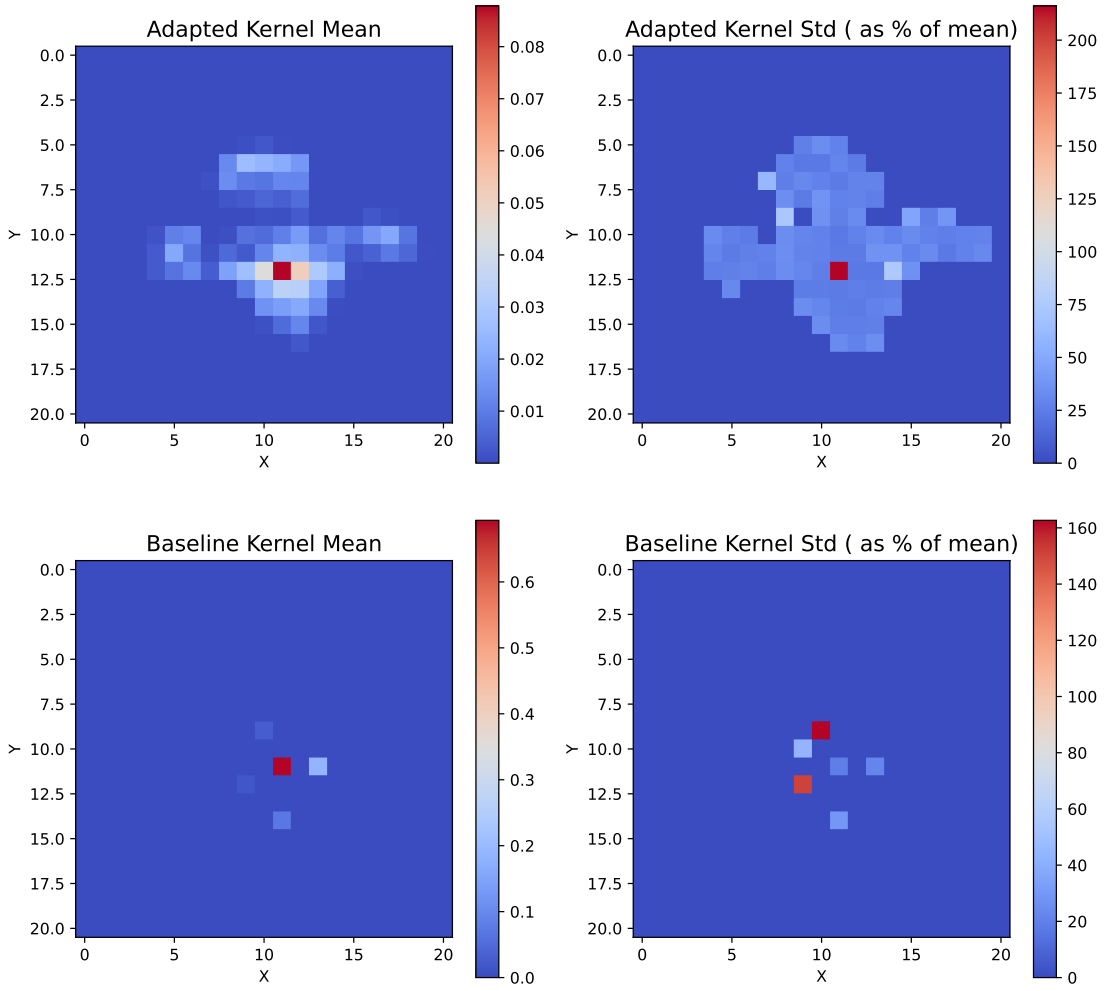

Figure 8.5: Mean and standard deviation of the estimated kernels for the baseline and adapted degradation model, using 2000 realizations of z_k. The standard deviation of each pixel is normalized by the corresponding mean value. kernel pixels with mean lower than 10^{-4} are considered with 0 std for clarity in the plot.

In the case of the noise, the experiment setup assumes that it depends on the pixel content and position. For that reason, two different characterizations were done. First, The stochastic component of the noise will be assessed by computing the SNR between the clean image $I_{\text{LR}}^{\text{clean}}$ and the output of the noise module for one HR input and 2000 realizations of z_n. This shows again the benefits of a stochastic model, several LR versions of an HR image can be generated, enriching a training dataset.

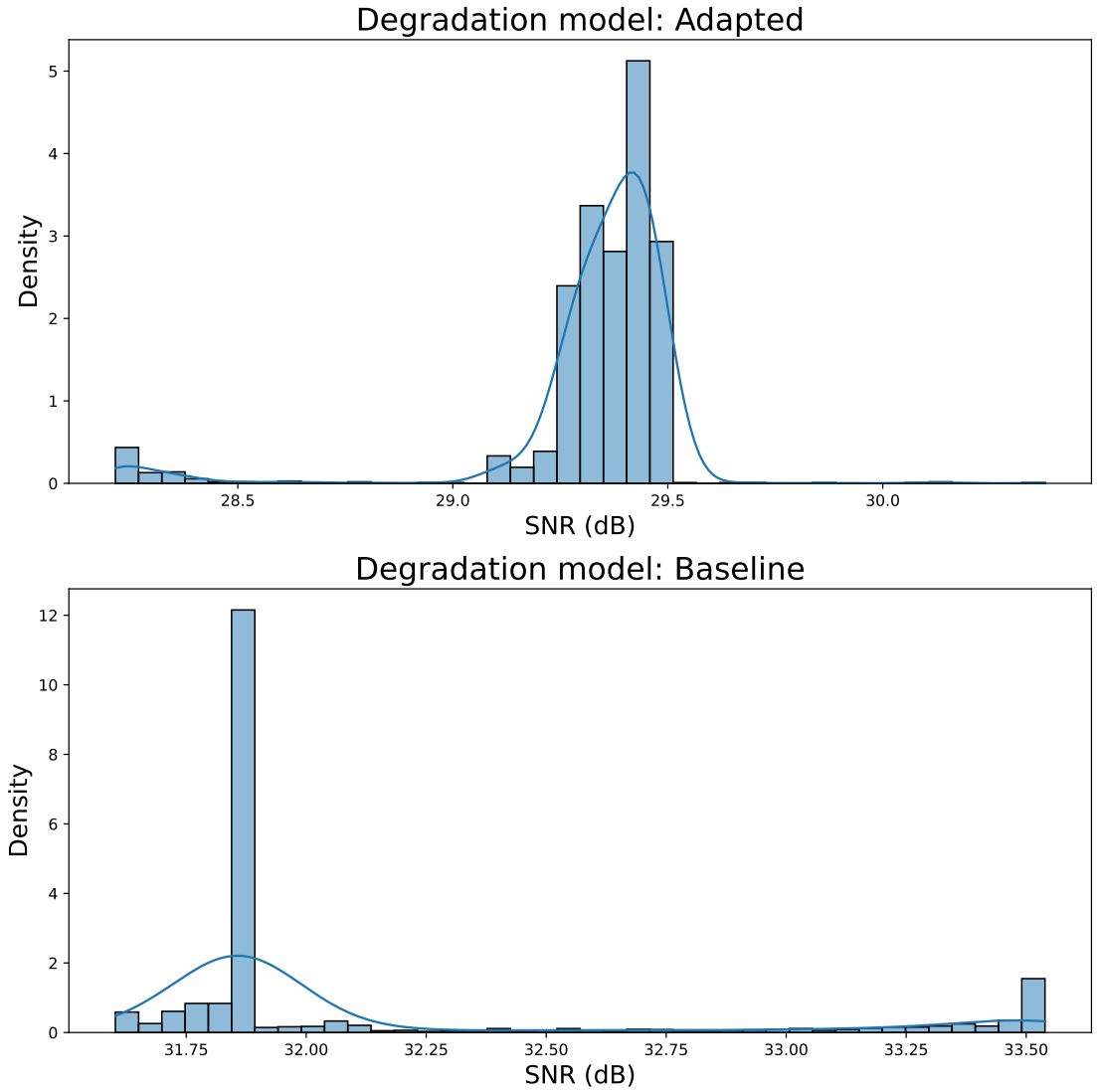

Figure 8.6: Distribution of SNR values using $I_{\text{LR}}^{\text{clean}}$, product of the convolution of the kernel and $I_{\text{HR}}^{\text{clean}}$, and the noise module output for both pipelines. The output noise is generated 2000 times, using different realizations of the random variable z_n for each iteration and the same input image.

