



LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

MSc Data Science

MASTER THESIS

Summer Semester 23

MULTI-IMAGE SUPER RESOLUTION AND DOMAIN ADAPTATION
TECHNIQUES APPLIED TO THERMAL REMOTE SENSING

June 30, 2023

Student:

Massaro, Patricio

< p.massaro@campus.lmu.de >

Acknowledgments

Contents

1	Introduction	1
1.1	Storyline	1
1.2	Wildfire monitoring using thermal remote sensing	2
1.2.1	Spatio-temporal trade-off	2
2	Super resolution	2
2.1	Single-Image Super Resolution	3
2.1.1	SR Resnet	3
2.2	Multi-Image Super Resolution	3
2.2.1	Multi-spectral super resolution	3
2.2.2	Importance of interframe correlation	4
2.3	The domain gap problem	4
2.4	Blind image Super Resolution	5
3	Methodology	7
3.1	Baseline Degradation model	7
3.1.1	Blurring Kernel	7
3.1.2	Radiometric error correction	8
3.2	Models Architecture	10
3.2.1	SRResNet	10
3.2.2	RAMS	10
3.2.3	Probabilistic Degradation Model	10
3.3	Referenced image quality metrics	13
3.3.1	pixel-wise losses	13
3.3.2	Peak Signal-to-Noise Ratio (PSNR)	13
3.3.3	Structural Similarity Index (SSIM)	14
3.3.4	Learned perceptual image patch similarity (LPIPS)	14
3.3.5	Adjusting measures to a multi-image framework	15
3.4	Non-referenced Image quality metrics	15
3.4.1	Naturalness Image Quality Evaluator (NIQE)	15
3.4.2	Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE)	16
3.4.3	Frequency Domain Analysis	16
3.4.4	Gradient Distribution analysis	18
4	Datasets	20
4.1	Obtaining a high resolution dataset	20
4.1.1	The ECOSTRESS mission	20
4.1.2	Accessing ECOSTRESS Scenes	21
4.1.3	Selecting the best scenes	21
4.1.4	Data Processing	23
4.2	Obtaining FOREST-2 data	25
4.3	Datasets	25
4.3.1	Synthetic FOREST - Degraded Synthetic FOREST	25
4.3.2	Synthetic FOREST - real FOREST (Unpaired)	26

4.3.3	Synthetic FOREST- real FOREST (Paired)	26
5	Experiment Setup	27
5.1	Training	27
6	Results and discussion	28
6.1	Source domain	28
6.1.1	LR comparison	33
6.1.2	Effects of the degradation model in SR	35
6.2	Target domain	36
6.3	The domain gap goes both ways	40
6.4	Domain gap using non-references image quality assessment	43
7	conclusions and future work	44

1 Introduction

This thesis delves into the intricate realm of MISR and Domain Adaptation techniques, tailored for the domain of thermal remote sensing, with an emphasis on the challenges and innovations within the unpaired dataset context. Thermal imagery, with its unique sensitivity to temperature variations, offers an invaluable perspective for phenomena such as wildfire tracking and climate change studies. However, the native resolution of such thermal images often falls short of the detail necessary for fine-grained analysis, propelling the need for advanced SR methods.

1.1 Storyline

- Forest fires are dynamic events that change quickly over time. They could be monitored using remote sensing LWIR -*i* LST
- Spatio-temporal trade off: Big missions have high resolution but bad revisit frequency. This is where forest plays a role. But can we improve the resolution using post-processing techniques?
- Super resolution is an ill-posed problem, supervised deep learning techniques are generally used in the current literature.
- MISR leverages on subpixel differences present in several images taken from the same scene, potentially giving more information to generate an SR image. The potential increase in performance comes with the cost a data-processing overhead.
- One of the biggest challenges in super resolution is to create proper datasets for model training. Usually the degradation model used to generate HR-LR pairs is very simplistic compared to real cases. This problem is called domain gap.
- To bridge the gap, techniques like domain adaptation using gans could be used to estimate the degradation process and generate more realistic datasets, that will translate in better production-ready models.

This thesis covers three main questions:

- How does the performance of Multi-Image Super Resolution (MISR) compare to Single-Image Super Resolution (SISR), considering the pre-processing burden associated with MISR?
- How do traditional baseline degradation models, such as gaussian blurring, compare against a probabilistic model that aligns the distribution of a source domain (e.g., ECOSTRESS) to our target domain (FOREST-2)?
- What impact do the different degradation models have during the inference stage, when the SR models are used in real data.

1.2 Wildfire monitoring using thermal remote sensing

1.2.1 Spatio-temporal trade-off

2 Super resolution

Single image Super-Resolution (SR) is the task of increasing the resolution of a given image as well as sharpening its content by predicting the high-frequency component and the missing information

Super resolution refers to an image processing process of recovering a corresponding high-resolution image from a low-resolution version of it, with applications that range from natural images [1], [2] to satellite [3] and medical imaging [4]. 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.

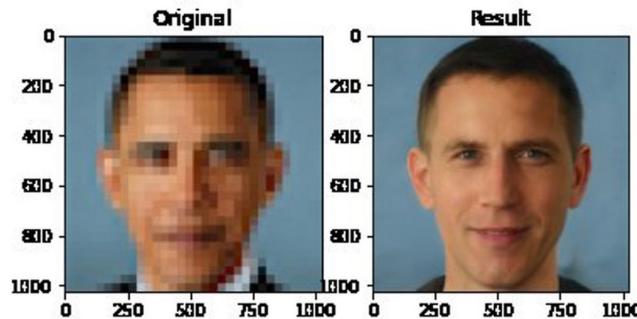


Figure 2.1: Example of super resolution as an ill posed problem. A blurry picture of Barack Obama can be generated from an HR image of another person.

Depending on the different number of low-resolution inputs, image super resolution reconstruction can be divided into single-image or multi-image. ASDFAGDSGADFS-GAFSFAD

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 understand the improvement of a super resolution network.

- what is super resolution?
- Interpolation based vs reconstruction-based vs learning-based

2.1 Single-Image Super Resolution

In a typical SISR framework, the LR image I^{LR} is modeled as follows:

$$I^{LR} = (I^{HR} * k) \downarrow_s + n \quad (1)$$

Where $I^{HR} * k$ 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. Super resolution objective is to solve this equation and obtain I^{HR} which as stated before, is an extremely ill-posed problem. Super resolution was first proposed in the 1960s, while the first use of multiple images dates of 1989. 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 extraction and representation. The rise of the technology wave around 2010 changed the way of solving problems in different branches. Instead of piecing together individual functional modules to form a system, the focus is to optimize parameters by global training after the whole system is designed, what is called end-to-end training.

2.1.1 SR Resnet

2.2 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 extensive 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.

In 2019, the European Space Agency (ESA) organized an SR challenge [5] 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.

2.2.1 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.

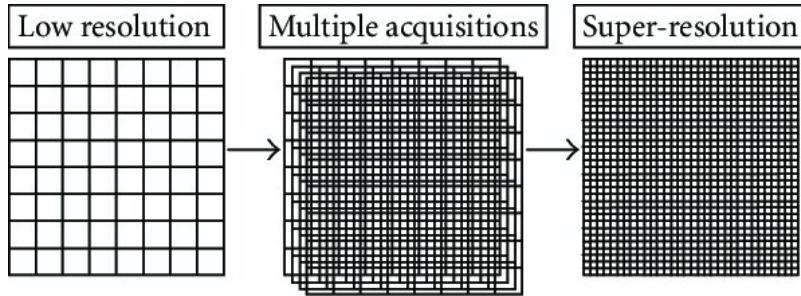


Figure 2.2: Multi-image super resolution algorithms combine multiple low-resolution image acquisitions into a high-resolution image. Source: [6]

In the context of MSSR, each spectral band, corresponding to a specific wavelength 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 reasonable approach.

Traditional pan-sharpening algorithms could be considered as deterministic MSSR algorithms. It is usually used to increase the resolution of a multi-spectral RGB image using the panchromatic band. The overlap between the wavelengths of both wavelength ranges makes this algorithm straightforward. However, it is ill-suited for Thermal Infrared (TIR) data due to the disjointed spectral domains of the visible and TIR bands. In [7], A deep learning 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 incorporated enhanced spatial resolution details captured from the visible bands.

2.2.2 Importance of interframe correlation

2.3 The domain gap problem

SR is a supervised problem, the super resolved image is compared to the HR ground truth and the pixel-by-pixel differences drive the gradients of the neural network to minimize the loss, in a fully supervised manner. Additionally, deep learning based SR methods are known to consume large quantities of training data. 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, such as bicubic. 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 [8].

The domain gap problem is the difference between the training data and the real-world data. The training data is usually generated synthetically, using bicubic down-sampling, which is a very simple degradation model. However, real-world degradations are usually too complex to be modelled with an explicit combination of multiple degradation types, as shown in Fig.3(c). Therefore, implicit modelling attempts to circumvent the explicit modelling function. Instead, it defines the degradation process f implicitly

through data distribution, and all the existing approaches with implicit modelling require an external dataset for training. Typically, these methods utilize data distribution learning with Generative Adversarial Network (GAN) [16] to grasp the implicit degradation model possessed within training dataset, like CinCGAN [8]. Great attempts have been made in the last several years on the generation of real pairs of LR-HR images, but the process remains costly. Additionally, SR networks trained on the collected datasets are not able to generalize to images captured in other conditions.



Figure 2.3: Effects of different degradation models on one HR image. Source: [9]

2.4 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 2.3. A schematic diagram of the problem is shown in Fig. 2.4. Non-blind SR 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 downsampled LR image, and thus the non-blind SR methods are not successful, moving to a different place in the HR space than the natural images. 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.

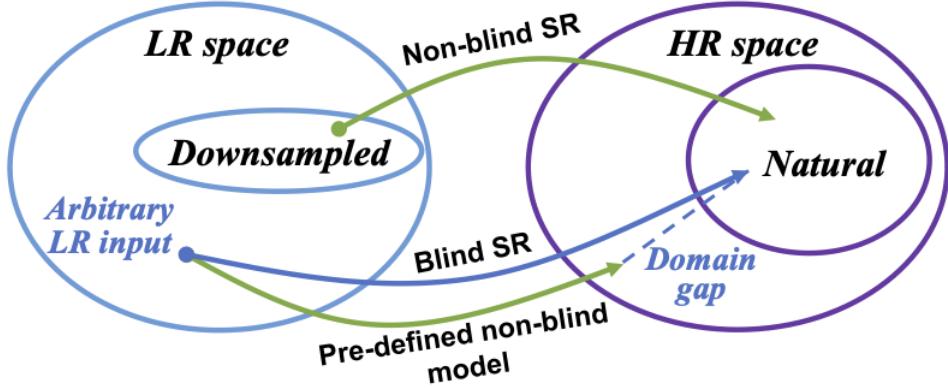


Figure 2.4: Domain interpretation of differences between non-blind and blind SR. Source: [9]

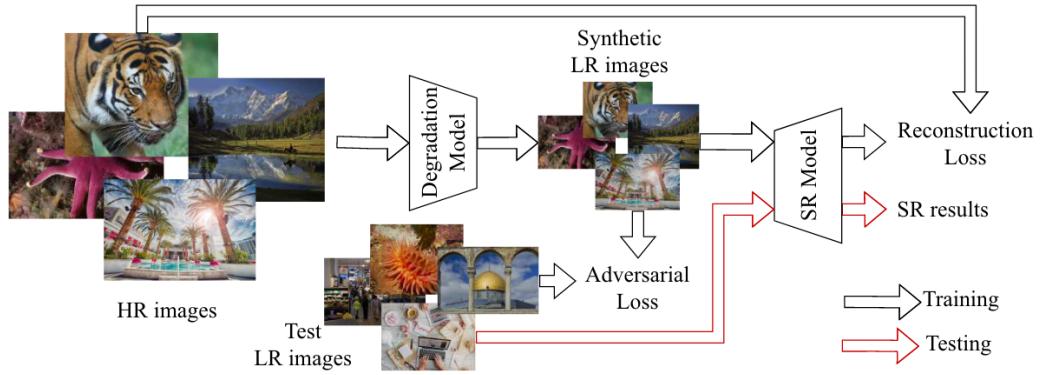


Figure 2.5: In Degradation-learning-based methods, adversarial training is used to encourage the degradation model to produce images in the same domain as the test LR images. After training, the SR model is directly used to super resolve the inputs. Source [10].

