Universidad Internacional de la Rioja (UNIR)

Escuela Superior de Ingeniería y Tecnología

Master in Artificial Intelligence

Enhancing Spatial Resolution of Sentinel-2 Satellite Images Using Multispectral Fusion and Super-Resolution Models

Final Master's Project

Presented by: Julio Cesar Contreras Huerta

Supervised by: Miguel Angel Navarro Burgos

City: Valencia

Date: February 2024

TABLE OF CONTENTS

1. INTRODUCTION

As the world becomes increasingly interconnected, the ability to observe and analyze our environment from space has emerged as one of the most important tools in addressing global challenges. From this perspective, satellite missions, through their advanced sensor systems, allow the collection of information about the Earth's surface, providing the most detailed and extensive view of the processes of change on our planet. Among the key disciplines that enable these functions, remote sensing stands out, transforming the data obtained by these sensors into critical information for solving environmental and social issues. This discipline provides specific analyses of the transformations and characteristics of certain geographic areas. With advancements in AI (Artificial Intelligence), analytical capabilities have reached a new level, allowing almost any type of satellite image to be analyzed with unprecedented precision and efficiency.

Spatial resolution is a fundamental parameter for the analysis of remote sensing images. It is defined as the minimum ground distance that separates two independent objects that can be distinguished. The factors that determine it include altitude, distance, and the quality of the instruments used (Alparone et al., 2015). Another important factor related to spatial resolution is Ground Sampling Distance (GSD), which is the portion of the Earth's surface represented by each pixel (Lillesand et al., 2015).

One of the most widely used sensors due to its high spatial resolution is the Sentinel-2 MSI, operated by the European Space Agency (ESA, 2019). Since its launch in June 2015, it has provided open-access multispectral images, generating significant interest in the scientific community and various industries, becoming a key tool for providing data in applications such as land use monitoring, change detection, and vegetation analysis. The sensor is equipped with 13 spectral bands distributed in three spatial resolutions: 4 spectral bands with a resolution of 10 m, 6 spectral bands with 20 m, and 3 spectral bands with 60 m, designed for the collection of various topographic parameters (Lanaras et al., 2018).

This spectral diversity enables detailed studies on a wide range of terrestrial phenomena by combining specific bands, from water quality assessment to snow surface monitoring.

Another sensor that stands out for its higher resolution is the National Agriculture Imagery Program (NAIP), which provides aerial images with a resolution of up to 1 meter. Although its coverage is limited to the continental United States, it offers aerial images with a spatial resolution of up to 1 meter, making it an important source for studies requiring a higher level of detail. NAIP captures images during the peak agricultural activity months, from June to August, and provides updates every three years (previously every five years before 2009). Since 2011, the images have consistently included RGB and NIR bands, improving their quality and applicability in areas such as urban planning and agricultural monitoring. Although NAIP and Sentinel-2 have different approaches and scopes, the combination of images from both sensors allows for a greater richness of data, which is crucial for studies requiring both extensive coverage and exceptional detail.

There are various techniques to address the limitation of spatial and spectral resolution in satellite images within the field of remote sensing, from traditional image fusion methods to more recent approaches based on artificial intelligence. Among the latter, deep learning has proven to be especially promising (Gargiulo et al., 2019). One of the most effective applications of this technique is super-resolution, which seeks to increase the spatial resolution of images by generating additional details from existing data. These machine learning techniques allow for the reconstruction of higher-resolution images from lower-quality versions, optimizing both spatial precision and spectral coherence.

1.1 Motivation

The development of techniques to improve the resolution of images obtained by the Sentinel-2 sensor has been a significant challenge in the scientific community. In this process, advanced deep learning architectures such as Generative Adversarial Networks (GANs) and Convolutional Neural Networks (CNNs) play a crucial role, efficiently managing the data (Salgueiro Romero et al., 2020). Achieving an increase in spatial resolution

while maintaining spectral coherence and optimizing data quality opens new frontiers in satellite image analysis.

The availability of free and high-quality images, such as those provided by Sentinel-2, is a valuable resource for scientific research. Although there are other ways to obtain high-resolution images, they often do not cover as large areas or have the necessary temporal resolution for certain studies. Thanks to its global coverage and high revisit frequency, this sensor offers a significant advantage in this regard. Improving its resolution would allow for more detailed studies without the need to resort to costly sensors, making the available data even more impactful.

Additionally, this work seeks to contribute to the advancement of AI in remote sensing applications. By using innovative architectures, this research has the potential to open new possibilities for improving satellite image resolution. This approach not only benefits current science but also has the potential to inspire future research, allowing the continued development of these techniques in new contexts.

1.2 Problem Statement

Despite the availability of advanced super-resolution models (Navarro & Sánchez, 2020; Salgueiro Romero et al., 2020), the main challenge lies in ensuring that the resulting images not only improve spatial resolution but also maintain spectral coherence, a crucial factor in scientific applications. To overcome the spatial resolution limitations of sensors like Sentinel-2, this work proposes the use of advanced super-resolution techniques based on deep learning models. Super-resolution has become a powerful tool for enhancing image quality, using architectures such as Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) to generate high-resolution images while preserving spectral and spatial coherence. These techniques, applied to Sentinel-2 multispectral images, have the potential to produce results comparable to those obtained with higher-resolution sensors like NAIP, but with the advantage of working with more accessible data.

In various studies, the Wald Protocol has been followed, a standard in the field of remote sensing, to validate the quality of super-resolved images. This protocol establishes rigorous criteria that ensure the enhanced images maintain spectral and spatial consistency with the original images. In this work, a super-resolution (SR) model is proposed that transforms the 10m and 20m bands of Sentinel-2 into higher-resolution images (2.5m), preserving critical spectral details for precise studies. The enhancement process includes an intermediate step of super-resolution from 10m to 40m, which strengthens the model's robustness.

Furthermore, the use of the Wald Protocol ensures that the fused images retain both spatial and spectral integrity, validating their quality. This approach, which separates the image fusion and resolution enhancement processes through super-resolution, allows each stage to work independently, increasing the versatility and applicability of the proposed method for different types of sensors and applications.

This research not only aims to advance the field of image super-resolution but also offers a viable and accessible solution for projects that require high-resolution images but must operate within the economic and technical limitations of current sensors.

1.3 Structure of the Work

In **Chapter One**, the purpose and motivation of the work are established. The problem addressed focuses on the need to improve the spatial resolution of satellite images, specifically those from Sentinel-2, through super-resolution techniques. The chapter also describes the organization of the thesis, outlining the main sections and objectives.

Chapter Two delves into a detailed analysis of the context, examining the current state of satellite imaging technologies and the demand for high-resolution data in various fields, such as environmental monitoring and urban planning. The benefits and challenges of improving spatial resolution through multispectral data fusion and super-resolution models are discussed, with a focus on the limitations of existing techniques.

Chapter Three reviews the state of the art in super-resolution technologies applied

to satellite images. The chapter provides an overview of key methods such as pansharpening, image model inversion, and deep learning approaches. It also highlights the importance of the Wald Protocol as a validation framework to ensure spectral and spatial consistency in the generated high-resolution images.

In **Chapter Four**, the general and specific objectives of the project are defined. The main objective is to develop a super-resolution model capable of improving Sentinel-2 images from 10m and 20m resolution to 2.5m resolution using a combination of Convolutional Neural Network (CNN) models and multispectral data fusion techniques. The methodology section details the process followed for model training, dataset preparation, and evaluation using the Wald Protocol.

Chapter Five presents the functional and non-functional requirements of the super-resolution system. Functional requirements include the ability to handle multispectral images of various resolutions and generate consistent, high-quality results with a resolution of 2.5m. Non-functional requirements focus on the system's efficiency, scalability, and robustness in handling large datasets.

In **Chapter Six**, the development of the super-resolution model is explained in detail. This includes the training process, the design of the neural network architectures (Fusion X2 and X4 models), and the integration of deep learning techniques. The limitations encountered during the training phase, such as memory constraints and model performance, are discussed along with potential optimizations.