Second, the signal-to-noise (SNR) ratio between the clean image $I_{\text{LR}}^{\text{clean}}$ and the output of the noise module of the generator will be computed one time for the whole validation dataset. An estimated density function of the SNR for each degradation model is shown in Fig. 8.7. The SNR is in general bigger when using the baseline model compared to the adapted one. This implies that in the output of the generator, the noise has more energy compared to the clean image when the target domain is composed of scenes coming from the FOREST-2 satellite. The distribution of the noise can not be assessed using this method, but it was observed in Fig. 8.10 that it seems to have a higher dependency on the pixel value of the input.

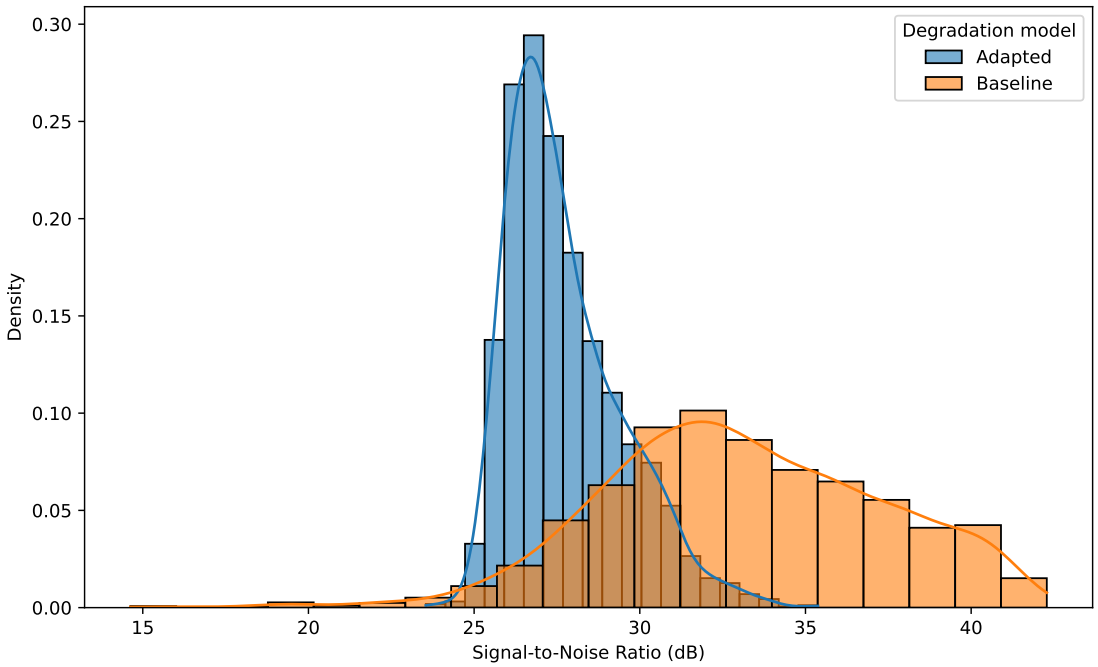

Figure 8.7: Comparison of the SNR expressed in dB of the low resolution images generated by the baseline and adapted degradation model.

The observed effects support what is observed in the Fig. 8.10. The adapted pipeline produces broader kernels and noise with more energy, compared to the baseline pipeline. This leads to more blurry and more noisy generated LR images. The kernel imposes the diminishing of frequency components in the signal, and the noise degrades the ratio of clean signal energy in the image. Both components will create a more difficult scenario for the super resolution model, which will try to go back to the HR image from the generated LR image.

8.1.2 Low resolution images comparison

A quantitative analysis of the LR images obtained by the generator of each pipeline is performed. Fig. 8.8 shows 3 supervised performance metrics obtained by comparing the LR images obtained by the pipelines with the gaussian blurring + bicubic downsampling degradation. In this case, a consistent higher PSNR and SSIM means that the baseline-LR image is closer to the gaussian blurring + bicubic downsampling LR image than the one generated by the adapted pipeline. A lower LPIPS means that even using perceptual metrics, the baseline-LR image is also closer. This is consistent with the results shown in Fig. 8.1, where the adapted LR image is more blurry and noisy, suggesting that the unknown degradation is far from a baseline degradation model.

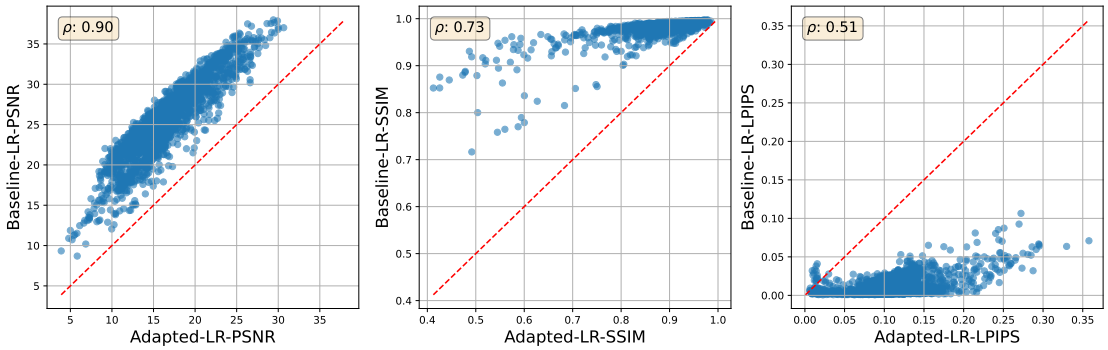

Figure 8.8: Performance metrics between the LR images obtained by the pipelines vs the gaussian blurring + bicubic downsampling degradation. On the left, the PSNR is displayed. On the middle and the right, SSIM and LPIPS are represented respectively.

An alternative way to evaluate the differences in the degradations is by analyzing the frequency domain of the LR images. An analysis of the whole validation dataset is performed by calculating the FFT of each LR image and comparing them with the gaussian blurring + bicubic downsmapling degradation model. The results are displayed shown in Fig. 8.9. In (a) the log magnitude of the FFT across different spatial frequency values for the degraded images is shown. The spatial frequency is obtained from the radial distance to the center of the FFT, as shown in 5.5.3. In (b), the amplification of each generated LR image with respect to a simple gaussian blurring + downscaling is shown. The results for the whole dataset show that the LR images generated by the adapted pipeline yield a reduction in the higher frequency components consistently across all samples, with a $\pm$ 1 standard deviation interval between -4 and -6 dB from between 0.3 and 0.7 cycles per 210m pixel.