3 Methodology

3.1 Baseline Degradation model

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 [11]. 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 degradation model, which includes a set of multiple blurring kernels and several random noises configurations. The larger degradation space grants these models better generalization capabilities and 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.

3.1.1 Blurring Kernel

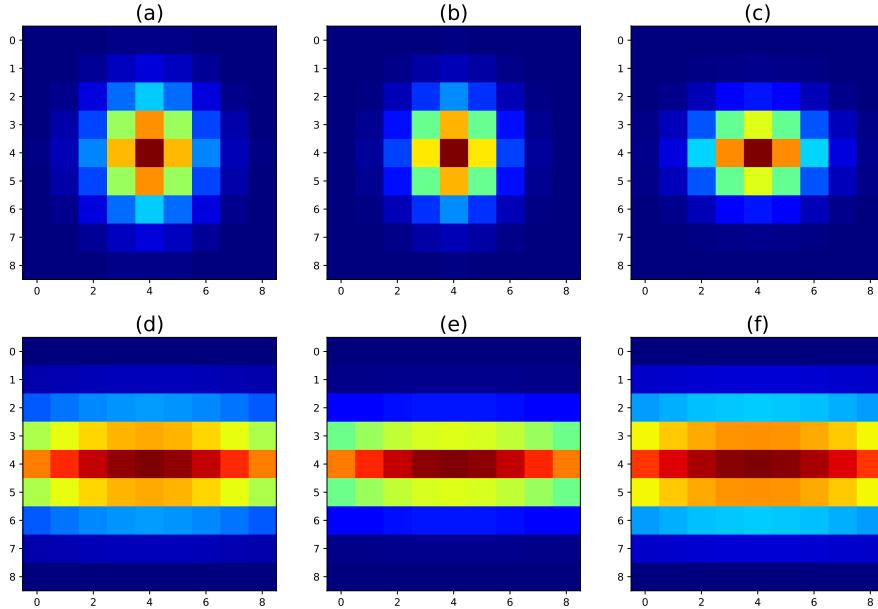


Figure 3.1: Example of kernels used in a stochastic degradation model. (a),(b) and (c) are generated using a symmetric variance on the x and y axis. (d) (e) and (f) are generated using an asymmetric variances, resulting in much more anisotropic kernels.

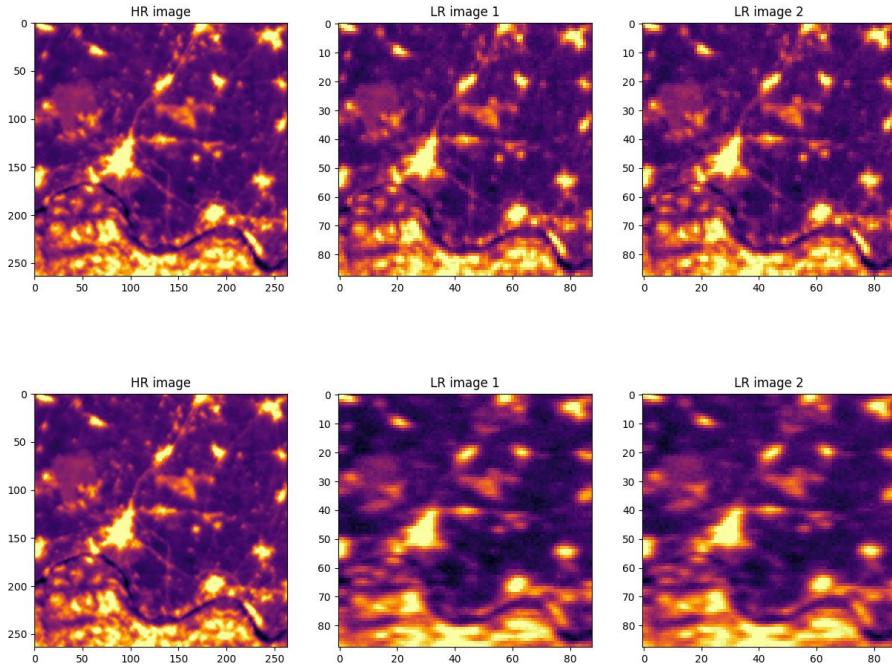


Figure 3.2: Effects of different blurring kernels on the HR-LR generation. The upper row contains images generated using blurring kernels with symmetric distributions. The lower rows contain images generated using asymmetric distributions for the variances, resulting in highly anisotropic kernels.

3.1.2 Radiometric error correction

As reported by the ECOSTRESS instrument sheet,[12] its nominal radiometric accuracy at 300K is 0.5K. FOREST-2 target radiometric accuracy is 1K. This difference in accuracy should be taken into account. To align these accuracies, we first calculate the additional error required using the following equation:

$$e_{\text{forest}} = \sqrt{e_{\text{eco}}^2 + e_{\text{extra}}^2} \quad (2)$$

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:

$$\frac{\partial B}{\partial T} = \frac{B(\lambda, T + \delta T) - B(\lambda, T)}{\delta T} \quad (3)$$

By multiplying the results of equations 2 and 3, 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/m²/μm and 1.1444×10^{-1} W/sr/m²/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:

$$\begin{aligned} \varepsilon_{\text{final}} &= (1 - \phi) \times \varepsilon_{\text{radiance}} + \phi \times \eta \times \varepsilon_{\text{radiance}} \\ \eta &\sim \mathcal{N}(0, 1) \end{aligned} \quad (4)$$

The effects of the error correction is shown in Fig. 3.3. As the target radiometric error increases with respect to ECOSTRESS scenes, the loss of information is more noticeable.

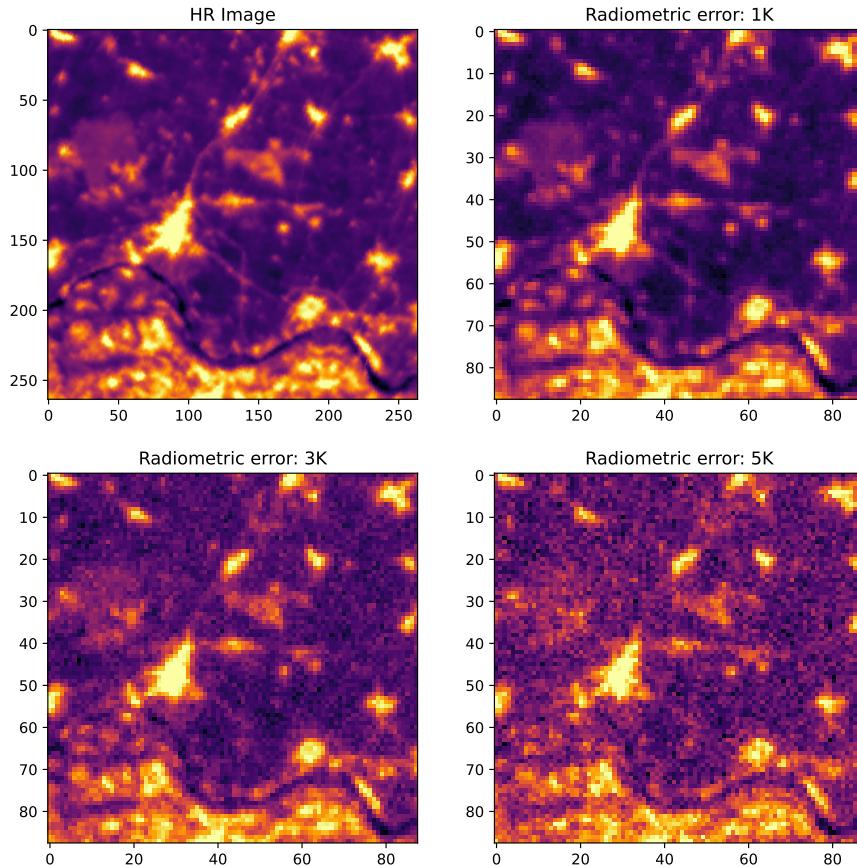


Figure 3.3: Effects of different radiometric error corrections on the HR-LR generation.

3.2 Models Architecture

3.2.1 SRResNet

Introduced in 2017 [13], SRResnet leverages on residual networks [14] that employ skip connections to solve the super resolution task. The architecture is detailed in Fig. CITE. Specifically, 16 residual blocks consisting two convolutional layers, followed by batch-normalization layers and ParametricReLU activation functions. The convolutional layers have 3×3 kernels and 64 feature maps. To increase the resolution of the input image, two trained sub-pixel convolution layers are used.

As this work focuses on having super resolved images with high physical consistency and not on the perceptual superiority of the images, improvements introduced in the publications like the Generative Adversarial Network (SRGAN) and the perceptual loss for gradient calculation are not used.

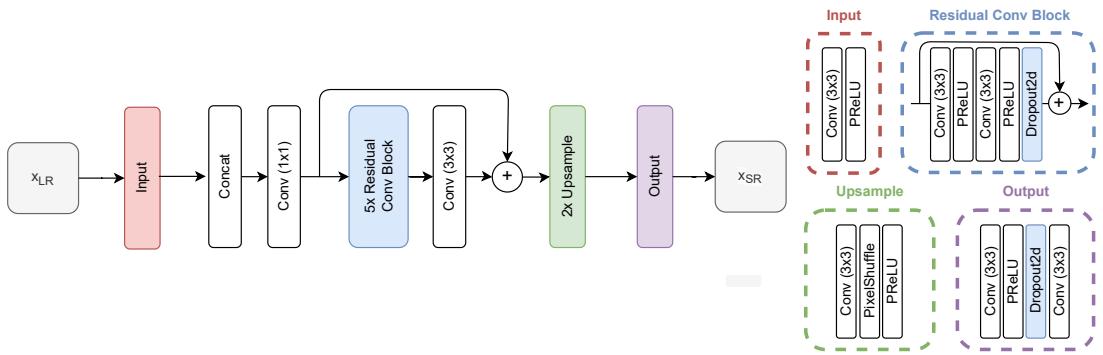


Figure 3.4: SRResNet architecture. X_{LR} represents the low resolution input image, X_{SR} the super resolved image, which is then compared to the ground truth X_{HR} .

3.2.2 RAMS

3.2.3 Probabilistic Degradation Model

To avoid the domain gap between synthetic and test images, most previous methods try to adaptively learn the degradation process via a deterministic model. However, some degradations in real scenarios are stochastic and cannot be determined by the content of the image. These deterministic models 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 paper, we propose a probabilistic degradation model (PDM), which studies the degradation D as a random variable, and 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 thus prevent the SR model from over-fitting to specific ones.

Most previous degradation-learning-based SR methods have a common drawback: their degradation models are deterministic, and each HR image can only be degraded to a certain LR image. It implies 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 or blur caused by random shakes of cameras. 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.