Chapter Seven evaluates the performance of the developed model, comparing it with existing techniques. The evaluation is based on both qualitative and quantitative metrics, focusing on the spatial and spectral fidelity of the generated high-resolution images. The chapter also includes an analysis of the system's usability and its potential impact on real-world applications, such as precision agriculture and land use monitoring.

Chapter Eight concludes the work by summarizing the main contributions of the thesis and proposing future research directions. Possible improvements include refining the deep learning architecture, exploring other fusion techniques, and applying the model

to different types of satellite images.

Appendix I includes output images generated by the super-resolution model, demonstrating the improved resolution for various Sentinel-2 bands.

Appendix II presents the results of the Wald Protocol validation tests, showing the system's performance in maintaining spectral and spatial consistency.

Appendix III provides detailed documentation of the software tools and libraries used for the model's implementation.

Appendix IV contains the research paper associated with this thesis, submitted for publication.

2. CONTEXT ANALYSIS

2.1 Datasets

In the realm of super-resolution algorithms, the choice of the dataset can significantly influence the algorithm's training and ultimately its performance. Three key datasets in this domain are the OpenImage Dataset, the WorldStrat Dataset, and the Sen2Venus Dataset. Upon contrasting these datasets, it becomes evident that there are several trade-offs to consider. These trade-offs, depending on the nature of the super-resolution tasks to be performed, could significantly influence the choice of the dataset. The OpenImage Dataset, for instance, excels with its HR image pairs that reach up to 0.5 meters. However, its LR pairs are synthetically generated and are geographically limited to the United States. Consequently, its global applicability may be limited, which could potentially constrain its utility in worldwide scenarios. The WorldStrat dataset, on the other hand, stands out with its diversity, offering a wide range of scenes and landscapes. This diverse sampling can enhance the generalization capabilities of SR algorithms. However, its approach of not filtering low-resolution revisits by their cloud coverage introduces an additional layer of complexity. Additionally, it is quite small in comparison to the size if the models. Lastly, the Sen2Venus Dataset is noted for its well-managed pre-processing that assures good correspondence between LR and HR images in the spectral domain. However, the HR image has only double the spatial resolution of the LR image and may lack the diversity found in the WorldStrat dataset. Below, we provide a more detailed explanation of each dataset.

2.1.1 OpenImages

OpenImages is a popular computer vision dataset that provides a large collection of labeled images for training and evaluating machine learning models. The dataset covers a wide range of visual concepts and objects, including people, cars, houses, trains, animals,

etc. It is widely used in tasks such as object detection, image classification, and visual relationship detection. For SR purposes, the amount and quality of the annotations are not of importance, but rather that it is an easily accessible and freely available dataset (CC-BY) containing over millions images. The original authors of the latent-diffusion model used this dataset to train their checkpoints.

Since it is a computer vision dataset, it only contains RGB images of every-day scenes taken with standard digital cameras. Not only are therefore the objects depicted in the images completely unrelated to the remote sensing domain, their spectral characteristics are vastly different as well. In order to train 4 band models, a 4th band is artificially created from the image intensity and appended in the band dimension. The LR image is created by interpolating the HR version to the desired size of $128x128$ pixels.

The data in question is not directly connected to the research problem in terms of its spectral, spatial, or domain characteristics. However, the large volume of images proves valuable for training a freshly initialized model on fundamental image attributes and connections. Surprisingly, when applying super-resolution techniques to remote sensing imagery using autoencoders and denoising U-Nets trained on the OpenImages dataset, remarkably good results have been achieved. This demonstrates that the knowledge acquired through training on the OpenImages dataset can be transferred to the remote sensing domain and create a solid base for finetuning on more sophisticated datasets.

Figura 1.

Example of an OpenImages LR and HR image pair.

Note. Example of an OpenImages LR and HR image pair.

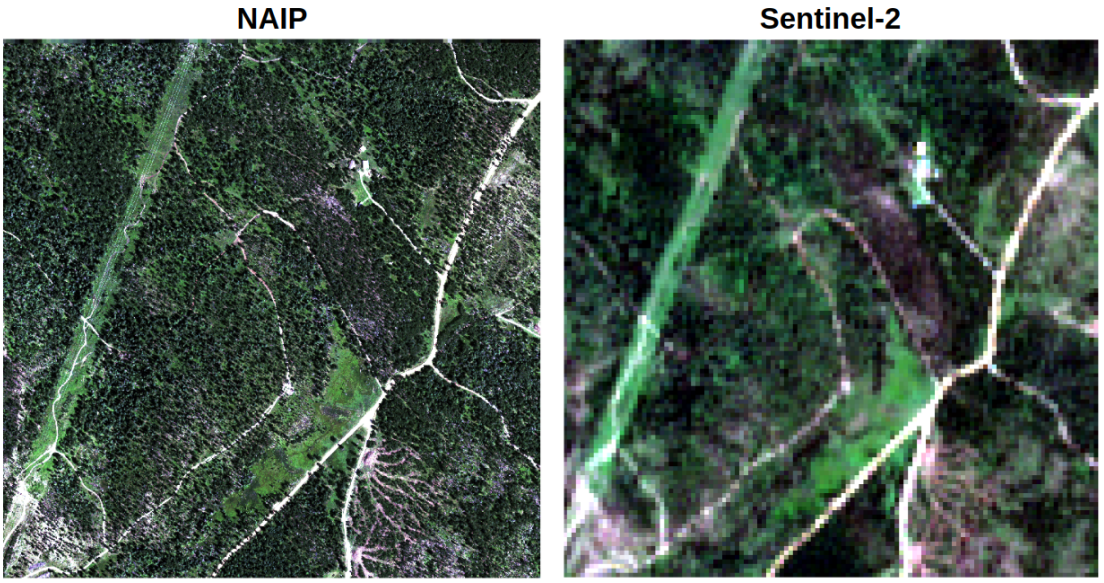
2.1.2 Sentinel-2 Dataset

2.1.3 NAIP Dataset

The National Agriculture Imagery Program (NAIP) acquires high-resolution aerial imagery of agricultural areas in the United States. It provides current and accurate imagery to support agricultural applications and decision-making. NAIP captures imagery at a spatial resolution of 0.6 meters, allowing for detailed analysis of agricultural activities. Using specialized aerial platforms with multispectral sensors, NAIP collects data in various spectral bands, including red, green, blue, and near-infrared. The NAIP dataset covers diverse agricultural regions and is regularly updated. It is freely available to the public, facilitating its use in agricultural research, precision farming, land management, and related applications. A comparison between Sentinel-2 and NAIP is shown in Figure 2.

Figura 2.

Comparison between Sentinel-2 and NAIP near Rapid City, South Dakota.



Note. Comparison between Sentinel-2 and NAIP near Rapid City, South Dakota. The Sentinel-2 image (rigth) has a spatial resolution of 10 meters, while the NAIP image (left) has a spatial resolution of 0.6 meters.

2.1.4 OpenSR dataset

The OpenSR dataset is an extensive dataset designed to facilitate the development of standard and reference single-image super-resolution algorithms for Sentinel-2 imagery. The high-resolution (HR) images in this dataset are sourced from the National Agriculture Imagery Program (NAIP), which provides aerial imagery covering the entire contiguous United States.

NAIP collects data across various spectral bands, including red, green, blue, and near-infrared, at a resolution of 0.6 m/pixel.