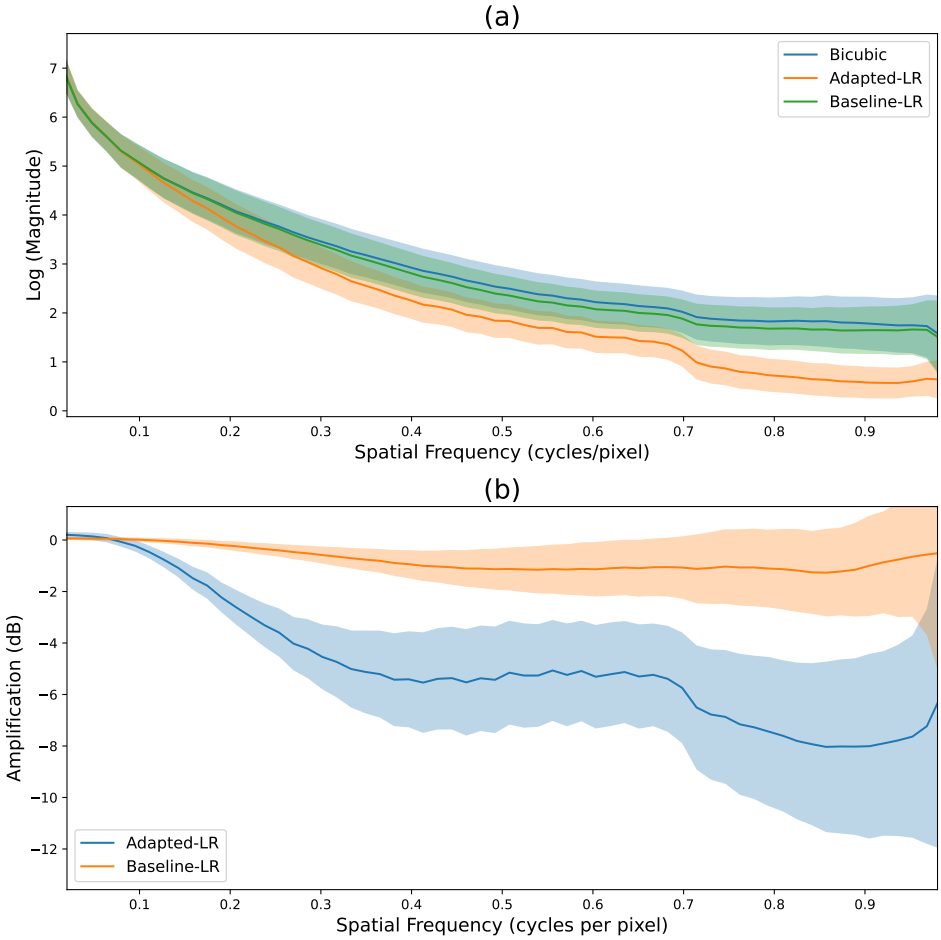

Figure 8.9: Frequency domain analysis of the LR images obtained by applying different degradation models on the HR sample displayed in Fig. 8.1. In (a), the log of the magnitude of the FFT for the LR images is shown, while in (b), the amplification with respect to a simple gaussian blurring + downscaling is shown. The painted area represents the ±1 standard deviation of the radial profiles and the amplification.

8.1.3 Effects of the degradation model in SR

Another subject of interest is how the degradation model affects the performance of the super resolution process. Fig. 8.10 shows the performance obtained by super resolving the output of each pipeline generator for the whole validation dataset. In the upper row (a), the corresponding SR model of each pipeline is used to obtain the super resolved images. The performance, both in PSNR and SSIM, are very similar for both pipelines. The LPIPS shows a consistent behavior too. In the lower row (b), the SR model is discarded and a simple bicubic upsampling is used to super resolve the degraded images of each pipeline. In this case, using the baseline LR version as input consistently yields better results than the adapted LR version, in all metrics. This suggests that the learned degradation model from FOREST-2 images loses more information than the baseline, resulting in a lower effective ground sampling distance than what was is specified in FOREST-2 fact sheet. Consistent with what was found in the frequency domain analysis observed in Fig. 8.4, the SR model is able to recover most of the information, as the

performance when employing the SR models is very similar.

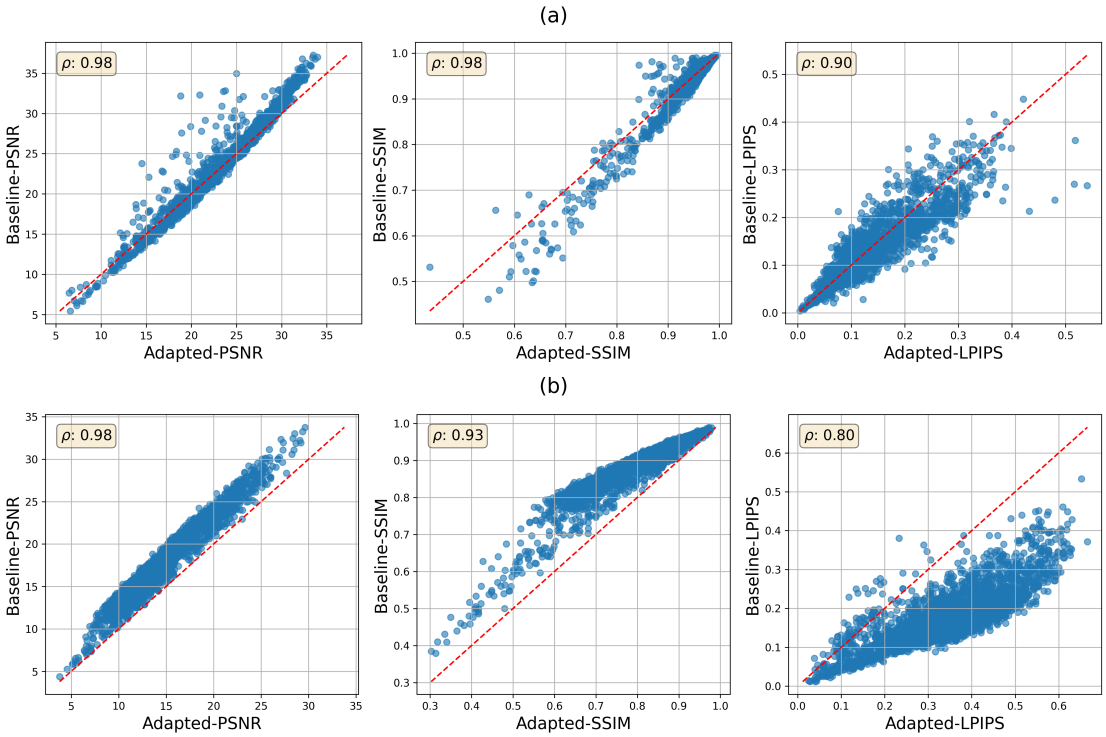

Figure 8.10: Performance obtained by super resolving the degraded images coming out of the generator. In (a), the corresponding SR model of each pipeline is used. In (b), a simple bicubic upsampling is used to super resolve the degraded images.

Fig 8.10 proves the relevance of the domain gap in super resolution, the SR model is able to estimate the inverse of the degradation function, if given the correct data. The problem relies on that in most experiments, the wrong degradation is shown to the model, forcing it to learn the inverse of an incorrect function. This plays an essential role when deploying super resolution model in real production environments, where the degradation model may not be known.