We parameterize the degradation with two random variables, i.e., the blur kernel k and random noise n , by formulating the degradation process as the linear function from Eq. 1. It can be divided into two linear steps [15]:

$$\begin{aligned} I_{\text{clean}}^{\text{LR}} &= (I^{\text{HR}} * k) \downarrow_s \\ I^{\text{LR}} &= I_{\text{clean}}^{\text{LR}} + n \end{aligned} \quad (5)$$

The model is trained in a adversarial framework, and the distribution of the LR image could be automatically learned during the training

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 .

$$p_D(D) = p_{k,n}(k, n) = p_k(k)p_n(n). \quad (6)$$

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 :

$$k = \text{net}K(z_k), \quad z_k \sim \mathcal{N}(0, 1), \quad (7)$$

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

$$z_k \in \mathbb{R}^{f_k \times h \times w}, \quad k \in \mathbb{R}^{(k \times k) \times h \times w}, \quad (8)$$

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 appropriate 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:

$$k = \text{net}N(z_n), \quad z_n \sim \mathcal{N}(0, 1), \quad (9)$$

$$z_n \in \mathbb{R}^{f_n \times h \times w}, \quad n \in \mathbb{R}^{h \times w \times c}, \quad (10)$$

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 [16], 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.

$$n \sim \mathcal{N}(0, \sigma_{\text{read}} + \sigma_{\text{shot}} \cdot I_{\text{clean}}^{\text{LR}}), \quad (11)$$

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

$$k = \text{net}N(z_n, I_{\text{clean}}^{\text{LR}}), \quad z_n \sim \mathcal{N}(0, 1), \quad (12)$$

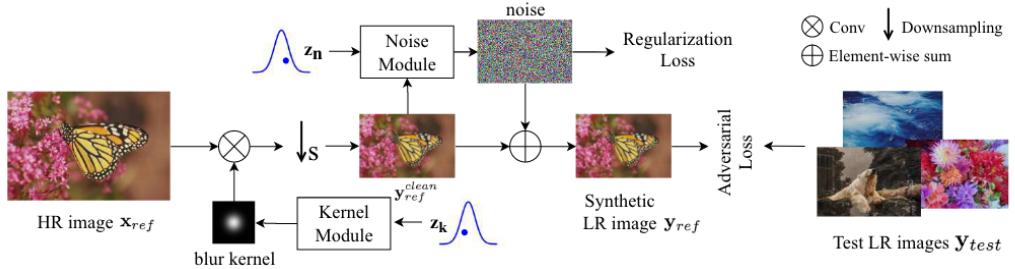


Figure 3.5: Schematic of the probabilistic degradation module. The discriminator is left out for a more intuitive description

The probabilistic degradation model is optimized via adversarial training, which encourages the output of the generator to be similar with the test images [17]. 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.

$$l_{\text{total}} = l_{\text{adversarial}} + 100\|\mathbf{n}\|_2^2. \quad (13)$$

PDM is able to model more diverse degradations. It allows one HR image to be degraded into multiple LR (4) the priori knowledge about the degradations can be easily incorporated into PDM, which may encourage it to learn the degradations better.

in our method, since PDM is better constrained and easier to be trained, it works well to train PDM and the SR model simultaneously. In this way, PDM can be integrated with any SR model to form a unified framework for blind SR, which is called PDM-SR

At last, PDM 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 could focus on learning the degradations. In most previous methods, to ensure that y_{ref} has consistent content with x_{ref} , it is usually guided by a bicubically downsampled version of x_{ref} [4, 40]. However, this guidance may be inappropriate, especially when the test images are heavily blurred.

Instead, in our PDM, due to the well-constrained blur kernels and noises, the content of y_{ref} is inherently consistent with x_{ref} . As a result, PDM could avoid the limitation of extra guidance and focus on learning the degradations.

In [12] and [40], the training of the degradation model and the SR model are separate, i.e. they firstly train a degradation model and then use the trained degradation model to help train the SR model. This two-step training method is time-consuming but is necessary for their method, because their highly nonlinear degradation models will produce undesirable results at the beginning of the training, which may mislead the optimization of the SR model. However, in our method, since PDM is better constrained and easier to be trained, it works well to train PDM and the SR model simultaneously. In this way, PDM can be integrated with any SR model to form a unified framework for blind SR, which is called PDM-SR

3.3 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 built 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.

3.3.1 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 network gradients during training. In a general form, the L_1 and L_2 losses are defined as follows:

$$\mathcal{L}_{L_k} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|^k \quad (14)$$

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.

3.3.2 Peak Signal-to-Noise Ratio (PSNR)

Peak Signal-to-Noise Ratio (PSNR) measures the quality of a reconstructed or super-resolved image in comparison to the original high-resolution image. It quantifies 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 the original and reconstructed images, 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:

$$PSNR = 10 \cdot \log_{10} \left(\frac{I_{MAX}^2}{\mathcal{L}_2} \right) \quad (15)$$

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

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.

3.3.3 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:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (16)$$

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.

3.3.4 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 [18]) 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:

$$LPIPS(I, I') = \sqrt{\sum_{i=1}^N w_i \|f_i(I) - f_i(I')\|^2} \quad (17)$$

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.

3.3.5 Adjusting measures to a multi-image framework

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 [5]. 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. 18

$$\mathcal{L}^*(I^{HR}, I^{LR}, d) = \min_{u,v \in [0,2d]} \mathcal{L}(I_{u,v}^{HR}, I_{u,v}^{SR}) \quad (18)$$

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:

$$\begin{aligned} \mathcal{L}^*(I^{HR}, I^{LR}, d) &= \min_{u,v \in [0,2d]} \mathcal{L}(I_{u,v}^{HR}, (I_{u,v}^{SR} + b)) \\ b &= \frac{1}{(W-d)(H-d)} \sum_{x,y} (I_{u,v}^{HR} - I_{u,v}^{SR}) \end{aligned} \quad (19)$$

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

3.4 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 distortion may alter these statistics [19]. 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. 3.6.

3.4.1 Naturalness Image Quality Evaluator (NIQE)

The Naturalness Image Quality Evaluator (NIQE) [19] is a no-reference image quality assessment metric that quantifies the perceptual quality of images based on their

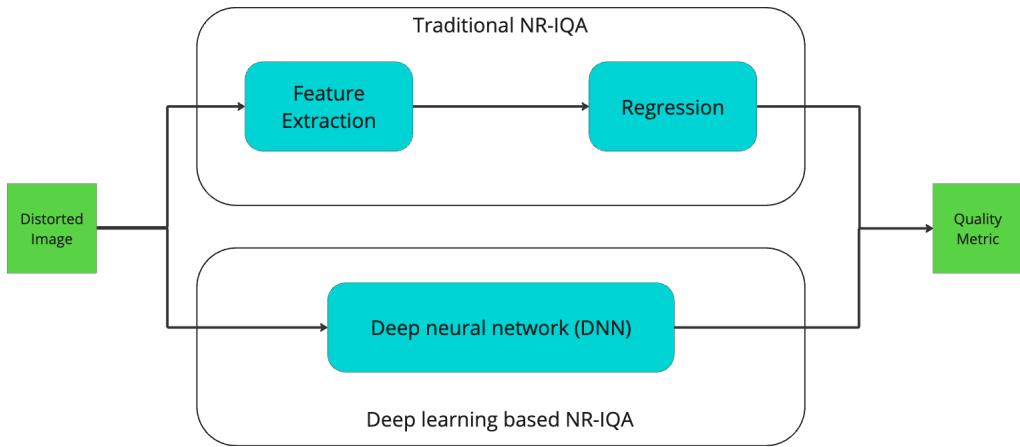


Figure 3.6: Workflow of a NR-IQA model.

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 an objective 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.

3.4.2 Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE)

3.4.3 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 downsampled 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 [20]. 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:

$$f(r) = \frac{r}{N} \cdot f_{\text{Nyquist}}, \quad (20)$$

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.

Through FFT we acquire a depiction of the frequency-based amplification or attenuation 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 for any noise or artifacts generated by the SR, and should be used in combination to other supervised metrics.

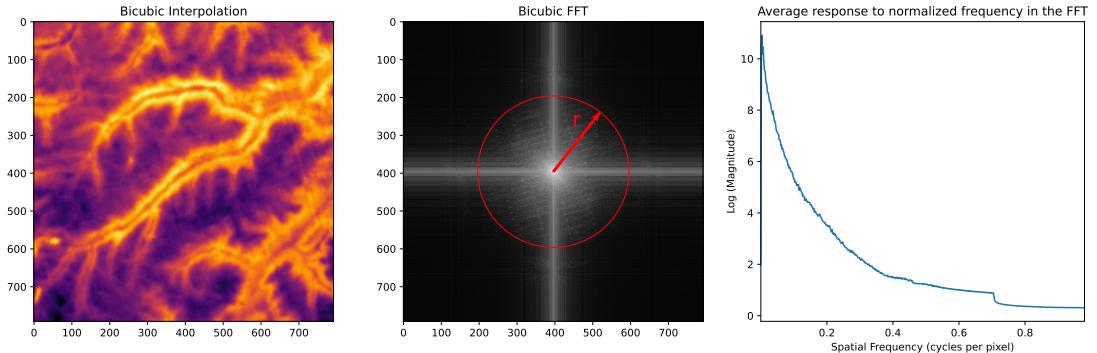


Figure 3.7: Steps of the frequency domain analysis. The Center image shows the log magnitude of the shifted FFT of a bicubic upsampled FOREST scene and an example of a radial profile, the average of all the points that have the same r is calculated. The right image displays the log magnitude obtained for every radial profile, translated into spatial frequency.

3.4.4 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. 21 [21]. These kernels are designed to respond maximally to edges running vertically and horizontally relative to the pixel grid.

$$\hat{G}_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \quad \hat{G}_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad (21)$$

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:

$$|G| = \sqrt{G_x^2 + G_y^2} \quad (22)$$

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.

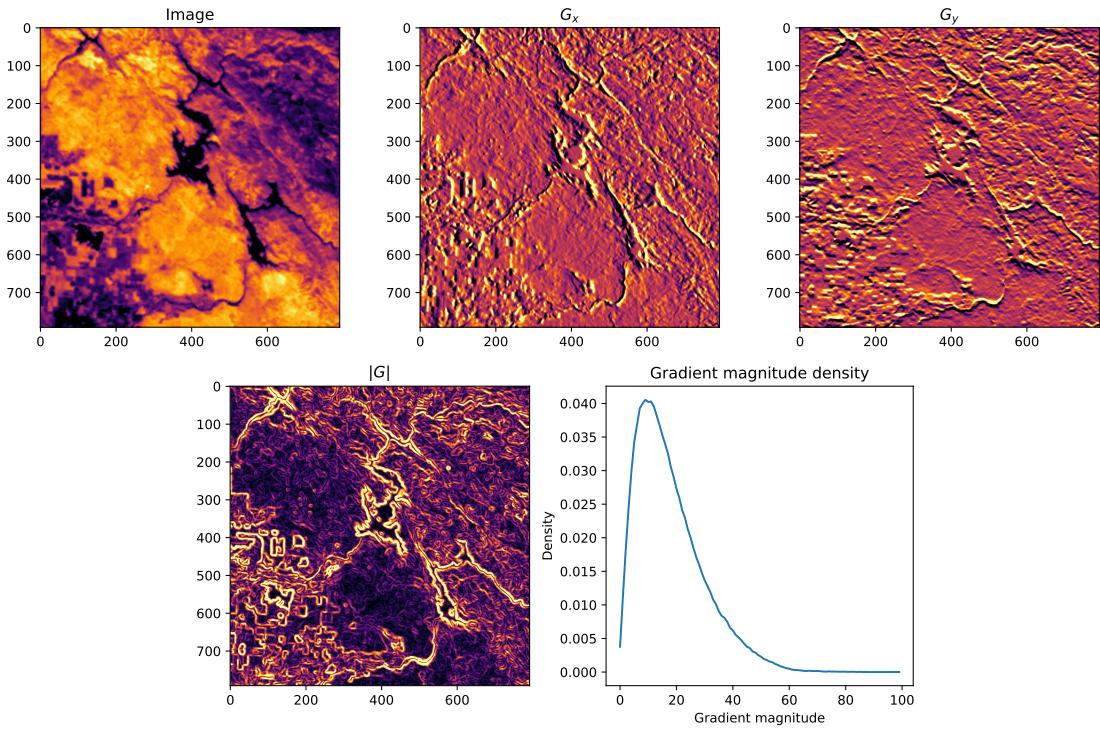


Figure 3.8: Steps to obtain a gradient magnitude density. Using the sobel operators, G_x and G_y are obtained from an image. The magnitude $|G|$ of each pixel is calculated using Eq. 22. The density can be estimated afterwards, using 100 bins in this case.