2.1.5 WorldStrat Dataset

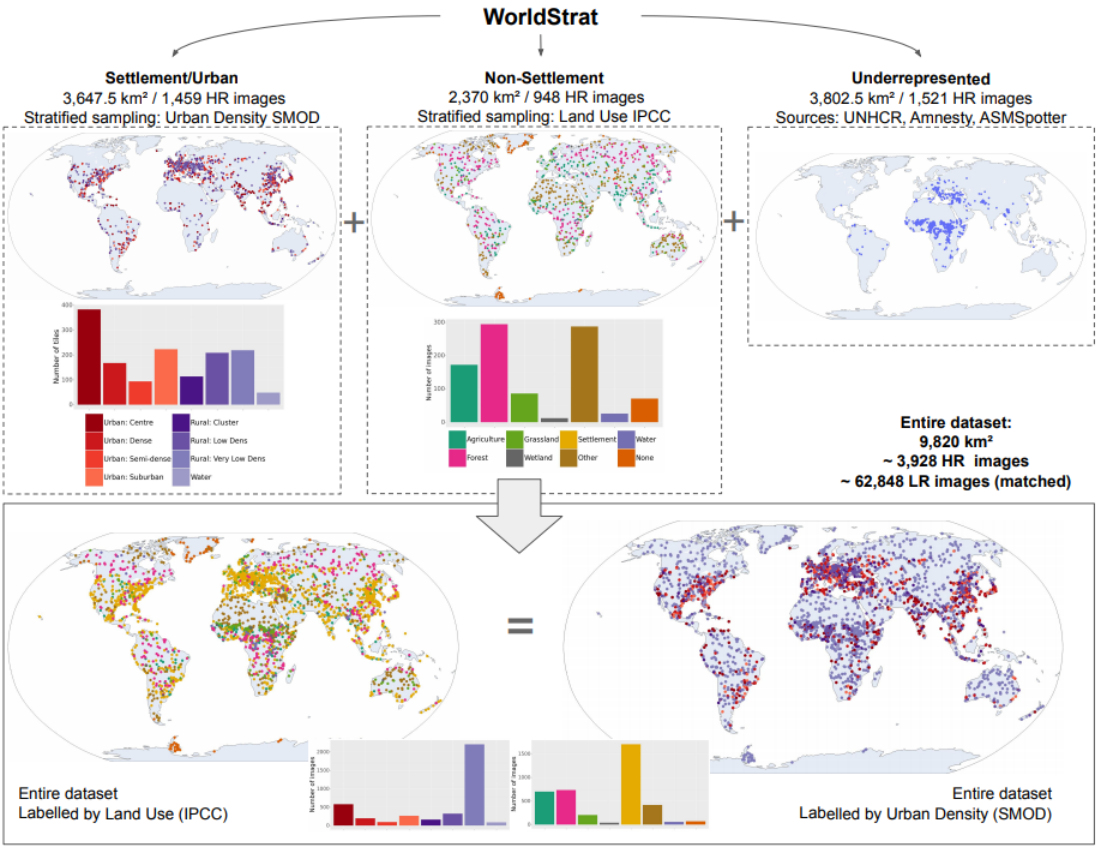
The World Stratified Dataset, also known as WorldStrat, is a curated and diverse dataset. Its primary purpose is to aid in the development of multi-frame super-resolution algorithms specifically designed for Sentinel-2 imagery. A key feature of this dataset is the inclusion of high-resolution (HR) imagery from the SPOT 6/7 satellites. The SPOT imagery encompasses five distinct spectral bands. The panchromatic (PAN) band is at 1.5 m/pixel and the remaining bands, include Red, Green, Blue, and Near Infrared (RGBNIR) at 6 m/pixel.

The dataset covers approximately 10000 square kilometers and includes 3504 distinct areas of interest 3, curated for the highest diversity of possible uses. The image acquisition budget for the dataset is divided into three parts:

- **Settlement/Urban:** This part covers 3,647.5 square kilometers with 1,459 high-resolution images. The images are stratified sampling based on urban density (SMOD).
- **Non-Settlement**: This part covers 2,370 square kilometers with 948 high-resolution images. The images are stratified sampling based on land use (IPCC).
- **Underrepresented**: This part covers 3802.5 square kilometers with 1521 high-resolution images. The sources for these images include UNHCR, Amnesty, and ASMSpotter.

Figura 3.

Summarizing the construction and classes of the WorldStrat dataset.



Note. Summarizing the construction and classes of the WorldStrat dataset. Obtained from Cornebise et al. (2022)

2.1.6 Sen2Venus Dataset

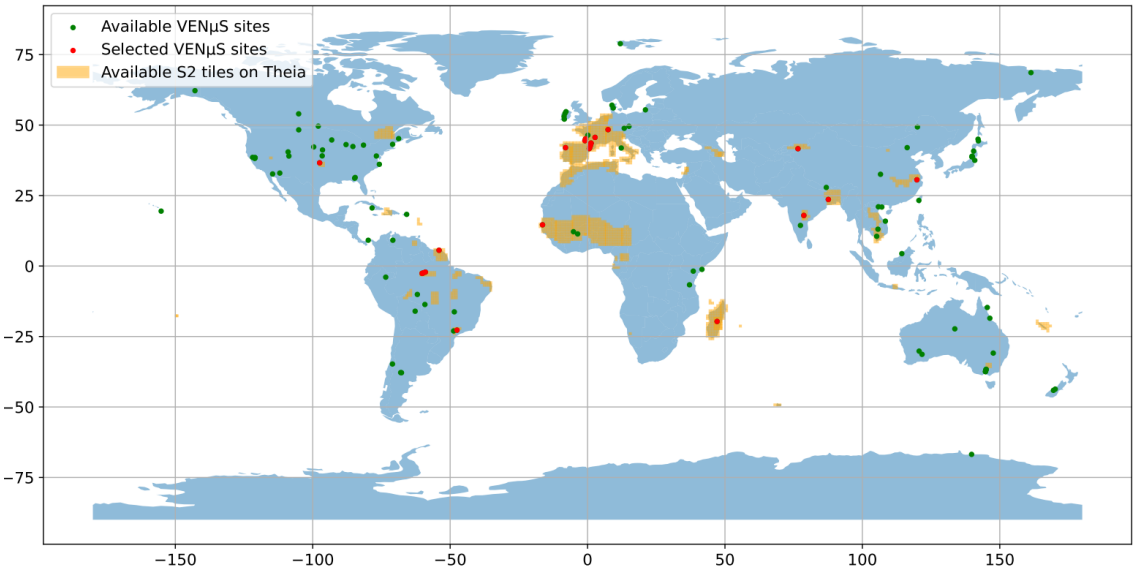
SEN2VENµS is a large dataset specifically designed to improve the spatial resolution of eight Sentinel-2 bands to 5 meters. It consists of cloud-free surface reflectance patches captured by Sentinel-2 L2A at 10 meters and 20 meters, along with corresponding reference patches acquired by the VENµS satellite at 5-meter resolution on the same day. The dataset covers 29 different locations worldwide, encompassing a total of 132,955 patches, each with a size of 256 × 256 pixels

SEN2VENµS is a large dataset specifically designed to improve the spatial reso-

lution of eight Sentinel-2 bands to 5 meters. It consists of cloud-free surface reflectance patches captured by Sentinel-2 L2A at 10 meters and 20 meters, along with corresponding reference patches acquired by the VENμS satellite at 5-meter resolution on the same day. The dataset covers 29 different locations worldwide, encompassing a total of 132,955 patches, each with a size of 256 × 256 pixels (Figure 4).

Figura 4.

Map of Sentinel-2 coverage on Theia (orange), available VENμS sites (green).