8.2 Target domain

This subsection will show the results from the experiments performed on the target domain, which is the equivalent of the red arrows flow described in fig. 5.4. In this case, the GAN trained for the degradation model is discarded and only the super resolution model is used. The input images are real FOREST-2 images, and the output images are super resolved versions of them. Due to the unpaired nature of the dataset, the performance of the SR model can not be evaluated using metrics like PSNR and SSIM. Other alternatives will be presented, and a qualitative analysis will be performed. Additionally, a quantitative analysis will be discussed using a very small sample of paired data obtained by synchronizing the overpass of FOREST-2 with the route of ECOSTRESS.

In Fig. 8.11, the super resolution models were used with a 264x264 pixels crop of a real FOREST-2 image as an input. The results show that the baseline model, trained with $\mathcal{D}_{\mathrm{SF-SF}}$, has very similar results to bicubic upsampling. On the other side, the

adapted model, trained using real FOREST images as the target domain ($\mathcal{D}_{\mathrm{SF-RF}}$), is able to recover more details, producing sharper images. In the frequency domain, the effects of super resolution are clear, frequency components of interest are amplified compared to bicubic upsampling, without over-amplifying higher frequencies usually related to noise.

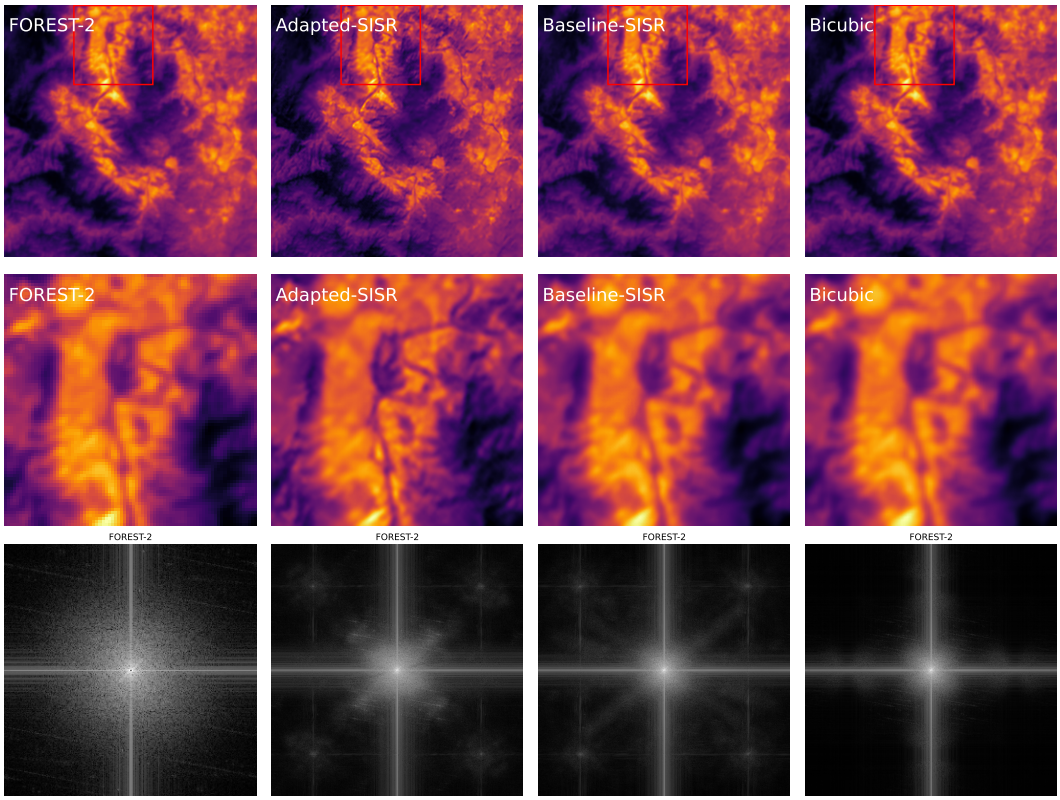

Figure 8.11: Super Resolved Forest-2 Scene using different SR models. In the upper row, the image is displayed. A detailed zoom is displayed below. The bottom row shows the log magnitude of the FFT for the images. The original image is displayed in the left, while the super resolved images are displayed afterwards.

Fig. 8.12 shows a more detailed analysis of the frequency domain of the SR images obtained by applying different SR models to the whole FOREST-2 validation dataset. In (a),the log magnitude of the FFT for the SR images is displayed, adding a shade that represent the interval of ± 1 standard deviations. Up until 0.3 cycles per pixel, the adapted model has a higher log magnitude than the baseline SR model or bicubic upsampling, also staying slightly higher in high frequency components. As higher frequencies are related to noise and artifacts, this suggests that the adapted model is able to recover more details than the baseline model, while minimizing undesired components. The amplification plot of the SR models against bicubic upsampling shows the same behaviour in a more clear way. Between 0.1 and 0.25 cycles per pixel, the amplification peaks between 6 and 8 dB on average with respect to bicubic, while the baseline model is between 0 and 2 dB. Such amplification, at a pixel size of 70m, corresponds to cycle frequencies between 300 $\frac{1}{m}$ and 700 $\frac{1}{m}$, which is consistent with the loss of components observed in 8.9. The variability of the amplification allows to conclude that this amplification is

consistently higher than the baseline-SISR along the dataset. On the other side, while the amplification is very similar in frequencies related to noise, the adapted model seems to step up a little bit compared to the baseline. This suggests that the adapted model is able to recover details from real FOREST-2 images, amplifying frequencies of interest, at the cost of a small increase in the overall noise of the image.

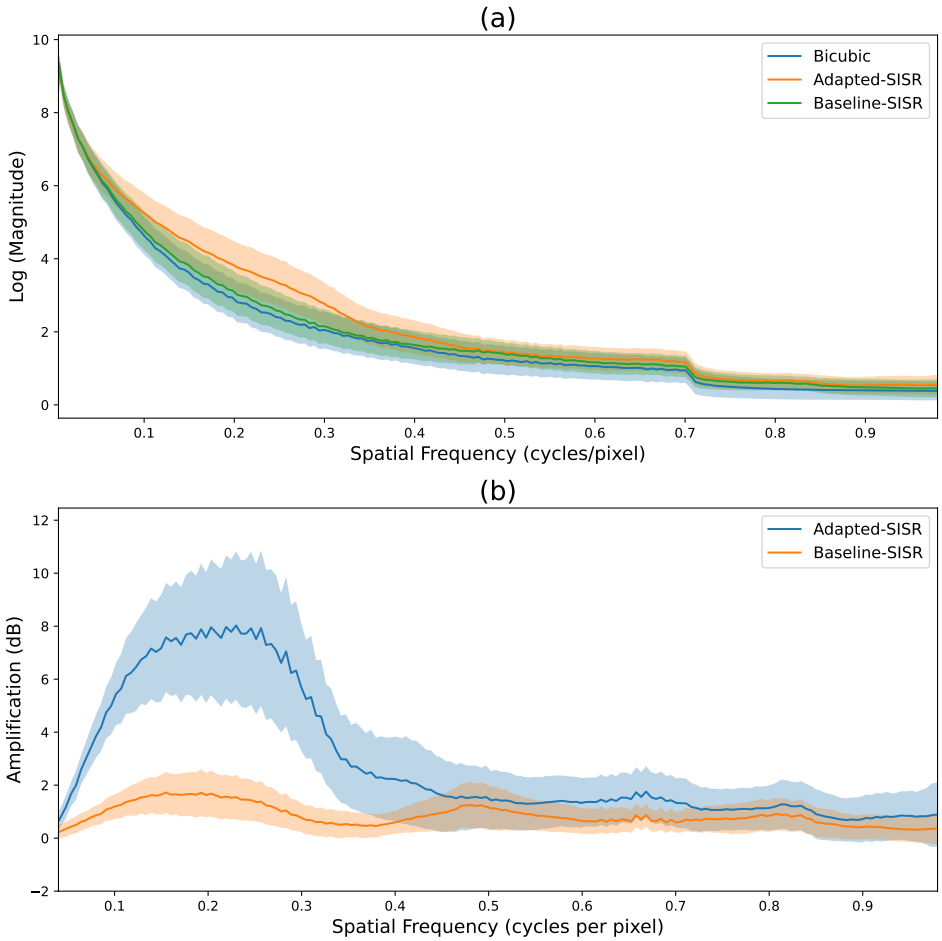

Figure 8.12: Frequency domain analysis of the SR images obtained by applying different SR models to the real FOREST-2 validation dataset. In (a), the log of the magnitude of the FFT for the SR images is shown, while in (b), the amplification with respect to a simple bicubic upsampling is displayed.