4 Datasets

4.1 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.

4.1.1 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 [22], 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 HypIRI Thermal Infrared Radiometer [23] 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 [22]. 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. 4.1 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.

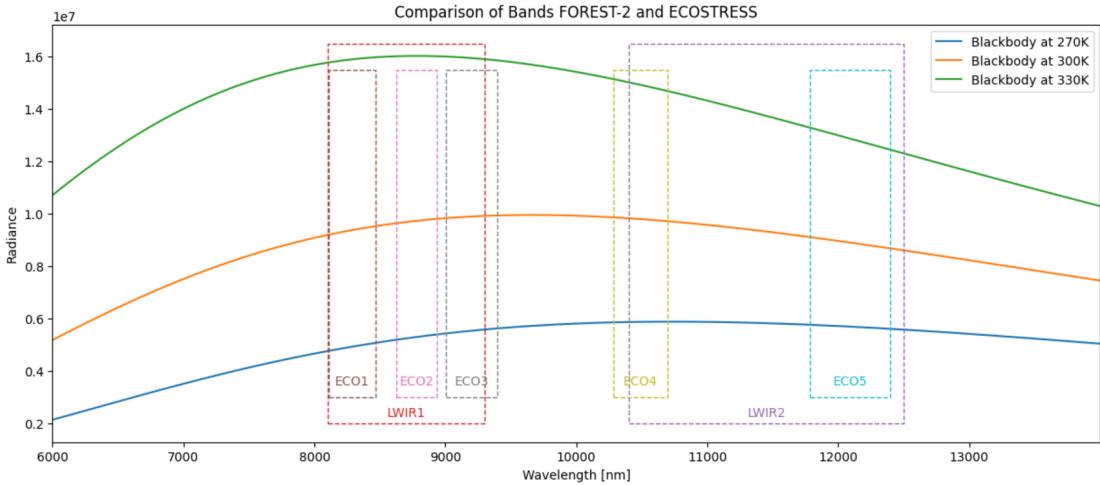


Figure 4.1: Wavelengths of the sensors in Ecostress and Forest satellites. The radiation spectrum of black-bodies at different temperatures are included for comparison.

4.1.2 Accessing ECOSTRESS Scenes

ECOSTRESS imagery is available via NASA’s Application for Extracting and Exploring Analysis Ready Samples (AppEEARS) [24]. This tool allows the request of area samples via vector polygons. Using the product’s API [25], Level 1 Mapped Radiance scenes of size 200x200 km with center on the locations provided in Fig. 4.1 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 [26].

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 4.1: Requests configuration

4.1.3 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 [27], the accuracy of radiance measurements is highly dependent on clear sky conditions; cloudy scenes typically yield negligible radiance emissions.

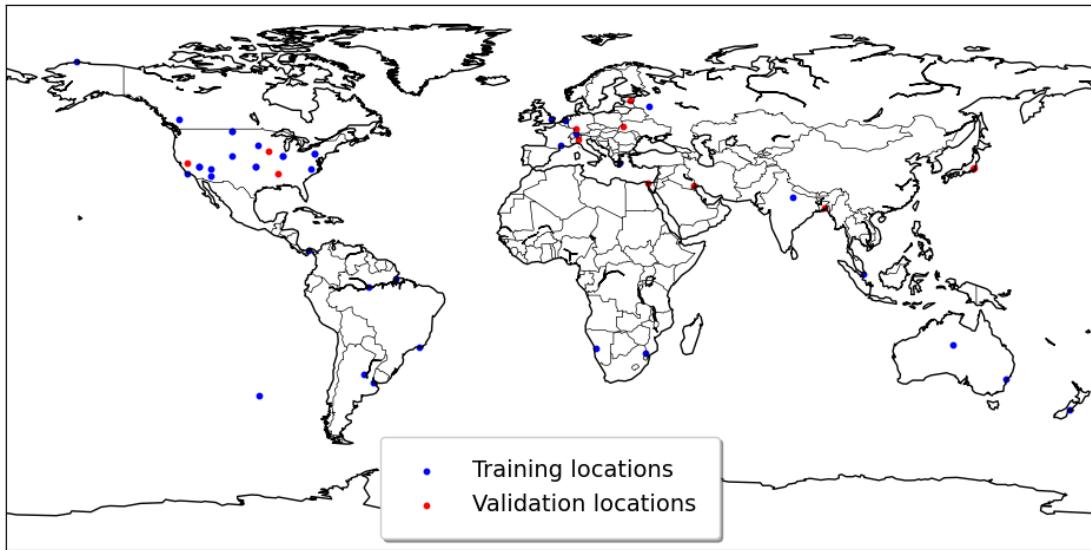


Figure 4.2: Location of the samples taken from ecostress.

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.

- 1: **QA statistics:**
 - 2: Get the average proportion of good pixels p_{gp} for the 5 radiances of the scene.
 - 3: Discard scenes where $p_{gp} < 60$.
 - 4: **Scene Statistics:**
 - 5: Get the biggest scene of each month.
 - 6: Calculate the proportion between the size each scene and the biggest of the month.
 - 7: Discard images which size proportion is smaller than 0.2.
 - 8: Calculate the median of the radiance values of the scene.
 - 9: **Selecting the scene of the month:**
 - 10: Merge the QA statistics and the Scene statistics.
 - 11: For each month, get the 3 scenes with the greatest p_{gp} .
 - 12: 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.

4.1.4 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. 4.3 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 its 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 4.2: 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. 4.4. 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.

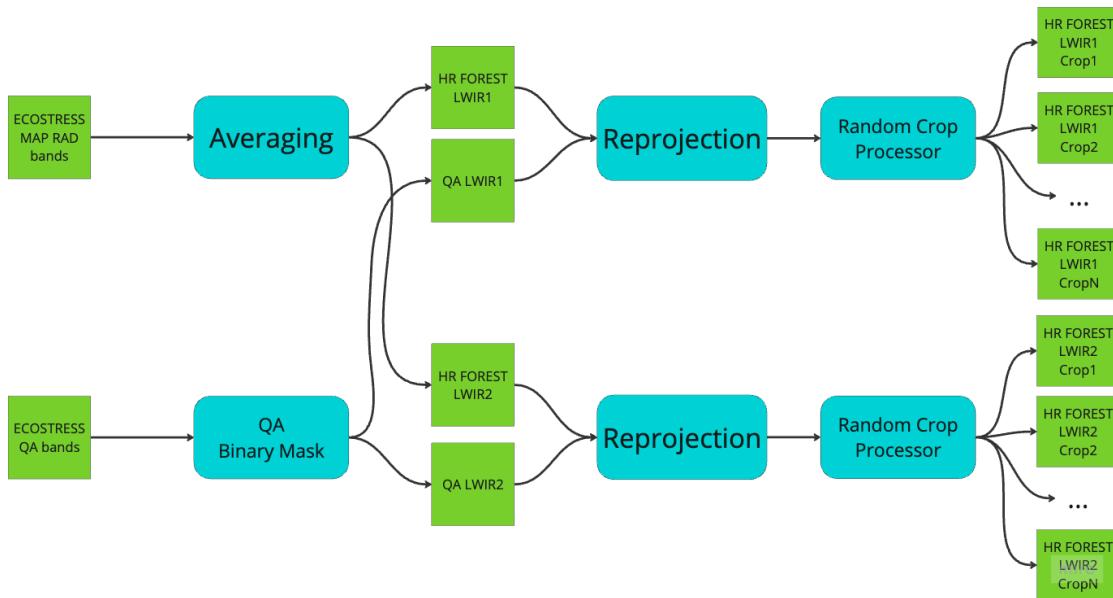


Figure 4.3: Data processing workflow

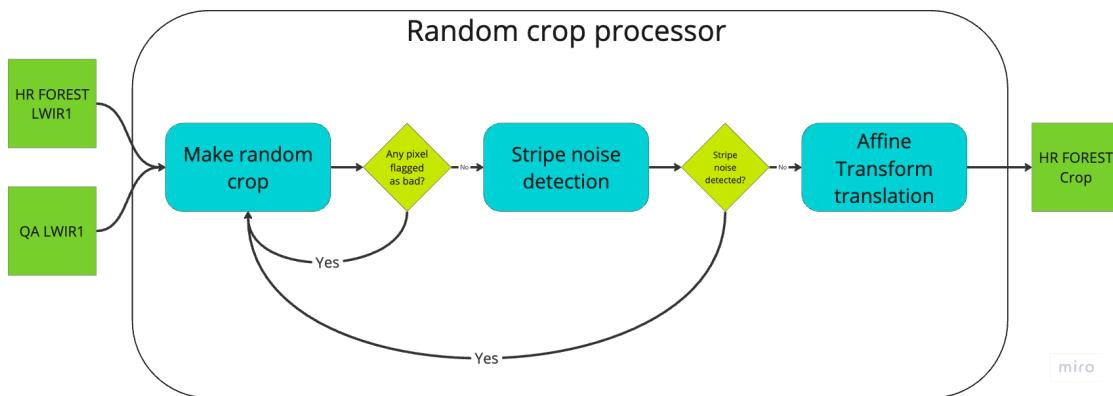


Figure 4.4: Random crop processor

4.2 Obtaining FOREST-2 data

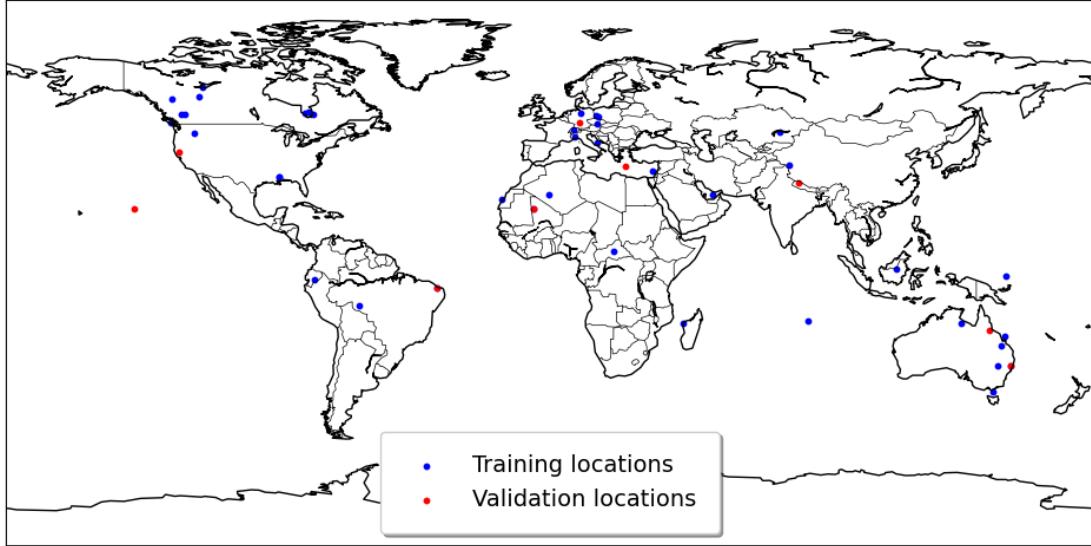


Figure 4.5: Location of the FOREST-2 scenes.