Note. Map of Sentinel-2 coverage on Theia (orange), available VENμS sites (green) and 29 selected sites (red) for the dataset. Obtained from Michel et al. (2022)

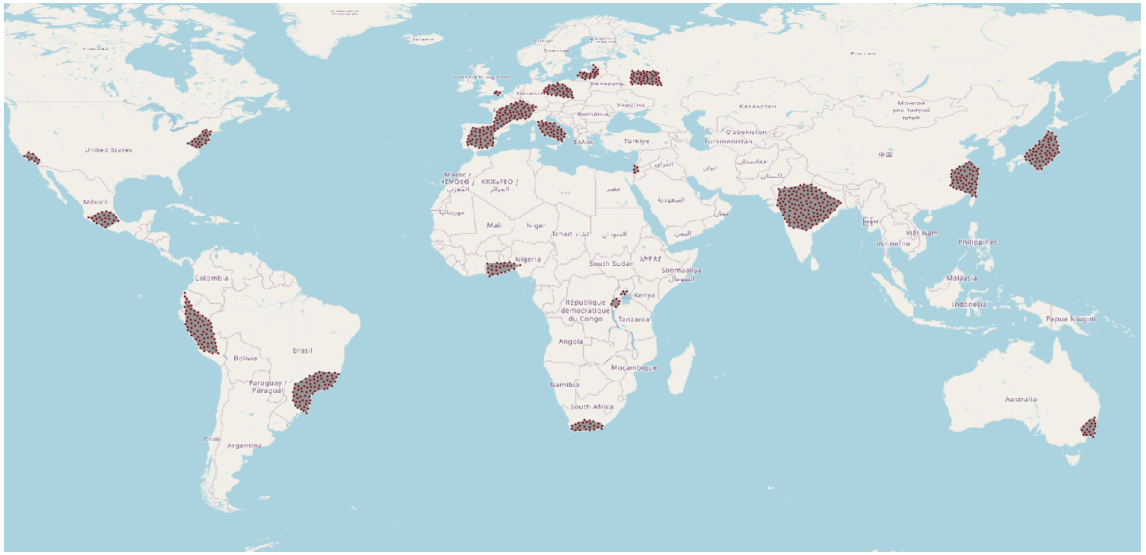
2.1.7 Degraded Sentinel-2 Dataset

The datasets introduced above all have certain qualities and drawbacks. One of the factors to balance is dataset size compared to the quality and suitability of the data. On one end of the spectrum, WorldStrat is of very high quality and closely connected to the type of SR we want to perform, while on the other end the OpenImages dataset is huge while having no connection to remote sensing imagery. In order to have images in the spectral domain of Sentinel-2 and from the remote sensing domain, a large amount of Sentinel-2 RGB-NIR imagery has been sampled all over the world (Figure 5).

The images of course have the standard Sentinel-2 resolution of 10 meters, which is used as the HR version. The LR version of the image pair is created by applying a degradation kernel to the HR image, producing a pair of LR-HR RGB-NIR images that have 40 and 10m spatial resolution respectively, for a SR factor of 4. Over 250,000 image pairs are created this way, providing a dataset that has no problems with a spectral, temporal or geometrical mismatch and is already in the domain of Sentinel-2. The only drawback is that the frequency of information is changed since the scale of objects on the ground is vastly different. Still, this dataset is very useful in training the networks either from scratch or for a rougher finetuning after being trained on natural image datasets. After that, the NAIP and worldstrat datasets can provide the final touch.

Figura 5.

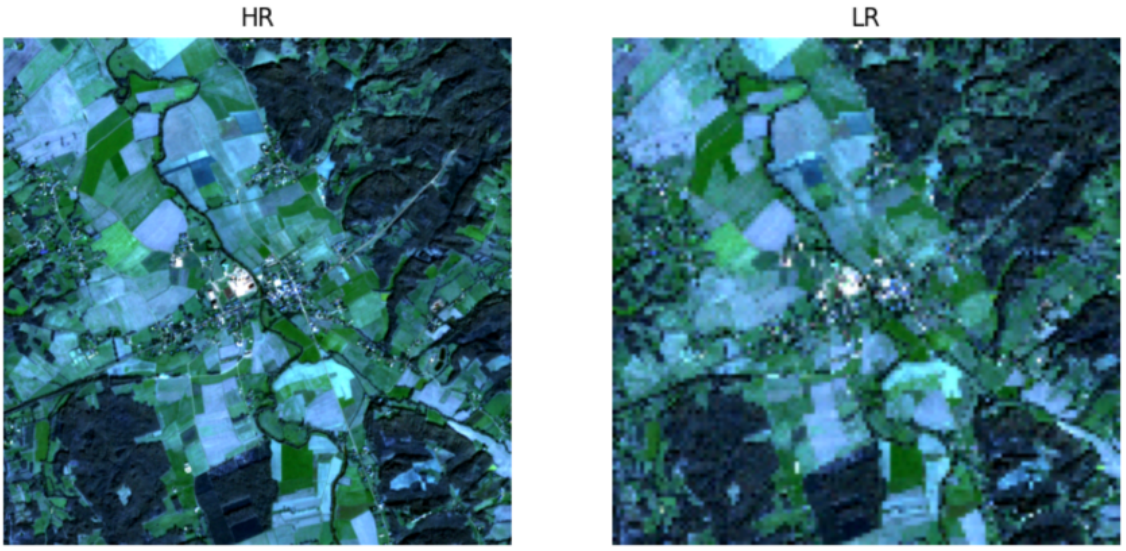
Sampling areas (gray) and locations (red) of cloud-free Sentinel-2 images.



Note. Sampling areas (gray) and locations (red) of cloud-free Sentinel-2 images.

Figura 6.

Example of the degraded Sentinel-2 dataset.



Note. Example of the degraded Sentinel-2 dataset (HR: 10m, LR: 40m). While the spectral properties as unmistakable Sentinel-2-like, the spatial information frequency is clearly different from 2.5m-10m datasets.

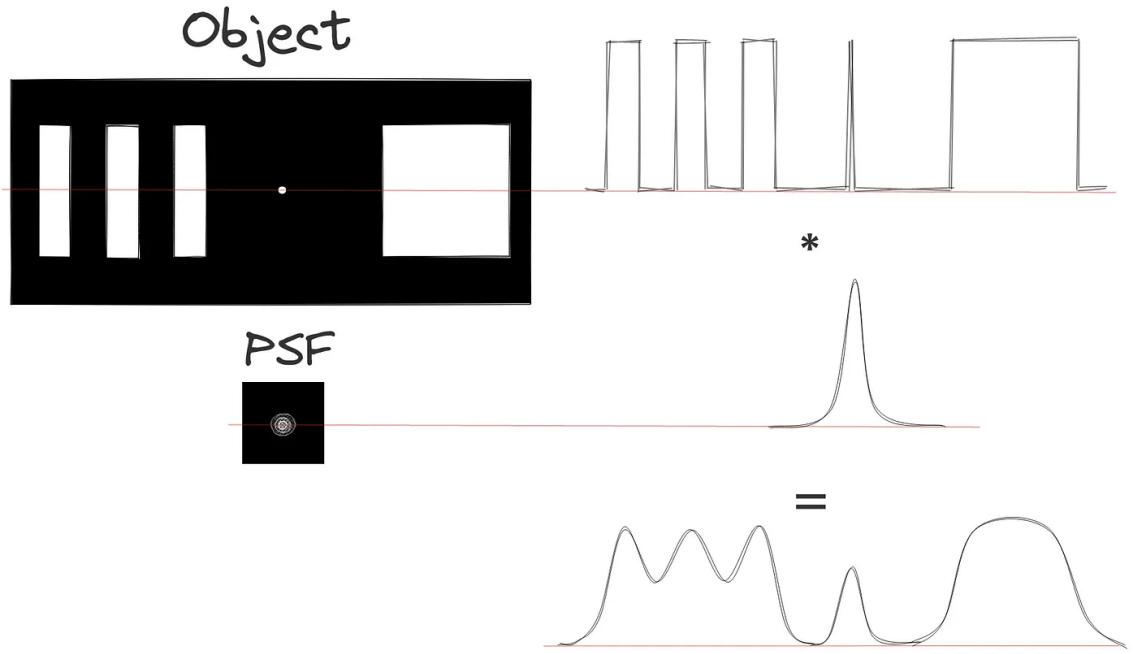
2.2 Data Harmonization

2.2.1 Effective spatial resolution and PSF

Understanding the relationship between effective spatial resolution and the Point Spread Function (PSF) is essential for accurate interpretation of super-resolution models. In our study, the term effective spatial resolution is employed to denote the spatial resolution as gauged by the ground sampling distance, rather than simply characterizing it by the pixel size. The concept of ground sampling distance (GSD) refers to the system's capability to delineate small objects within an image. The GSD is influenced by various factors, including sensor design, viewing angle, and image pre-processing techniques. These factors can be effectively modeled using single or multiple PSFs (Figure 7).