In Fig. 8.13, an example of the gradient analysis of the SR images is shown. Compared to the baseline SISR model, the adapted model shows higher gradient magnitudes, suggesting that the adapted model is able to recover more details than the baseline model. However, in the more dark sections of the gradient magnitude, some small background noise can be percieved, consistent with the results of the frequency domain analysis.

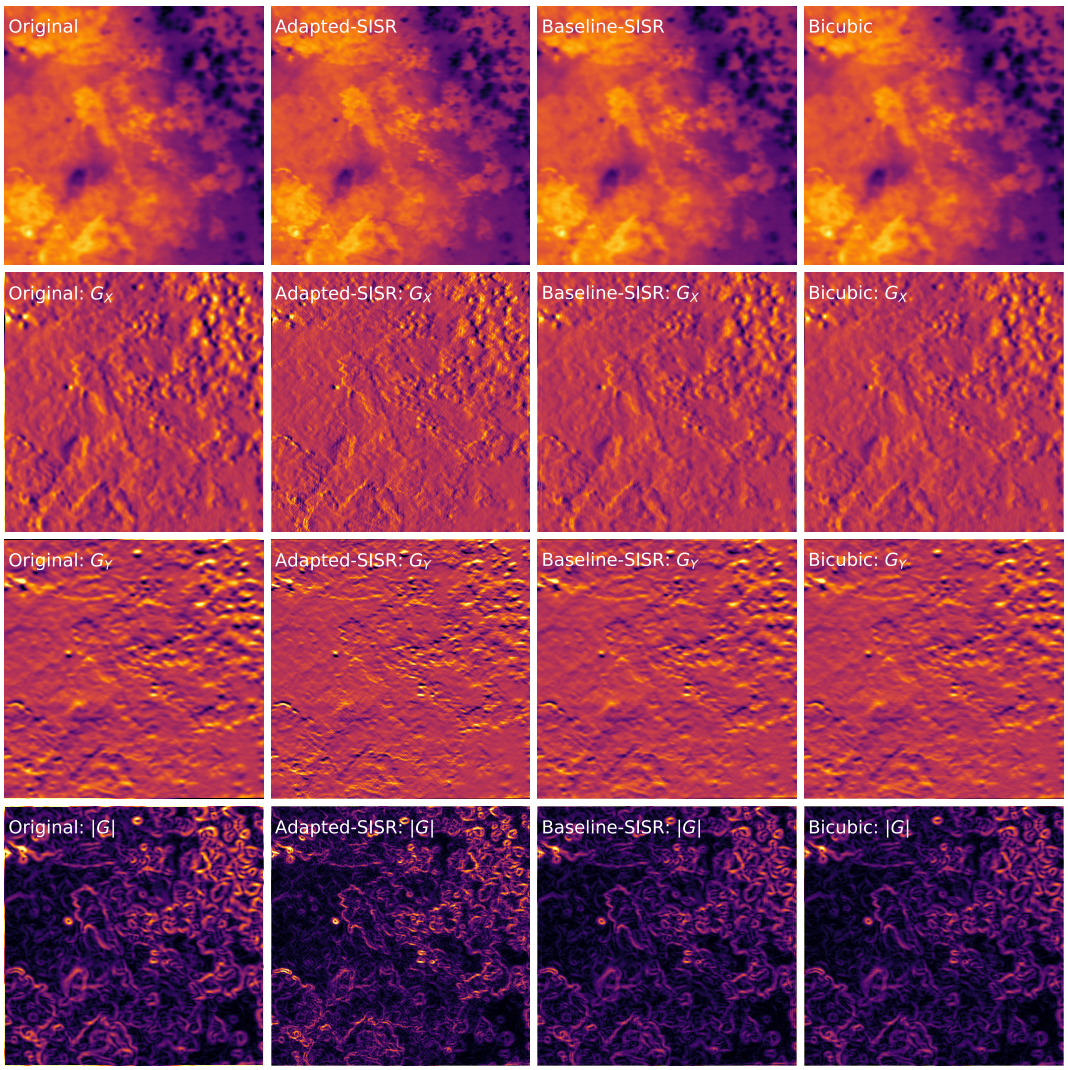

Figure 8.13: Gradient analysis of the super resolved images using different SR models for scenes coming from the real FOREST-2 validation dataset. In the upper row, the image is displayed. The gradients in the x and y direction (G_x and G_y respectiely) are displayed below. the gradient magnitude $|G|$ is displayed in the bottom row.

Fig 8.14 shows the distribution function of the gradient magnitudes of the whole validation dataset, estimated through a histogram. Both the adapted and the baseline model show a decrease in the number of pixels with low gradient magnitudes, suggesting that both models are able to recover more details than bicubic upsampling. However, the adapted model shows a higher number of pixels with high gradient magnitudes, implying that the adapted model is able to produce sharper edges than the baseline model. This is consistent with the observed results and the frequency domain analysis. However, it is important to note that the gradient magnitude is not a good measure of the performance of the SR model, as it does not take into account the noise and artifacts that may be present in the image. It represents only a complementary way to understand the effects of the SR model.

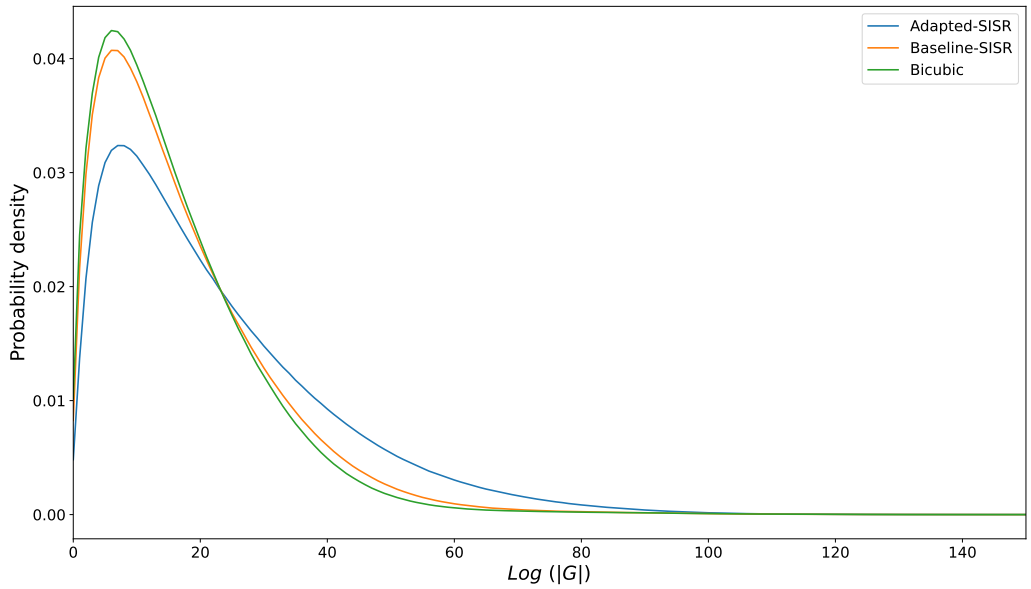

Figure 8.14: Histogram of the gradient magnitude $|G|$ for the whole validation real FOREST-2 dataset.