4.3 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 sources. 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}_{SF-SF}				\mathcal{D}_{SF-RF}				$\mathcal{D}_{SF-RF}^{Paired}$			
	Training		Validation		Training		Validation		Training		Validation	
Source	Target	source	target	source	target	source	target	source	target	source	target	source
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 4.3

4.3.1 Synthetic FOREST - Degraded Synthetic FOREST

The dataset \mathcal{D}_{SF-SF} is built by taking the HR synthetic FOREST crops and applying the baseline degradation model proposed in 3.1. 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 4.4: Parameters used in the degradation model employed to generate the $\mathcal{D}_{\text{SF-SF}}$ dataset.

4.3.2 Synthetic FOREST - real FOREST (Unpaired)

The dataset $\mathcal{D}_{\text{SF-RF}}$ 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.

4.3.3 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.

5 Experiment Setup

5.1 Training

Instance normalization

6 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 $\mathcal{D}_{\text{SF-SF}}$, using unpaired HR-LR pairs generated by applying the baseline degradation model described in 3.5 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 $\mathcal{D}_{\text{SF-RF}}$, 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 estimate it, generating LR versions of the synthetic FOREST images that come from the same distribution as the real FOREST images.

6.1 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. 2.5. 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. 6.1 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.

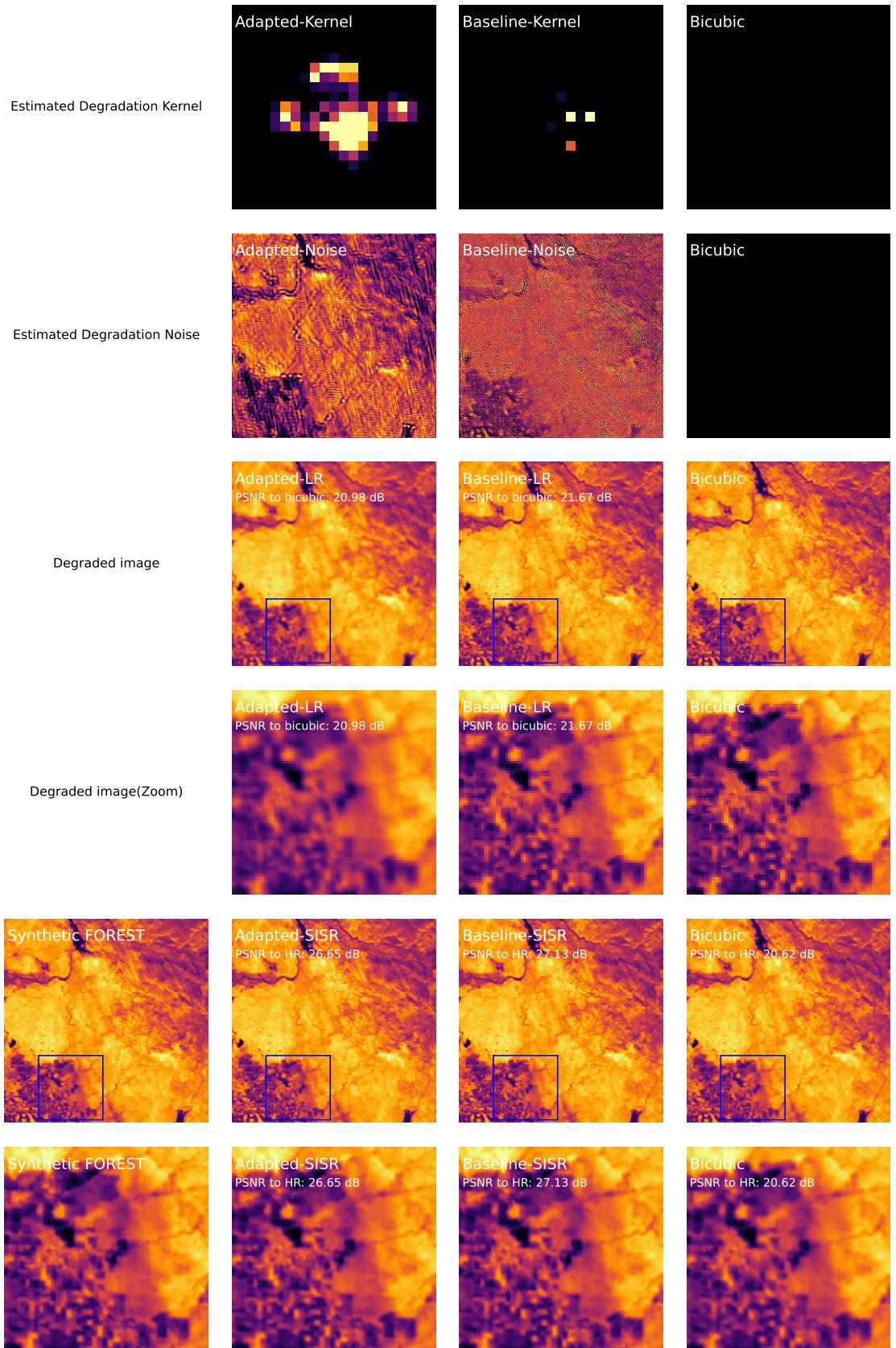


Figure 6.1: Applying different degradation models on an HR sample. The 2 most upper rows show the estimated degradation kernels and noise of each pipeline, the bicubic downsampling does not estimate a kernel or noise. The degraded LR images from each model and a zoom is displayed on the two subsequent rows. In this case, the PSNR₃₀ 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 6.2 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.

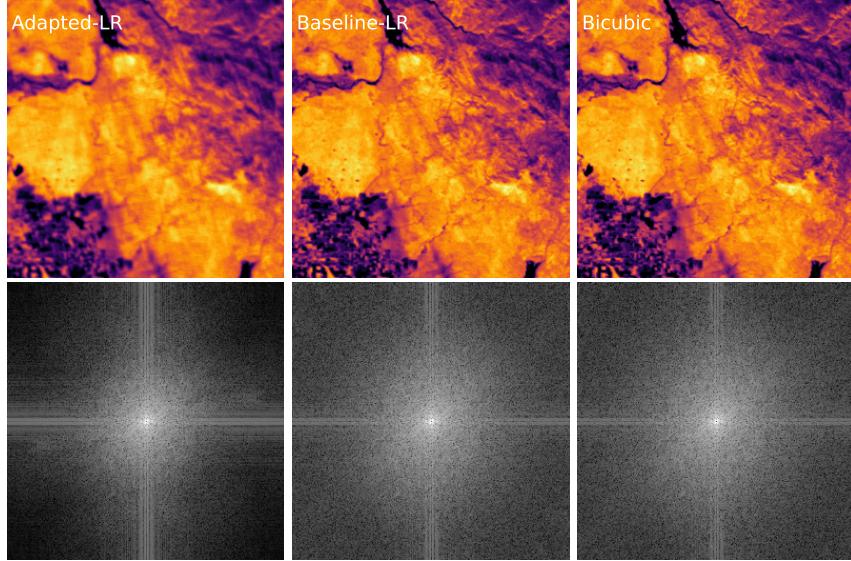


Figure 6.2: Log mangnitude of the FFT for the LR images obtained by the pipelines and the gaussian blurring + bicubic upsampling.

The radial profile of the log magnitude of the FFT for the LR images shown in Fig. 6.3 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 2100m, 0.3 cycles per pixel corresponds to 700m and 0.7 cycles per pixel to 300m. 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.

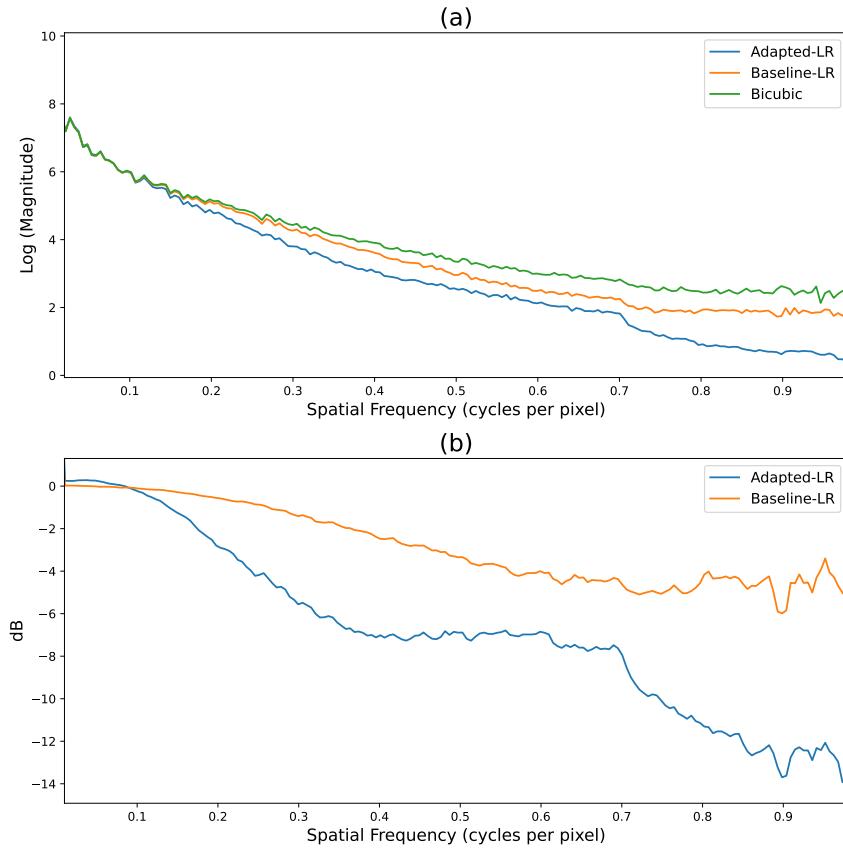


Figure 6.3: (a) Radial profile of the log magnitude across spatial frequency of the LR images obtained by the pipelines and the gaussian blurring + bicubic downsampling model. (b) Amplification in dB of each pipeline with respect to the gaussian blurring + bicubic downsampling.

When analyzing the super resolved images versus the ground truth in the frequency domain, a very similar frequency response is observed for the super resolved images of the adapted and the baseline pipelines. Moreover, the SR images are able to stay above -3dB, a common threshold 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.