Figura 7.

A graphical depiction illustrating the Point Spread Function (PSF) of an optical imaging system.



Note. A graphical depiction illustrating the Point Spread Function (PSF) of an optical imaging system. The extent of the spread directly impacts the spatial resolution, meaning that when it is larger, the ability to distinguish individual objects decreases.

The PSF characterizes the response of an imaging system to a point source or a single pixel. It describes how the energy from an object spreads out in the image, affecting image sharpness, resolution, and spatial accuracy. Through the modeling of the PSF attributes, it is possible to emulate the functionalities of low-resolution (LR) remote sensing systems using high-resolution (HR) counterparts.

2.2.2 Spatial Co-registration

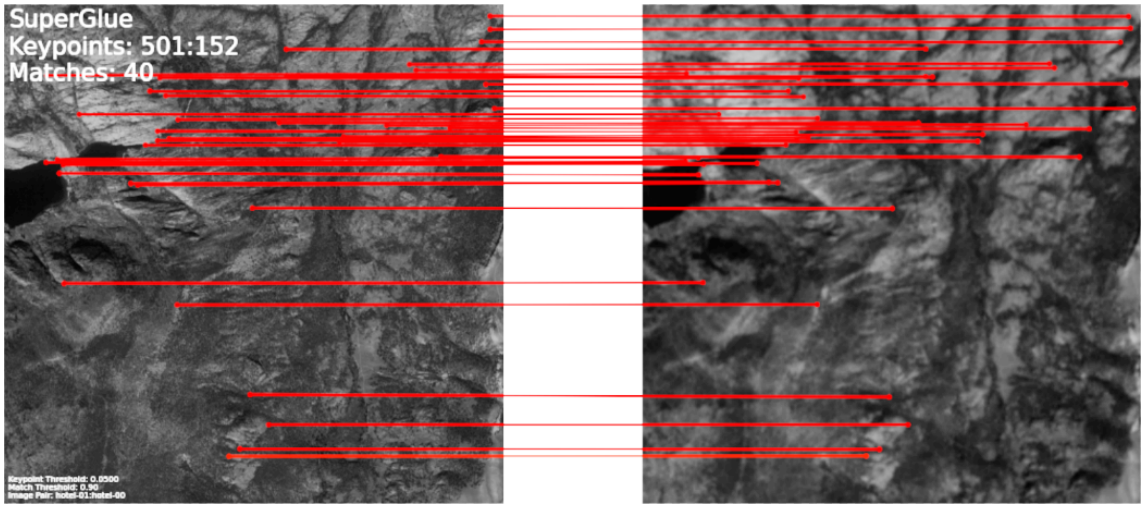
In the realm of super-resolution, spatial co-registration serves as an essential preprocessing measure to mitigate disparities between high-resolution (HR) and low-resolution (LR) sensors. In practical scenarios, even marginal misalignments can considerably in-

fluence the precision of the resultant metrics. Consequently, guaranteeing meticulous spatial co-registration becomes indispensable for deriving trustworthy and insightful super-resolution metrics.

Spatial co-registration algorithms generally encompass three primary stages. The first stage is keypoint detection, which seeks to identify prominent points within an image. The next stage involves the detection of matching points between the HR and LR images. Lastly, the misalignment errors are calculated through a polynomial model fitted to these matching points. Over recent years, a plethora of algorithms for automated image alignment have been presented. For instance, the SEN2VENμS dataset utilized the SIFT matching algorithm (Michel et al., 2022), while MuRA-T compare the alignment outcomes of Fast+VGG and SuperPoint+SuperGlue algorithms (Deshmukh et al., 2023).

Figura 8.

Matching points obtained by applying the SuperPoint + SuperGlue algorithm to a pair of images.



Note. Matching points obtained by applying the SuperPoint + SuperGlue algorithm to a pair of images, where the low-resolution (LR) image is from Sentinel-2 and the high-resolution (HR) image is from NAIP.

2.2.3 Cross-instrument calibration and harmonization

Even when two remote sensing sensors acquire data on the same day, significant variations in their values may still exist. These variations can be attributed to several factors, which include:

- **Sensor characteristics**: Each sensor has its own unique spectral response function, which can result in variations in the reflectance values, even when the atmospheric conditions and viewing angle are the same.
- **Atmospheric conditions**: The atmosphere can undergo significant changes within a few minutes or seconds. These variations can impact the quality and accuracy of the data captured by remote sensing sensors. Atmospheric scattering, for example, can introduce variations in the sensor signals, leading to differences in reflectance values.
- **Viewing geometry**: The difference between the angle at which the sensors observe the Earth's surface, can also lead to variations in the reflectance values.
- **Calibration and preprocessing**: Each sensor goes through its own calibration process to convert the raw sensor measurements into meaningful reflectance values. These variations need to be carefully accounted for and addressed during LR-HR comparison.

To minimize potential variations in reflectance between LR-HR pairs, an additional component is proposed to enhance the classical degradation model.

$$I_{LR} = \gamma[\delta(I_{HR}, \eta)] + \epsilon \tag{2.1}$$

Where I_{LR} is the LR image, I_{HR} is the HR image, η is a parameter from the degradation model δ, ϵ represent the noise, and the γ the harmonization model.

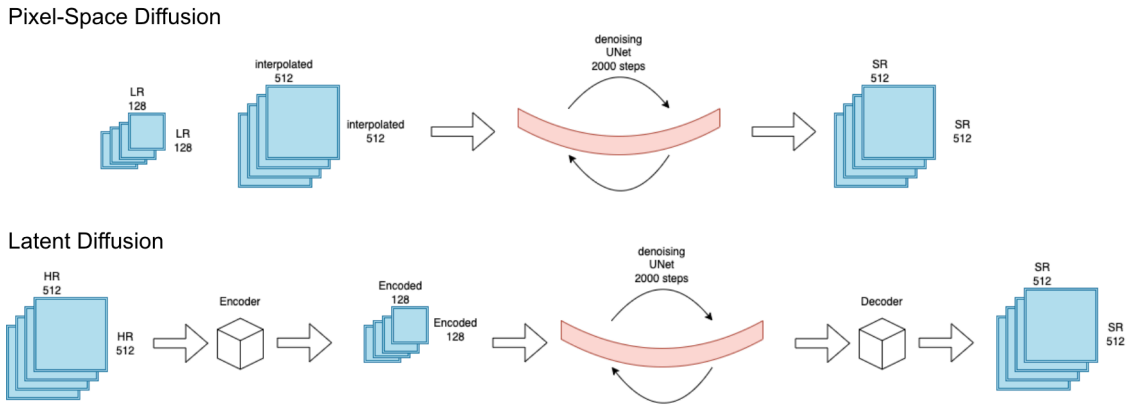
3. STATE OF THE ART

3.1 SR Design & Implementation

Diffusion (Ho et al., 2020) models have recently overtaken GAN (Goodfellow et al., 2014) models in the state-of-the-art of generative image methodologies (Croitoru et al., 2023). While GANs have been extensively used in image super-resolution (Z. Wang et al., 2021), they are unstable due to their very delicate training process. Diffusion models, which are also probabilistic models drawing results from a likelihood distribution, have proven to be easier to train and deliver better results, surpassing other methodologies in many reconstruction metrics (Saharia et al., 2021). This development has also been noticed in the remote sensing community, with recent super-resolution papers switching to diffusion-based SR methodologies (Duan et al., 2023; Jia et al., 2023; J. Liu et al., 2022).

Figura 9.