8.3 The domain gap goes both ways

The combination of the probabilistic degradation model and the SR model were proven helpful to bridge the domain gap and inprove the resolution of real FOREST-2 images. However, it is important to understand what happens when the target domain used in training does not match the conditions that will occur in the real world. While the common scenario is that the real degradation model is more complex than the one assumed in the dataset generation, the opposite can also occur. Assuming a more complex degradation model in the dataset could lead to LR images with more attenuation in higher frequency components, resulting in an SR model that "over-amplifies", producing noisy images with undesired artifacts. In this work, HR-LR pairs generated using a baseline degradation model exemplify an overly optimistic degradation scenario. When using the adapted SR model on these generated LR images, this scenario can be analyzed. As in this experiment the ground truth is known, the performance of the SR model can be evaluated using metrics like PSNR and SSIM.

The results are shown in Figs. 8.15 and 8.16. The performance of the adapted model on images with an optimistic degradation model is catastrophic, producing several artifacts and yielding a PSNR difference of approximately 10dB, which represent a 10x difference in terms of MSE.

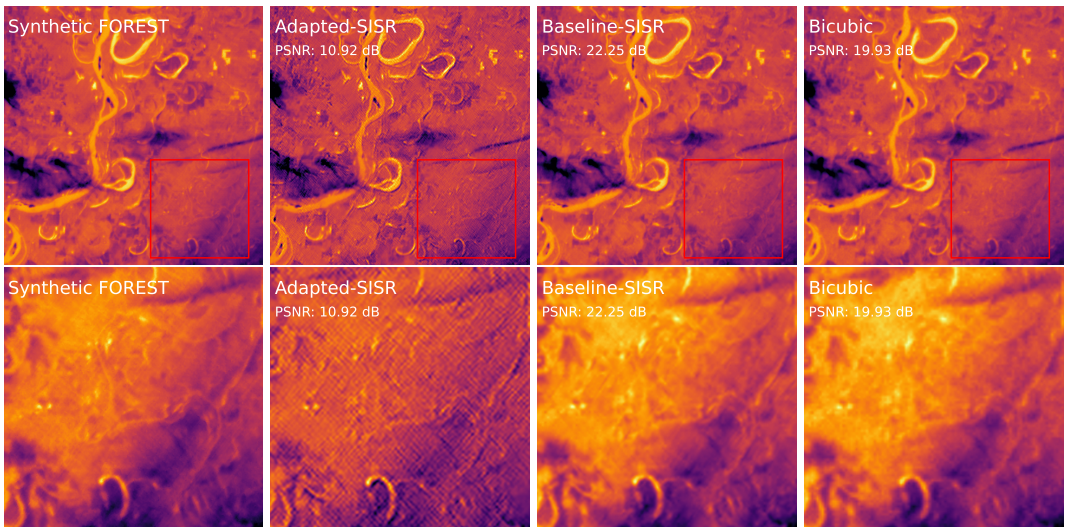

Figure 8.15: Effects of using a model trained with on different domain than at inference time. When using an Synthetic FOREST image degraded with the baseline degradation model as an input, the model trained using real FOREST-2 data as the target domain generates several artifacts and underperforms severly in terms of PSNR.

In the frequency domain, the results are shown in Fig. 8.16. The adapted model adds amplification in the higher range of spatial frequency, related with noise and artifacts. The frequencies of interest are also amplified. This suggests that while the adapted model highlights edges and details, it also severely amplifies the noise and artifacts, resulting in a worse performance in terms of PSNR.

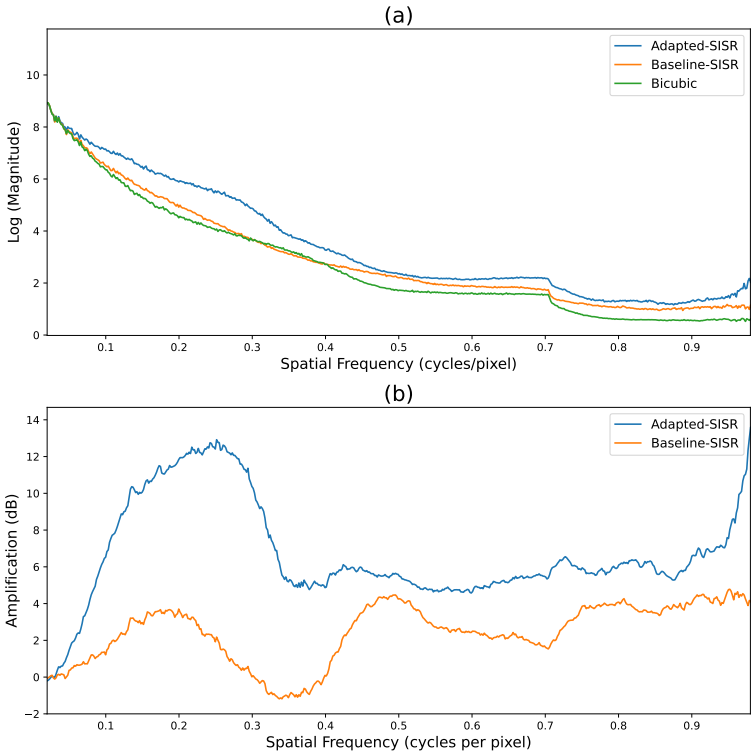

Figure 8.16: Effects of using a model trained with on different domain than at inference time. (a) shows the log magnitude of the radial average of the FFT for the SR images using different algorithms. (b) shows the amplification with respect to bicubic interpolation.

The performance results in terms of different metrics are shown in Fig. 8.17. In the conditions described above, the adapted super resolution model underperforms severly in every considered metric.

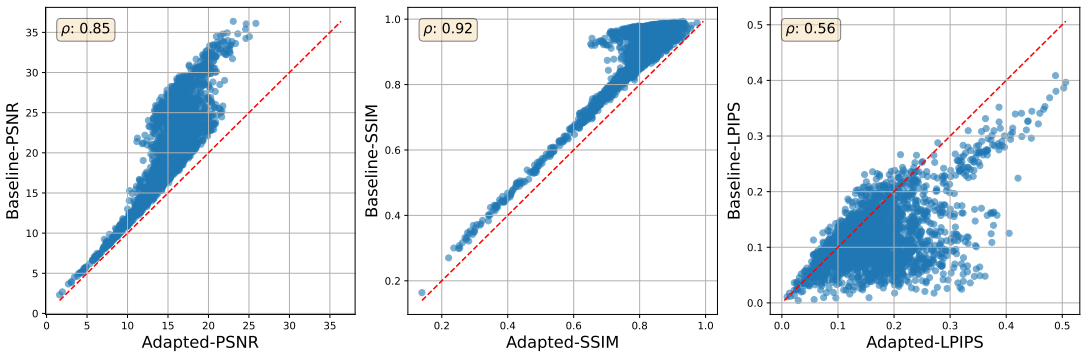

Figure 8.17: Performance obtained by super resolving the degraded synthetic FOREST images using different super resolution models.