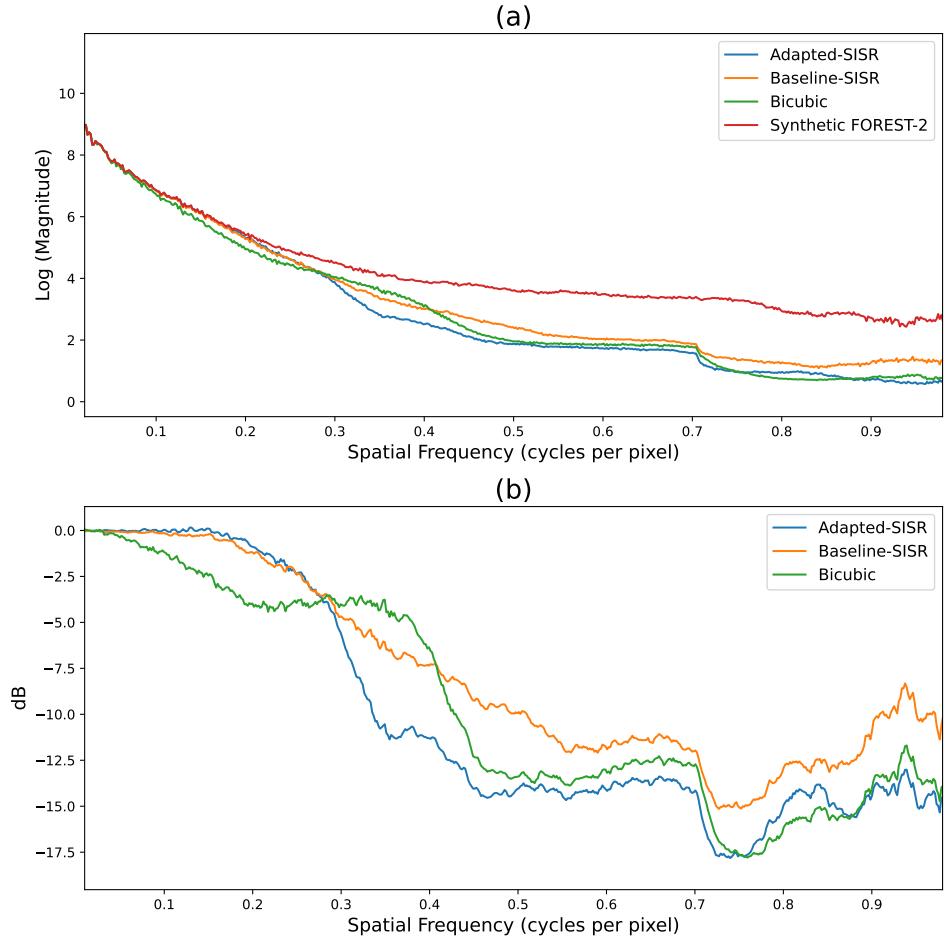


Figure 6.4: Frequency domain analysis of the SR images and the ground truth displayed in 6.1. In (a), the log of the magnitude of the FFT for the SR images and the ground truth is shown, while in (b), the amplification of each SR image with respect to the ground truth is shown.

6.1.1 LR comparison

A quantitative analysis of the LR images obtained by the generator of each pipeline is performed. Fig. 6.5 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. 6.1, where the adapted LR image is more blurry and noisy, suggesting that the unknown degradation is worse than the baseline degradation model.

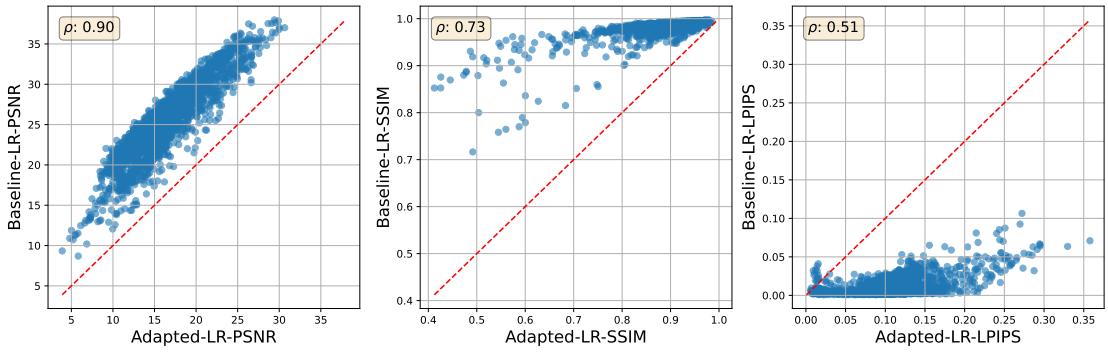


Figure 6.5: Performance metrics between the LR images obtained by the pipelines vs the gaussian blurring + bicubic downsampling degradation.

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 downscaling degradation model. The results are displayed shown in Fig. 6.6. 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 3.4.3. 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 ± 1 standard deviation interval between -4 and -6 dB from between 0.3 and 0.7 cycles per 210m pixel. This translates in blurrier images, and a more difficult starting point for a SR model to work with.

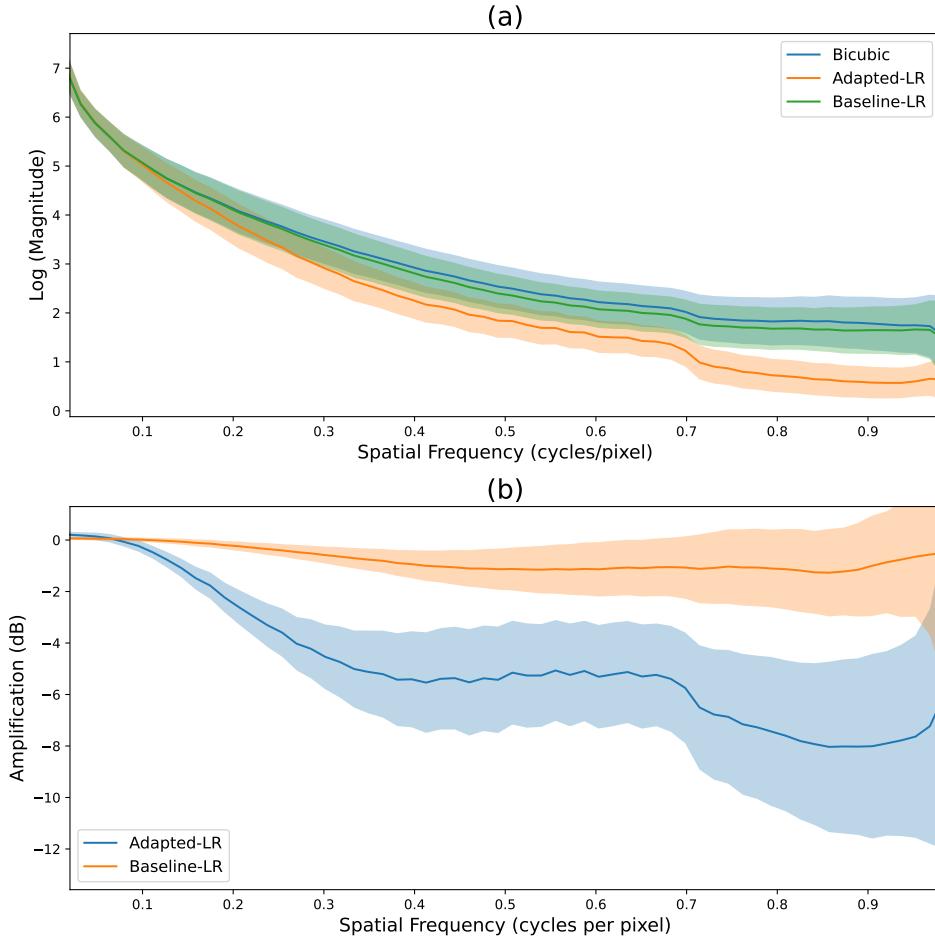


Figure 6.6: Frequency domain analysis of the LR images obtained by applying different degradation models on the HR sample displayed in Fig. 6.1. In (a), the log of the magnitude of the FFT for the LR images is shown, while in (b), the amplification with respect to a simple gaussian blurring + downscaling is shown. The painted area represents the ± 1 standard deviation of the radial profiles and the amplification.

6.1.2 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. 6.7 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 specified in FOREST-2 fact sheet. Consistent with what was found in the frequency domain analysis observed in Fig. 6.4, the SR model is able to recover most of the information, as the

performance after using the SR models is very similar.

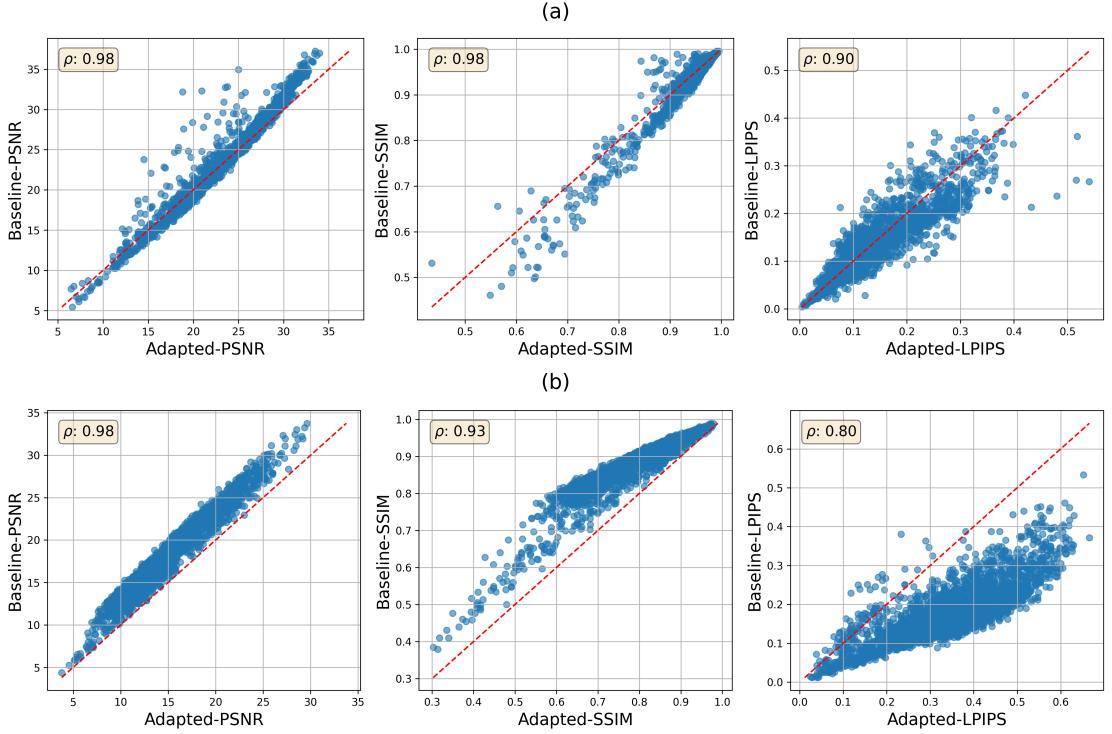


Figure 6.7: Performance obtained by super resolving the degraded images coming out of the generator. In (a), the corresponding SR model of each pipeline is used. In (b), a simple bicubic upsampling is used to super resolve the degraded images.

Fig 6.7 proves the relevance of the domain gap in super resolution, the SR model is able to estimate the inverse of the degradation function in most cases, 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.

6.2 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. 2.5. 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. 6.8, 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 $\mathcal{D}_{\text{SF-SF}}$, has very similar results to bicubic upsampling. On the other side, the adapted

model, trained using real FOREST images as the target domain ($\mathcal{D}_{\text{SF-RF}}$), 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.