Pixel-space (above) and latent-space (below) diffusion models.



Note. Pixel-space (above) and latent-space (below) diffusion models.

3.1.1 Latent-Diffusion vs Pixel-Space Diffusion

Most SR models found in the literature use pixel-space diffusion (see Figure 9, top). In this workflow, the LR image is interpolated to the desired size. The denoising U-

Net, which is cycled through *n* times, removing a small amount of noise in each iteration. This method is very computationally expensive, since each image that needs to be super-resoluted has to pass many times through the whole network to receive the output. In general, 500 to 2000 of these iterations are used.

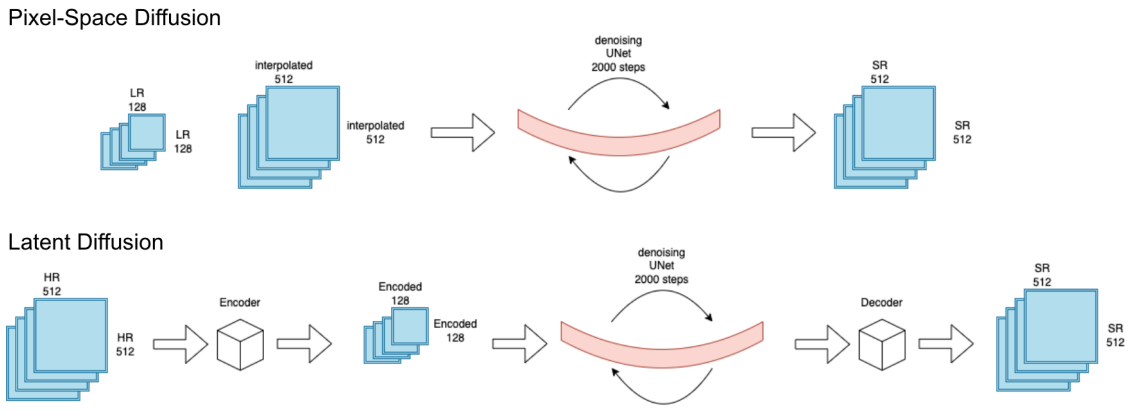
Latent diffusion models on the other hand perform the computationally expensive denoising steps on a lower-dimension representation of the input image (see fig. 9, bottom). A separate network, the autoencoder, generates a latent space and smaller than the original input space and therefore makes the denoising step much faster. After the denoising, the image is decoded and inflated again to the original desired dimensionality of the SR product. It has been shown that performing latent diffusion significantly reduces computational requirements while preserving the powerful SR capabilities of diffusion models (Rombach et al., 2022).

3.1.2 Codebase

Several code-bases related to super-resolution (SR) models have been developed and are available for exploration and implementation (see fig. 10). Among them, two unofficial implementations of the SR3 model (Saharia et al., 2021) stand out: one hosted on GitHub by Janspiry and another by KiUngSong. While these repositories provide valuable insights into SR3's functionality, they focus on pixel-space diffusion and have received feedback from the community regarding certain design choices and limitations in code quality and expandability. Additionally, these implementations tend to yield medium-quality results.

Figura 10.

Codebases in use.



Note. Codebases in use.

Another noteworthy resource is the DiffusionFastForward repository, which serves as a clear tutorial-like implementation for both pixel-space and latent-space diffusion models. However, it is relatively simple and heavily reliant on external repositories, limiting its scalability and modifiability. Despite its simplicity, it offers a good understanding of the core diffusion strategies for SR.

Finally, the latent-diffusion repository, the official code-base of the highly successful latent-diffusion model (Rombach et al., 2022), offers a comprehensive solution that encompasses various tasks such as unconditional image generation, text-prompt-based image generation, inpainting, and super-resolution. This repository is known for its versatility and well-organized structure, making it highly adaptable for further modifications and expansions. The availability of pre-trained checkpoints provided by the authors also allows for immediate demonstration of the model's strong capabilities, positioning it as a leading resource for super-resolution projects.

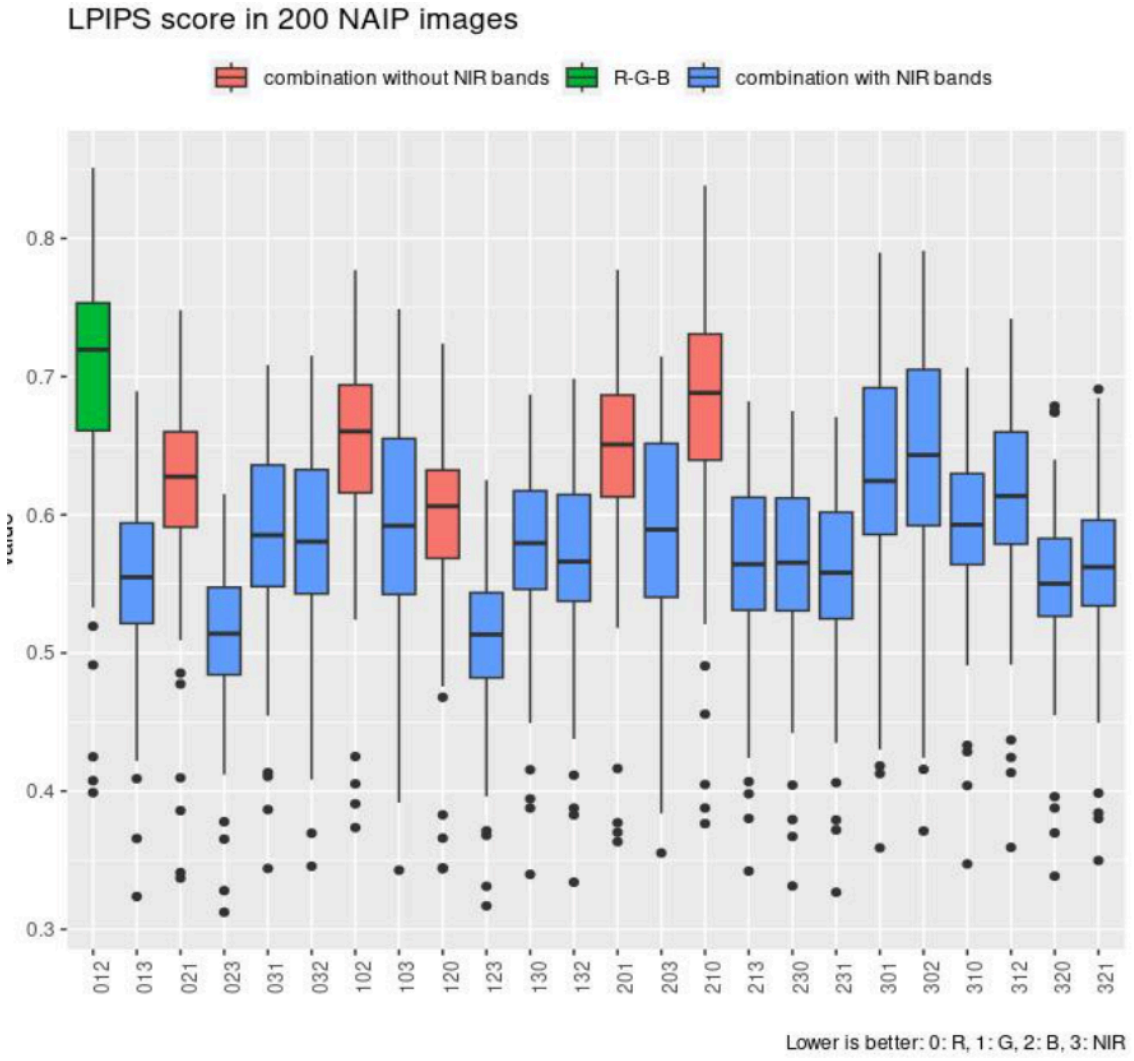
3.1.3 Autoencoder

The autoencoder under consideration is designed to compress the spatial dimension by a factor of 4, reducing tensors from $4\times512\times512$ to $4\times128\times128$ before decoding them back to their original dimensions. This approach is inspired by the *latent-diffusion* repository, as described in (Rombach et al., 2022). Ideally, such encoding and decoding processes aim to preserve image quality. Beyond reconstruction metrics, the latent space produced by the autoencoder plays a crucial role in super-resolution, as the denoising occurs after the encoder. This type of autoencoder has been effectively utilized in various applications, as demonstrated by the latent diffusion model and its adaptations in prior research (Esser et al., 2021).