This demonstrates that while this approach is very good to bridge a domain gap, it is not robust at all to domain shifts. This limitation is in sync with what is found in the literature seen in 4.4.3, implicit modelling for blind super resolution using GANs are not able to generalize to arbitrary domains not seen in the target domain.

8.4 Domain gap assessment using non-referenced image quality assessment

As in the target domain the ground truth is not known due to the lack of a paired dataset, the performance of the SR model can not be evaluated using metrics like PSNR and SSIM. Non-referenced image quality assessment (NR-IQA) metrics can help to understand the relative performance of the SR models when arbitrary LR images are used as an input.

The analysis was performed by taking the adapted and baseline SR models, trained on $\mathcal{D}_{\mathrm{SF-RF}}$ and $\mathcal{D}_{\mathrm{SF-SF}}$ respectively, and using them to super resolve synthetic degraded forest-2 images and real LR forest-2 images as an input. Then, the NIQE and BRISQUE scores are calculated.

The results are shown in Fig. 8.18. For both metrics, a large gap is observed between the adapted model and the rest when the input are real FOREST-2 data, suggesting that the adapted model is able to produce more natural images than the rest. This behaviour does not replicates when the input images come from $\mathcal{D}_{\mathrm{SF-SF}}$.

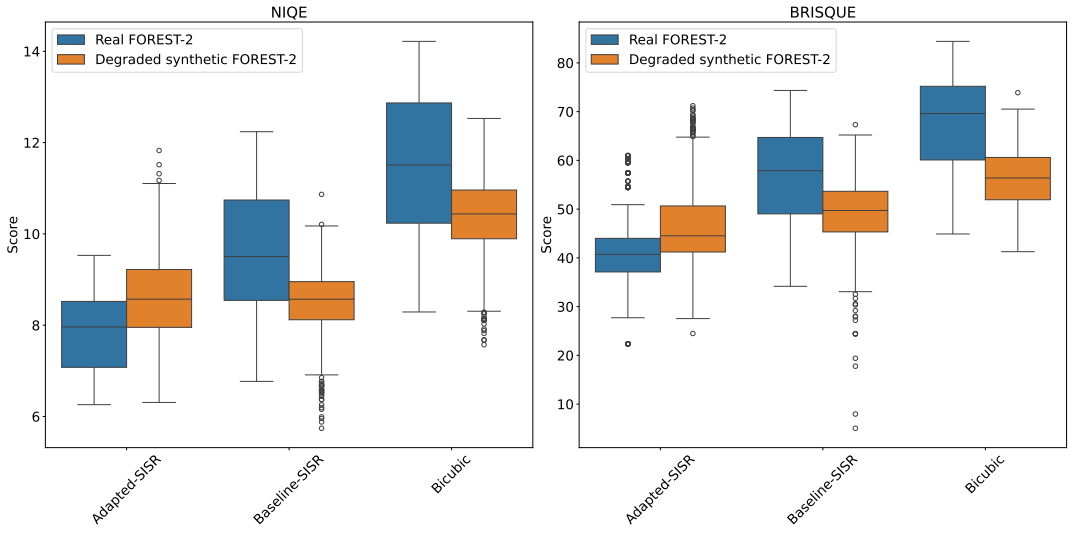

Figure 8.18: Image quality assessment metrics for the different SR models using different datasets as input. In both metrics, the lower the score, the better the image quality.

Moreover, for the adapted model, both metrics tend to get worse when the input images come from synthetic FOREST-2 images. The contrary happens for the rest of the models. This suggests that the SR model is able to produce more natural images only when the input images come from the same distribution as the target domain used in training.

However, it is important to note that:

1. NIQE and BRISQUE are calculated using a pre-trained model. The images used for the pre-training are not remote sensing images, and therefore, the results may not be representative. This could be circumvented by training a NIQE/BRISQUE model using with a more adequate dataset for the task.

2. NIQE and BRISQUE are a measure of image quality and naturalness, not physical consistency or reconstruction fidelity.

While the relative comparison of these results may help to understand the behaviour of the models, it is important to note that the results are not representative of their real world performance.

9 Conclusions and future work

- Flexible approach
- Models don't need to be very general, they just have to adapt between two very distinctive domains.
- Conditions of each missions are almost static
- Severe lack of paired data but abundance of unpaired data
- Just give me two datasets and the pipeline will find the way.

degradation model assumes complete independence between the noise and kernel components. It is a reasonable assumption but it may not be true in all cases.

The domain gap is not only being very optimistic when building the dataset. You can also be very pesimistic and lead to catastrophic results. Highlight on the difficulty of hand-picking the amount of degradation and the complexity of degradation modeling. Domain adaption seems to be very suitable due to:

Training of non-referenced image quality assesment for remote sensing MSSR More paired datasets would allow a more robust training of the model, because we could try to maximize the PSNR in those cases instead of how we do it now.

References

[1] J.-M Lefevre, C. Quentin, and Danièle Hauser. Land surface temperature retrieval techniques and applications : Case of the avhrr. *Measuring and Analysing the directional spectrum of ocean waves.*, 01 2005.

[2] C. O. JUSTICE J. R. G. TOWNSHEND. The 1 km resolution global data set: needs of the international geosphere biosphere programme. *International Journal of Remote Sensing*, 15(17):3417–3441, 1994.

[3] François Becker and Zhao-Liang Li. Becker f, li z. towards a local split window method over land surfaces. international journal of remote sensing. *International Journal of Remote Sensing - INT J REMOTE SENS*, 11:369–393, 03 1990.

[4] J.C. Jimenez-Munoz, J. Cristobal, J.A. Sobrino, G. Soria, M. Ninyerola, and X. Pons. Revision of the single-channel algorithm for land surface temperature retrieval from landsat thermal-infrared data. *IEEE Transactions on Geoscience and Remote Sensing*, 47(1):339–349, 2009.

[5] A. Karnieli Z. Qin and P. Berliner. A mono-window algorithm for retrieving land surface temperature from landsat tm data and its application to the israel-egypt border region. *International Journal of Remote Sensing*, 22(18):3719–3746, 2001.

[6] Zhao-Liang Li, Bo-Hui Tang, Hua Wu, Huazhong Ren, Guangjian Yan, Zhengming Wan, Isabel F. Trigo, and José A. Sobrino. Satellite-derived land surface temperature: Current status and perspectives. *Remote Sensing of Environment*, 131:14–37, 2013.

[7] N. Horning. Remote sensing. In Sven Erik Jørgensen and Brian D. Fath, editors, *Encyclopedia of Ecology*, pages 2986–2994. Academic Press, Oxford, 2008.

[8] United Nations Environment Programme. Spreading like wildfire: The rising threat of extraordinary landscape fires. https://www.unep.org/resources/report/spreading-wildfire-rising-threat-extraordinary-landscape-fires, 2021. Accessed: [2023-12-20].

[9] Christopher D Lippitt, Douglas A Stow, and Lloyd L Coulter. *Time-sensitive remote sensing*. Springer, 2015.

[10] Parwati Sofan, Fajar Yulianto, and Anjar Dimara Sakti. Characteristics of false-positive active fires for biomass burning monitoring in indonesia from viirs data and local geo-features. *ISPRS International Journal of Geo-Information*, 11(12), 2022.

[11] Atlantic Council. Extreme heat: Redefining business resilience for the climate crisis. Atlantic Council, August 2021.

[12] Angel Hsu, Glenn Sheriff, Tirthankar Chakraborty, et al. Disproportionate exposure to urban heat island intensity across major us cities. *Nat Commun*, 12(2721), 2021.