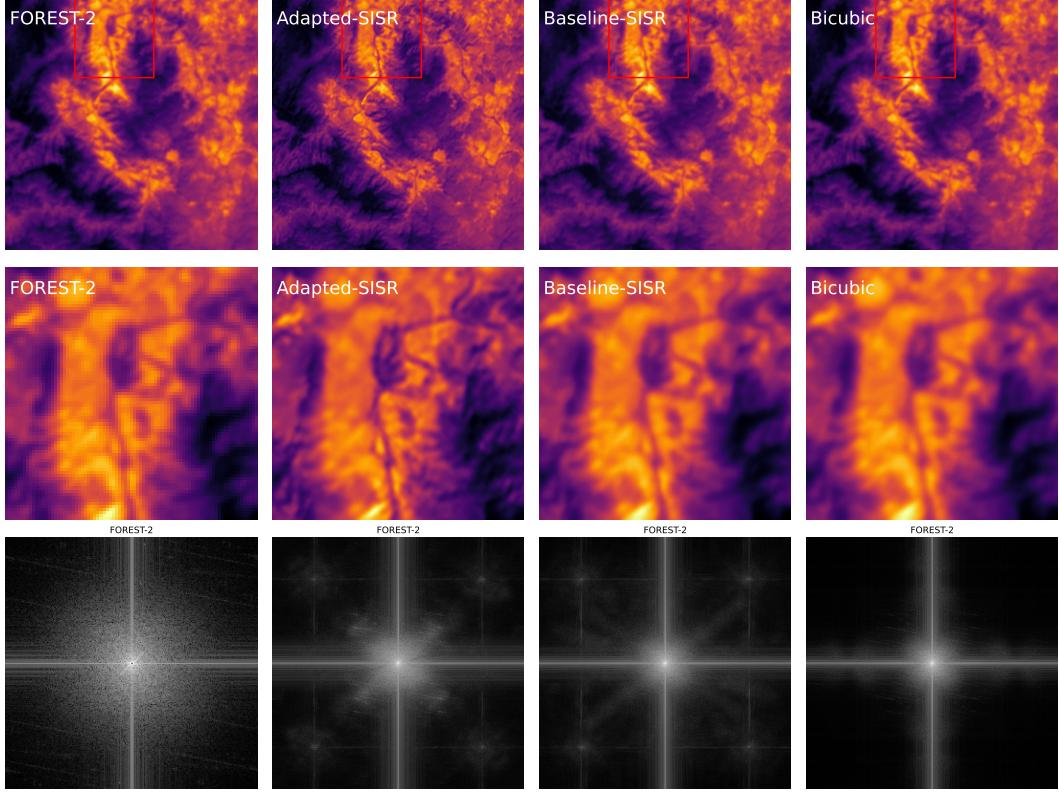


Figure 6.8: Super Resolved Forest-2 Scene using different SR models. In the upper row, the image is displayed. A detailed zoom is displayed below. The bottom row shows the log magnitude of the FFT for the images. The original image is displayed in the left, while the super resolved images are displayed afterwards.

Fig. 6.9 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 6.6. The variability of the amplification allows to conclude that this amplification is consistently higher than the baseline-SISR for all the validation 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.

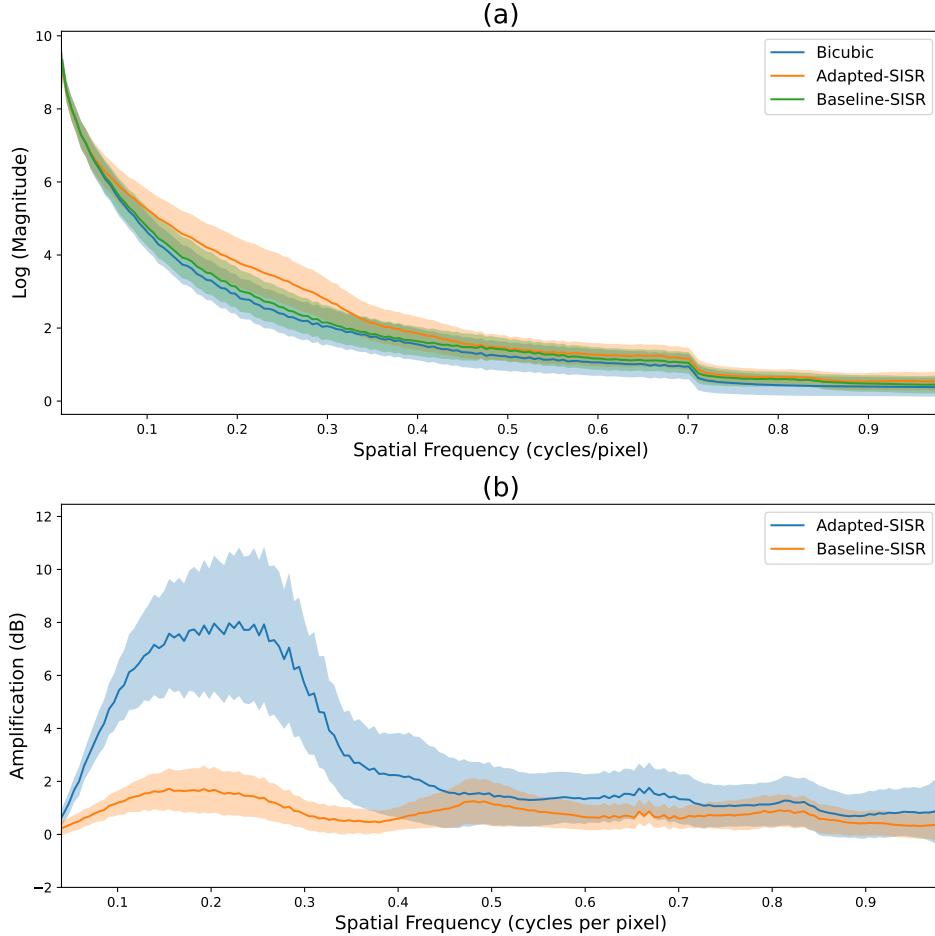


Figure 6.9: Frequency domain analysis of the SR images obtained by applying different SR models to the real FOREST-2 validation dataset. In (a), the log of the magnitude of the FFT for the SR images is shown, while in (b), the amplification with respect to a simple bicubic upsampling is displayed.

In Fig. 6.10, 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 perceived, consistent with the results of the frequency domain analysis.

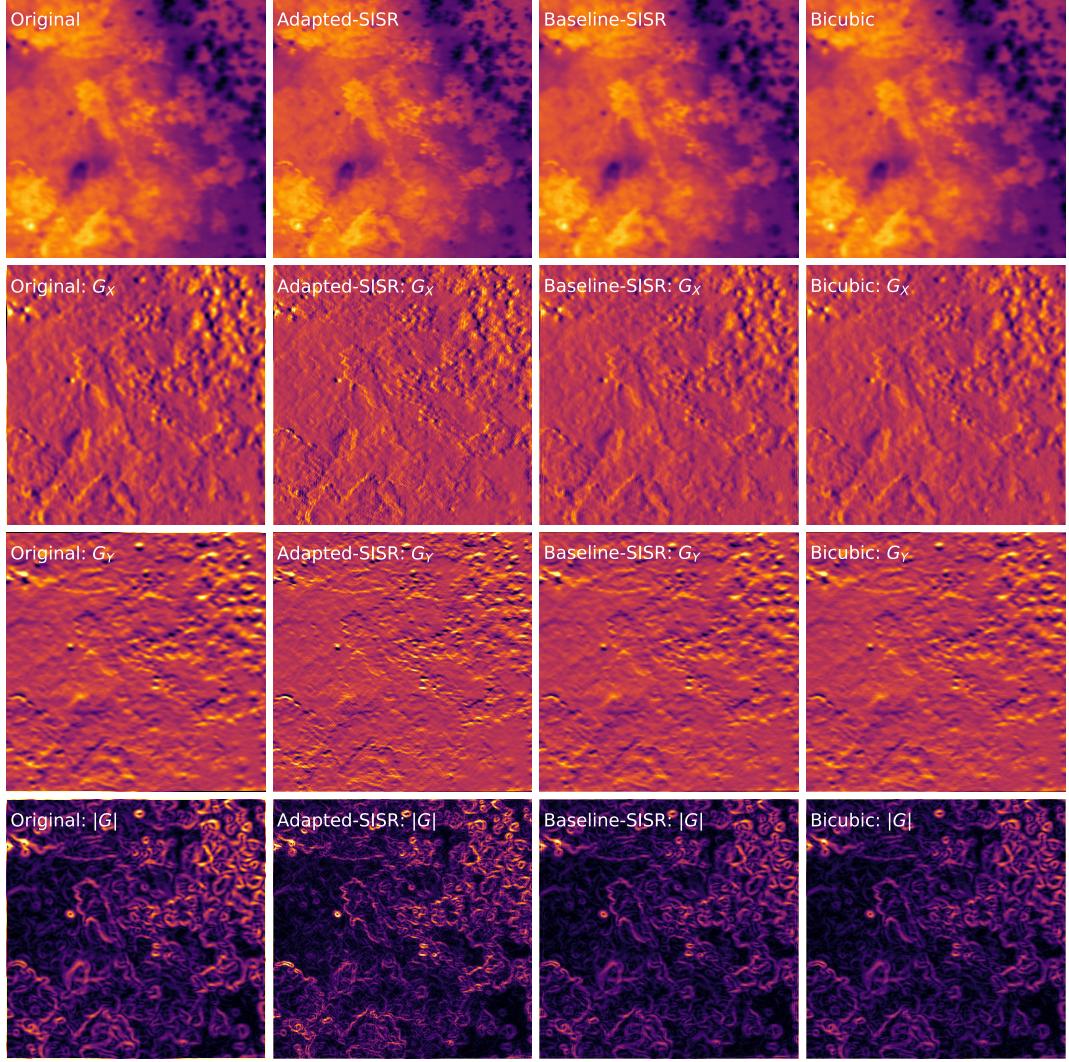


Figure 6.10: Gradient analysis of the super resolved images using different SR models for scenes coming from the real FOREST-2 validation dataset. In the upper row, the image is displayed. The gradients in the x and y direction (G_x and G_y respectively) are displayed below. the gradient magnitude $|G|$ is displayed in the bottom row.

Fig 6.11 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.

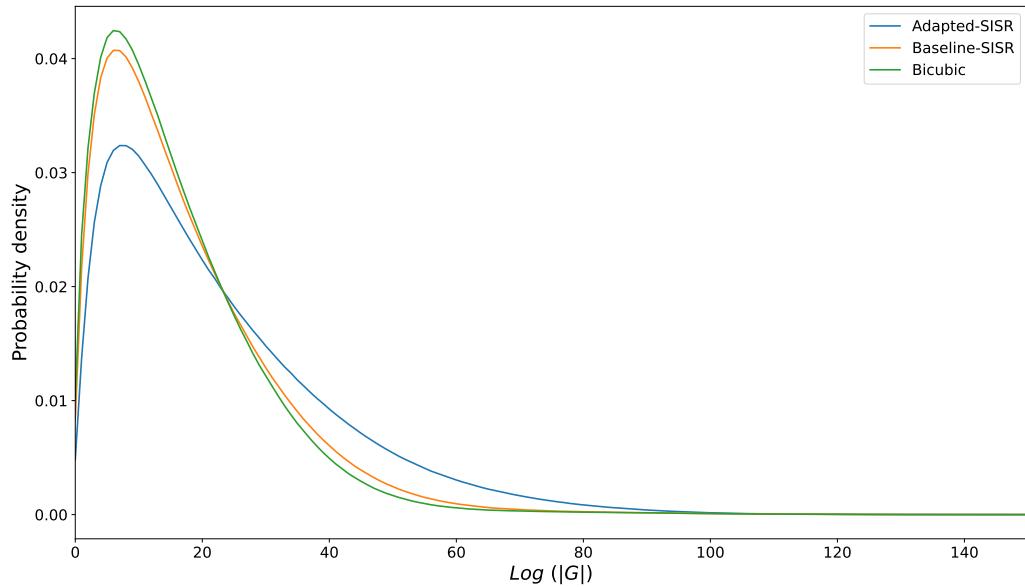


Figure 6.11: Histogram of the gradient magnitude $|G|$ for the whole validation real FOREST-2 dataset.