In this context, models are often trained independently from the denoising U-Net, using a combination of perceptual LPIPS and a GAN component. The GAN component typically employs a classifier that distinguishes between real images and reconstructed images, where the output logits are combined with the perceptual loss and then back-propagated through the autoencoder. The LPIPS metric (R. Zhang et al., 2018) is widely used in computer vision tasks, simulating human visual perception to return a similarity score between two images. Since LPIPS is based on a VGG network (Simonyan & Zisserman, 2015) and is primarily trained on RGB data, extending it to handle more than 3 bands presents challenges, especially for 4-band models like RGB-NIR. In such cases, a common workaround involves selecting 3 bands at random during each training step and feeding them into the LPIPS model. Although unconventional, this method has been validated, as random permutations of the 4-band tensor into 3 bands have shown no significant loss in LPIPS accuracy. Figure 11 demonstrates that the LPIPS boxplot results for 200 interpolated LR-HR image pairs remain consistent, even when using spectral band combinations that were not part of the original training, thus supporting the viability of this approach.

Figura 11.

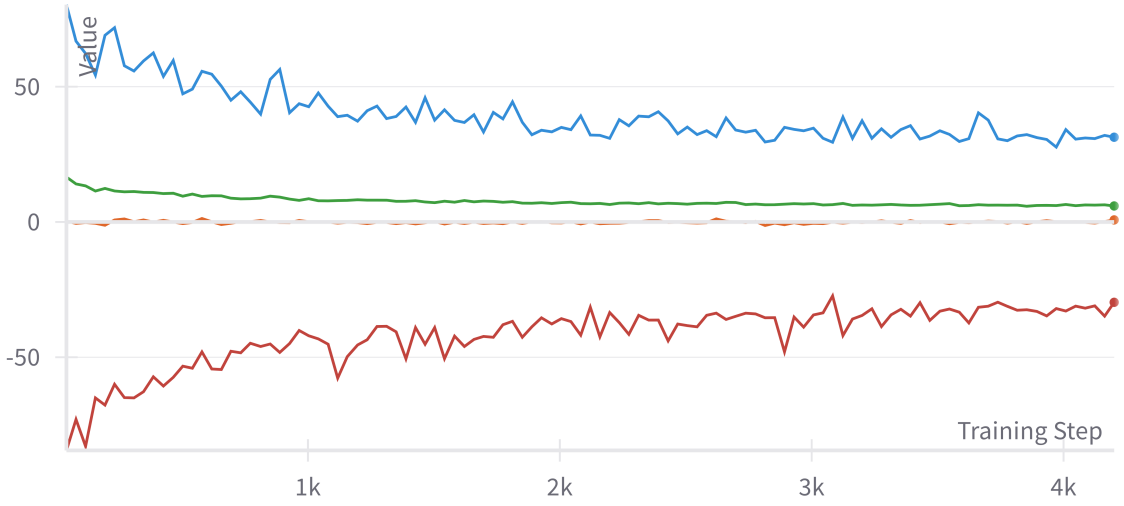
LPIPS boxplots for all possible 3 band permutations from the 4 band tensors.



Note. LPIPS boxplots for all possible 3 band permutations from the 4 band tensors.

Figura 12.

Minimum (light red), mean (orange), standard deviation (green), and maximum (blue) of the encoded value ranges.



Note. Minimum (light red), mean (orange), standard deviation (green), and maximum (blue) of the encoded value ranges during AE training. While we want the value range to be similar to the input images, if the value range is compressed too much we lose the ability to recover the images accurately.

Additionally, an autoencoder with Kullback-Leibler regularization is used to ensure that the latent space closely mirrors the input image distribution. This regularization penalizes the encoder when the encoded image strays too far from the original image distribution, which is particularly crucial during the inference stage, where the encoding process is skipped, and the image is directly denoised and super-resolved from 128×128 LR dimensions to 512×512 HR dimensions in the decoder. Maintaining the encoded image distribution as close as possible to the LR image (minus the noise) is key to achieving high-quality results.

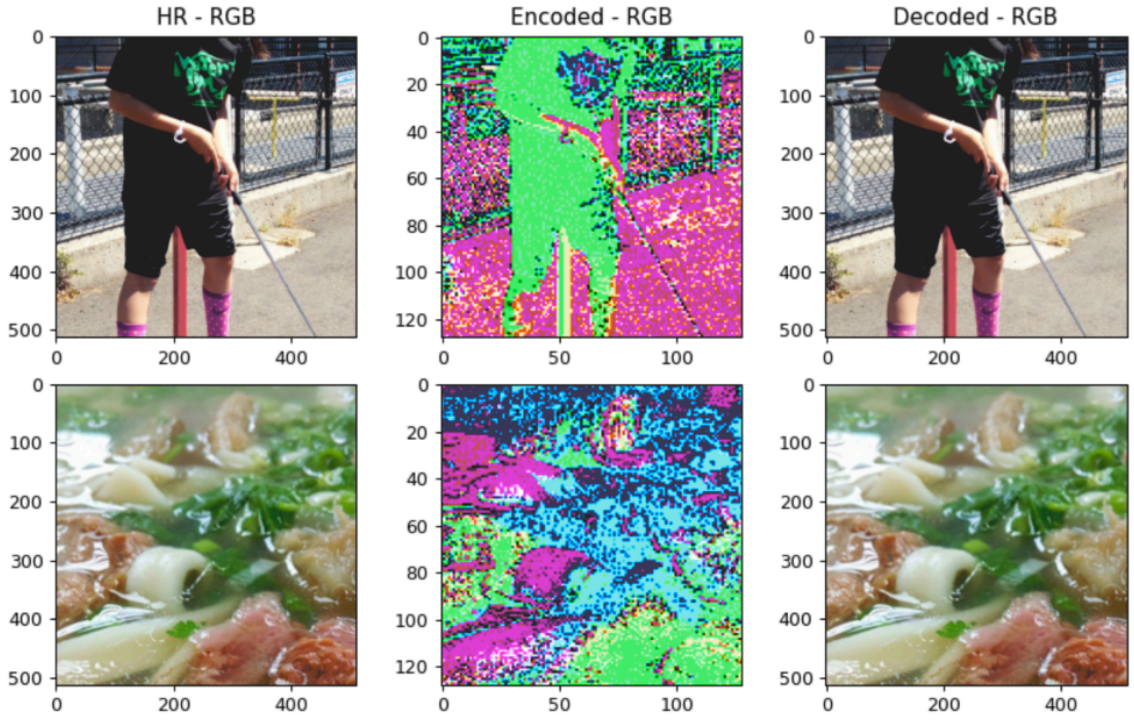
Balancing the regularization of the encoded values is essential to ensuring that the encoded image remains similar to the input, while also capturing enough variation. Over-regularization, such as reducing the extremes or standard deviation of the latent space

too much, can degrade reconstruction quality (see Figura 12). Empirical evidence suggests that a standard deviation around 5 strikes the optimal balance. By leveraging 32-bit floating-point numbers, this process effectively translates some of the high-resolution spatial complexity into numerical representation. Without this complexity, the reconstruction quality suffers.

Training the autoencoder (AE) is a delicate process. While the decoder's output converges rapidly—visually reconstructing the input image quite quickly—the resulting model checkpoint is not immediately ready for integration into the super-resolution (SR) workflow. The GAN component requires a warm-up period before its loss is weighted, and the regularization needs time to take effect. Furthermore, the autoencoder requires a large amount of satellite imagery data for training due to its size, with over 72 million parameters.