[13] K. Deilami, M. D. Kamruzzaman, and Y. Liu. Urban heat island effect: A systematic review of spatio-temporal factors, data, methods, and mitigation measures. *International journal of applied earth observation and geoinformation*, 67:30–42, 2018.

[14] A. A. Mohamed, J. Odindi, and O. Mutanga. Land surface temperature and emissivity estimation for urban heat island assessment using medium-and low-resolution space-borne sensors: A review. *Geocarto international*, 32(4):455–470, 2017.

[15] J. A. Sobrino, R. Oltra-Carrió, G. Sòria, R. Bianchi, and M. Paganini. Impact of spatial resolution and satellite overpass time on evaluation of the surface urban heat island effects. *Remote Sensing of Environment*, 117:50–56, 2012.

[16] B. Huang, J. Wang, H. Song, D. Fu, and K. Wong. Generating high spatiotemporal resolution land surface temperature for urban heat island monitoring. *IEEE Geoscience and Remote Sensing Letters*, 10(5):1011–1015, 2013.

[17] U.S. Geological Survey. Landsat Satellite Missions, 2023. Accessed: 2023-12-20.

[18] Terra – NASA's flagship earth observing satellite. https://terra.nasa.gov/, 2023. Accessed: [insert date here].

[19] W. Zhu, J. Sun, C. Yang, M. Liu, X. Xu, and C. Ji. How to measure the urban park cooling island? a perspective of absolute and relative indicators using remote sensing and buffer analysis. *Remote Sensing*, 13(16):3154, August 2021.

[20] Y. Shi, Y. Xiang, and Y. Zhang. Urban design factors influencing surface urban heat island in the high-density city of guangzhou based on the local climate zone. *Sensors*, 19(16):3459, August 2019.

[21] Chaobin Yang, Xingyuan He, Lingxue Yu, Jiuchun Yang, Fengqin Yan, Kun Bu, Liping Chang, and Shuwen Zhang. The cooling effect of urban parks and its monthly variations in a snow climate city. *Remote Sensing*, 9(10):1066, October 2017.

[22] John Wilson Rouse, Robert H. Haas, John A. Schell, and D. W. Deering. Monitoring vegetation systems in the great plains with erts. 1973.

[23] Gordana Kaplan, Ugur Avdan, and Zehra Yigit Avdan. Urban heat island analysis using the landsat 8 satellite data: A case study in skopje, macedonia. *Proceedings*, 2(7), 2018.

[24] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations, 2010.

[25] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, 2001.

[26] Diego Valsesia and Enrico Magli. Permutation invariance and uncertainty in multitemporal image super-resolution, 2021.

[27] Syed Muhammad Anwar Bashir, Yanning Wang, Murtaza Khan, and Yulei Niu. A comprehensive review of deep learning-based single image super-resolution, 2021.

[28] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution, 2015.

[29] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution, 2017.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[32] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network, 2017.

[33] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform, 2018.

[34] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[35] Marcus martens, Dario Izzo, Andrej Krzic, and Daniël Cox. Super-resolution of proba-v images using convolutional neural networks, 2019.

[36] Francesco Salvetti, Vittorio Mazzia, Aleem Khaliq, and Marcello Chiaberge. Multi-image super resolution of remotely sensed images using residual attention deep neural networks, July 2020.

[37] Andrea Bordone Molini, Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Deepsum: Deep neural network for super-resolution of unregistered multitemporal images, May 2020.

[38] John Kennedy, Ora Israel, Alex Frenkel, Rachel bar shalom, and Haim Azhari. Improved image fusion in pet/ct using hybrid image reconstruction and super-resolution, 01 2007.

[39] Christian Mollière, Julia Gottfriedsen, Martin Langer, Patricio Massaro, Christian Soraruf, and Matthias Schubert. Multi-spectral super-resolution of thermal infrared data products for urban heat applications, 2023.

[40] Andreas Lugmayr, Martin Danelljan, Radu Timofte, Namhyuk Ahn, Dongwoon Bai, Jie Cai, Yun Cao, Junyang Chen, Kaihua Cheng, SeYoung Chun, Wei Deng, Mostafa El-Khamy, Chiu Man Ho, Xiaozhong Ji, Amin Kheradmand, Gwantae Kim, Hanseok Ko, Kanghyu Lee, Jungwon Lee, Hao Li, Ziluan Liu, Zhi-Song Liu,

Shuai Liu, Yunhua Lu, Zibo Meng, Pablo Navarrete Michelini, Christian Micheloni, Kalpesh Prajapati, Haoyu Ren, Yong Hyeok Seo, Wan-Chi Siu, Kyung-Ah Sohn, Ying Tai, Rao Muhammad Umer, Shuangquan Wang, Huibing Wang, Timothy Haoning Wu, Haoning Wu, Biao Yang, Fuzhi Yang, Jaejun Yoo, Tongtong Zhao, Yuanbo Zhou, Haijie Zhuo, Ziyao Zong, and Xueyi Zou. Ntire 2020 challenge on real-world image super-resolution: Methods and results, 2020.

[41] Anran Liu, Yihao Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Blind image super-resolution: A survey and beyond, 2021.

[42] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution, 2018.

[43] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction, 2019.

[44] Zhengxiong Luo, Yan Huang, Shang Li, Liang Wang, and Tieniu Tan. Unfolding the alternating optimization for blind super resolution, 2020.

[45] Maria Zontak and Michal Irani. Internal statistics of a single natural image. In *CVPR 2011*, pages 977–984, 2011.

[46] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *2009 IEEE 12th International Conference on Computer Vision*, pages 349–356, 2009.

[47] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan, 2020.

[48] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning, 2017.

[49] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[50] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks, 2018.

[51] Zhengxiong Luo, Yan Huang, Shang Li, Liang Wang, and Tieniu Tan. Learning the degradation distribution for blind image super-resolution, 2022.

[52] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first, 2018.

[53] Yunxuan Wei, Shuhang Gu, Yawei Li, and Longcun Jin. Unsupervised real-world image super resolution via domain-distance aware training, 2020.

[54] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020.

[55] Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs, 2017.

[56] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018.

[57] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution, 2019.

[58] Simon Hook and Gerardo Rivera. ECOsystem Spaceborne Thermal Radiometer Experiment on Space Station (ECOSTRESS). https://ecostress.jpl.nasa.gov/instrument, 2023. Accessed: 28-November-2023.

[59] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[60] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer, 2013.

[61] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

[62] Dario Fuoli, Luc Van Gool, and Radu Timofte. Fourier space losses for efficient perceptual image super-resolution, 2021.

[63] Irwin Sobel and G. M. Feldman. An isotropic 3×3 image gradient operator, 1990.

[64] Jet Propulsion Laboratory. ECOSTRESS Fact Sheet, 2023. [Online accessed 28-November-2023].

[65] PhyTIR: Plant High Temperature InfraRed Viewer. https://phytir.jpl.nasa.gov/, 2023. [Online accessed 28-November-2023].

[66] Application for extracting and exploring analysis ready samples (AppEEARS). https://appeears.earthdatacloud.nasa.gov/, 2023. [Online; accessed 28-November-2023].

[67] Appeears api. https://appeears.earthdatacloud.nasa.gov/api/, 2023. [Online; accessed 28-November-2023].

[68] Land Processes Distributed Active Archive Center (LP DAAC). ECOSTRESS L1B Geolocated Radiance Data (ECO1BMAPRAD). https://lpdaac.usgs.gov/products/eco1bmapradv001/, 2023. [Online; accessed 28-November-2023].

[69] Ecostress faq. https://ecostress.jpl.nasa.gov/faq. Accessed: 2023-11-28.