6.3 The domain gap goes both ways

To simulate this situation, the model trained using real FOREST-2 images was employed to super-resolve synthetic FOREST images degraded using the baseline degradation model. The results, shown in Figs. 6.12 and 6.13, indicate the performance of the adapted model is catastrophic, producing several artifacts and yielding a PSNR difference of approximately 10dB, which represents a tenfold difference in terms of Mean Squared Error (MSE). This highlights the critical need for adaptable SR models that can effectively handle diverse and evolving real-world scenarios.

The combination of the probabilistic degradation model and the SR model were proven helpful to bridge the domain gap and improve 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 experiments, the opposite can also occur. Assuming a more complex degradation model in the dataset could lead to generated LR images with more attenuation in the frequency domain, 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 second 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. 6.12 and 6.13. 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.

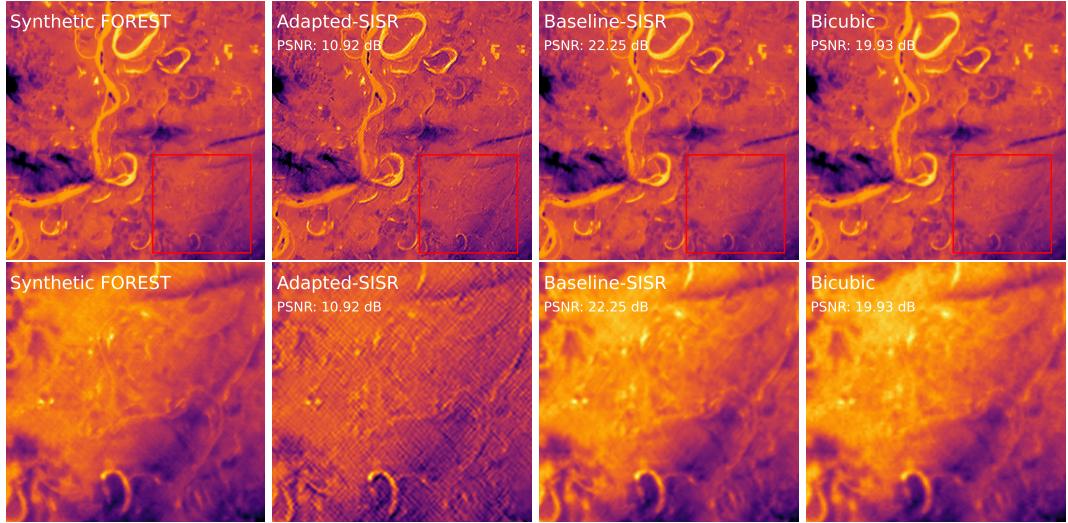


Figure 6.12: Effects of using a model trained with on different domain than at inference time. When using an Synthetic FOREST image degraded with the baseline degradation model as an input, the model trained using real FOREST-2 data as the target domain generates several artifacts and underperforms severely in terms of PSNR.

In the frequency domain, the results are shown in Fig. 6.13. 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.

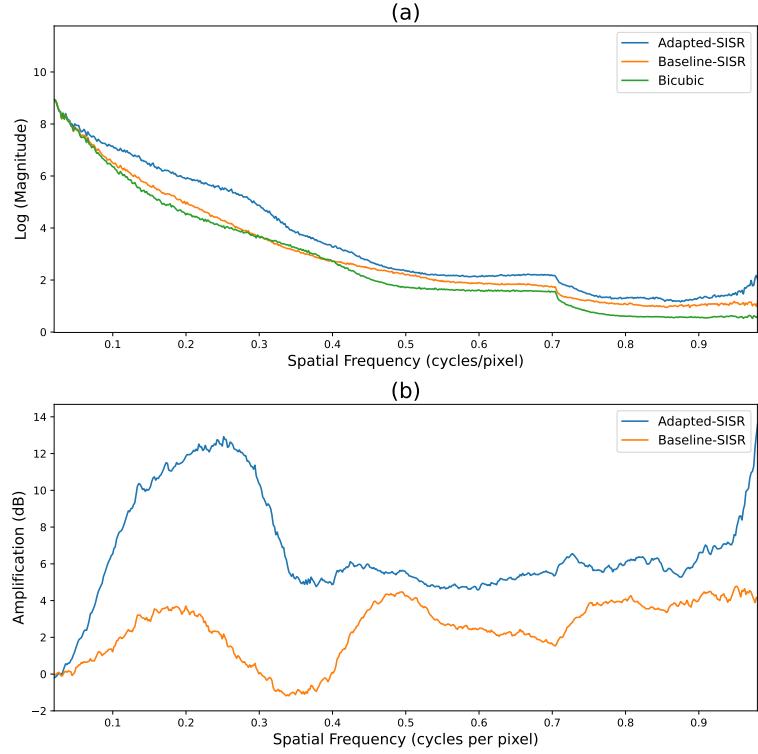


Figure 6.13: Effects of using a model trained with one domain than at inference time. (a) shows the log magnitude of the radial average of the FFT for the SR images using different algorithms. (b) shows the amplification with respect to bicubic interpolation.

The performance results in terms of different metrics are shown in Fig. 6.14. In the conditions described above, the adapted super resolution model underperforms severely in every considered metric.

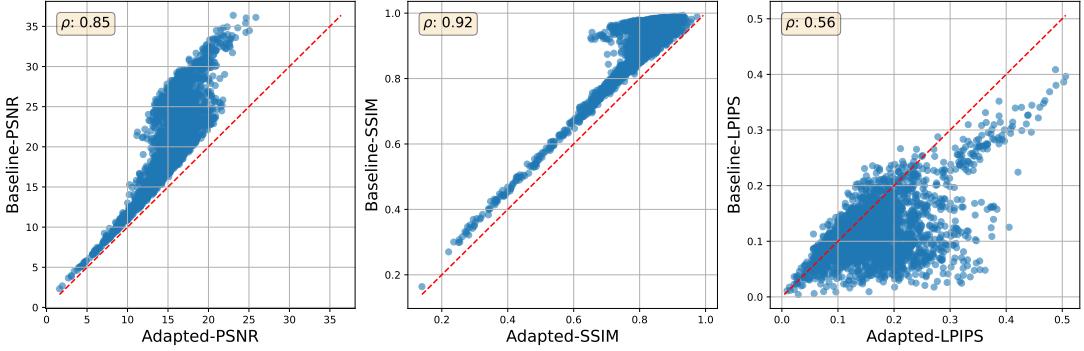


Figure 6.14: Performance obtained by super resolving the degraded synthetic FOREST images using different super resolution models.

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, where GAN approaches are not able to generalize to arbitrary domains DASDASD.

6.4 Domain gap using non-references 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. The image quality assessment metrics was done for \mathcal{D}_{SF-RF} and \mathcal{D}_{SF-SF} , calculating the NIQE and BRISQUE scores of all the super resolved images from each validation dataset.

The results are shown in Fig. 6.15. For both metrics, a large gap is observed between the adapted model and the rest for dataset \mathcal{D}_{SF-RF} , 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}_{SF-SF} .

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 any 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. The images used for training the NIQE and BRISQUE models are not remote sensing images, and therefore, the results may not be representative. This could be circumvented by training a NIQE/BRISQUE model using remote sensing images only.
2. NIQE and BRISQUE are objective measures of image quality, not of physical consistency or image fidelity.

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

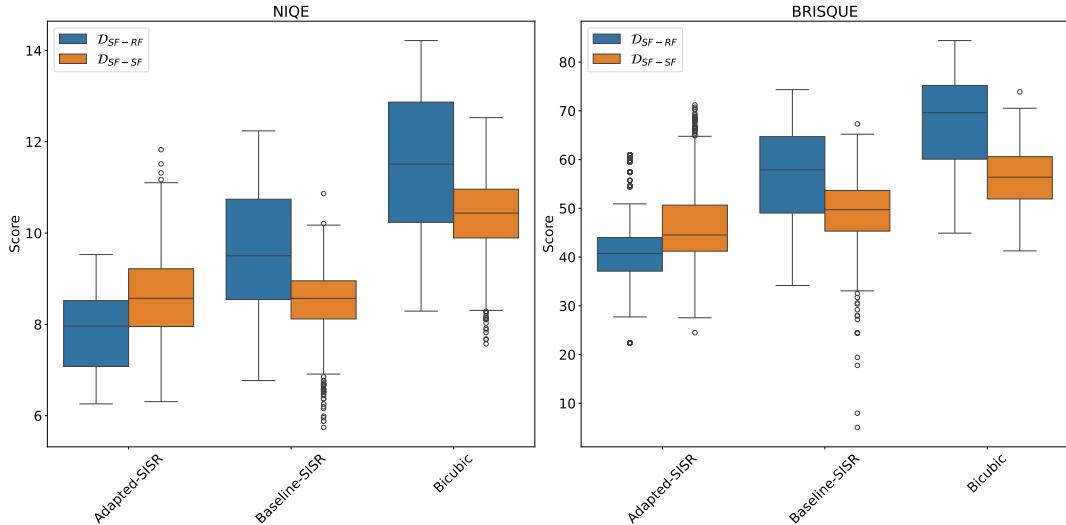


Figure 6.15: Image quality assessment metrics for the different SR models using different datasets as input. In both metrics, the lower the score, the better the image quality.

7 conclusions and future work

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:

- 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.

Training of non-referenced image quality assesment for remote sensing

References

- [1] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations, 2010.
- [2] 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.
- [3] Diego Valsesia and Enrico Magli. Permutation invariance and uncertainty in multitemporal image super-resolution, 2021.
- [4] Syed Muhammad Anwar Bashir, Yanning Wang, Murtaza Khan, and Yulei Niu. A comprehensive review of deep learning-based single image super-resolution, 2021.
- [5] Marcus martens, Dario Izzo, Andrej Krzic, and Daniël Cox. Super-resolution of proba-v images using convolutional neural networks, 2019.
- [6] 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.
- [7] Christian Mollière, Julia Gottfriesen, Martin Langer, Patricio Massaro, Christian Soraruf, and Matthias Schubert. Multi-spectral super-resolution of thermal infrared data products for urban heat applications, 2023.
- [8] 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 Micheiloni, 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.
- [9] Anran Liu, Yihao Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Blind image super-resolution: A survey and beyond, 2021.
- [10] Zhengxiong Luo, Yan Huang, Shang Li, Liang Wang, and Tieniu Tan. Learning the degradation distribution for blind image super-resolution, 2022.
- [11] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution, 2018.
- [12] 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.

- [13] 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.
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
- [15] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020.
- [16] Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs, 2017.
- [17] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first, 2018.
- [18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
- [19] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a “completely blind” image quality analyzer, 2013.
- [20] Dario Fuoli, Luc Van Gool, and Radu Timofte. Fourier space losses for efficient perceptual image super-resolution, 2021.
- [21] Irwin Sobel and G. M. Feldman. An isotropic 3×3 image gradient operator, 1990.
- [22] Jet Propulsion Laboratory. ECOSTRESS Fact Sheet, 2023. [Online accessed 28-November-2023].
- [23] PhyTIR: Plant High Temperature InfraRed Viewer. <https://phytir.jpl.nasa.gov/>, 2023. [Online accessed 28-November-2023].
- [24] Application for extracting and exploring analysis ready samples (AppEEARS). <https://appeears.earthdatacloud.nasa.gov/>, 2023. [Online; accessed 28-November-2023].
- [25] Appearss api. <https://appeears.earthdatacloud.nasa.gov/api/>, 2023. [Online; accessed 28-November-2023].
- [26] 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].
- [27] Ecostress faq. <https://ecostress.jpl.nasa.gov/faq>. Accessed: 2023-11-28.