Figura 13.

Visualization of the original image, the encoded image, and the recovered image (f.l.t.r.).



Note. Visualization of the original image, the encoded image, and the recovered image (f.l.t.r.).

Figura 13 illustrates that while the image is faithfully reconstructed, the latent space might not yet be optimized for SR with the U-Net. Extensive training, careful learning rate schedules, and warm-up periods, along with latent space regularization, are required to avoid this issue.

3.1.4 Denoising UNet

The denoising probabilistic model is designed to iteratively predict a denoised version of the input image by learning the input data distribution. This process can be described as a Markov chain, where each timestep determines both preceding and following states. By learning the distribution $p(x)$ of the input data, the model is able to reverse the chain and produce denoised images. The training minimizes the variational lower bound

on the log-likelihood of the data under the model, penalizing outputs that fall on the lower end of the probability scale with respect to the target distribution.

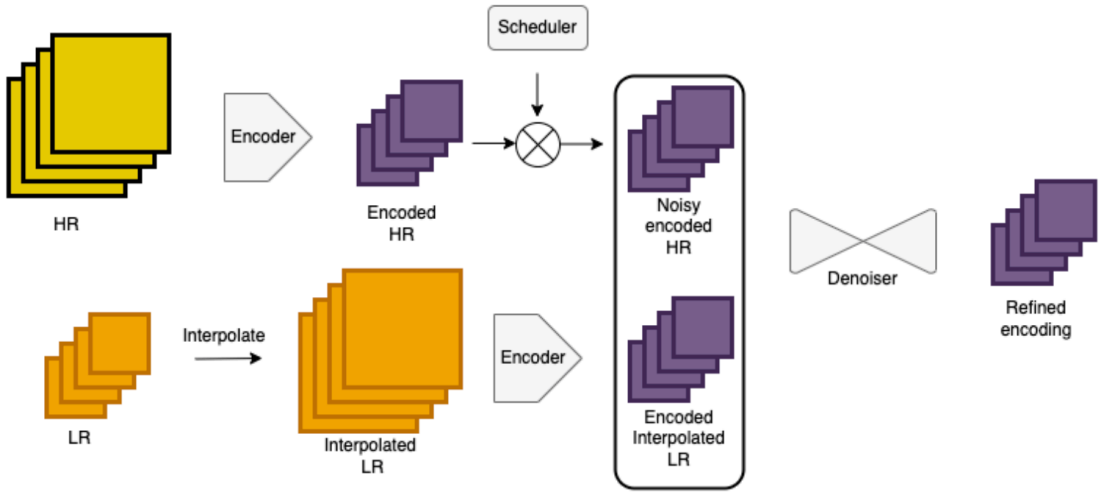
The denoising model uses a U-Net backbone, augmented with noise-adding functions that handle both the addition and removal of normally distributed noise. Since the autoencoder handles a 4x upscaling, the model can be conditioned on the LR image. Therefore, instead of starting from a pure noise image, the denoising process begins with the LR image and progressively removes noise, recovering lost quality before the image is upscaled by the autoencoder. This approach has been successfully demonstrated by (Rombach et al., 2022) for various tasks, such as conditional and unconditional image generation, inpainting, style transfer, and super-resolution. In this workflow, the U-Net is conditioned on the LR image in the original spectral domain, specifically in the RGB-NIR color space (see Figura 14).

3.1.4.1 UNet Adaptations. As the autoencoder's sampling process expects a distribution similar to that of its own training (see Figura 12), the U-Net implicitly learns to transform the input spectral distribution into the encoded spectral distribution. Although this transformation does not affect spatial feature reconstruction, it does impact the spectral consistency of the SR image. To mitigate this, we interpolate the LR image to the same dimensions as the HR image before encoding. This change relieves the U-Net from the burden of learning the mapping between the two spectral domains (Figura 14). During training, this results in a conditioning that concatenates the noise-encoded HR tensor with the same-size LR image, rather than mixing spectral domains. At inference, the encoded HR image is replaced by only the noise tensor from the scheduler.

Experimentation demonstrates that the interpolated LR image can be easily encoded and decoded by the autoencoder, even though the AE was trained on HR images (Figura 15). Although this adds an additional pass through the AE, it does not significantly increase processing time during training or inference.

Figura 14.

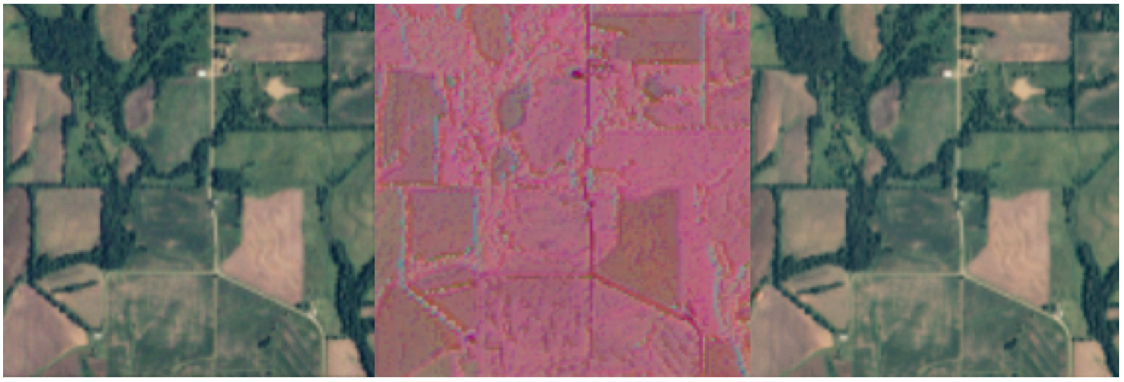
Schema of the changes to the UNet SR workflow.



Note. Schema of the changes to the UNet SR workflow. The denoising is conditioned on an interpolated and encoded LR image rather than the original LR image.

Figura 15.

Visualization of an interpolated (left), encoded (middle), and decoded (right).



Note. Visualization of an interpolated (left), encoded (middle), and decoded (right) version of the same image. This demonstrates that the AE, even though trained on HR imagery, can accurately process interpolated LR images.

3.1.5 Multispectral 20m Band SISR Methodology

The previous methodologies describe the performance of 4-band RGB-NIR SR on the S2NAIP dataset. To perform SR on the 20m bands of Sentinel-2, certain adaptations are necessary.

While this dataset does not exactly match the spectral characteristics of the 20m Sentinel-2 bands, it is already adapted between LR and HR, spatially co-registered, and degraded using degradation kernels optimized for this task. Since training is performed with random permutations and random ordering of the bands, the models ideally learn the degradation kernels and the spatial frequencies present in the 20m bands, while remaining agnostic to specific bands or values.

A 6-band autoencoder is trained on the described dataset, followed by the training of a 6-band U-Net to perform the actual SR task. These changes necessitate training from scratch. While the same spatial features and frequencies apply to both the 10m and 20m bands, they are not interchangeable and must not be used in the same model. The easiest and most efficient approach is to train separate models for each band type, as shown in Tabla 1.

Tabla 1.

Model types for the different Sentinel-2 MSI bands.

Band Number	Band Description	Resolution (m)	Model Type
B1	Coastal aerosol	60	-
B2	Blue	10	4-band
B3	Green	10	4-band
B4	Red	10	4-band
B5	Vegetation Red Edge	20	6-band
B6	Vegetation Red Edge	20	6-band
B7	Vegetation Red Edge	20	6-band
B8	NIR	10	4-band
B8A	Narrow NIR	20	6-band
B9	Water Vapor	60	-
B10	SWIR - Cirrus	60	-
B11	SWIR	20	6-band
B12	SWIR	20	6-band

Nota. Model types for the different Sentinel-2 MSI bands.

3.2 SISR Training Strategy

The datasets used in this study and their properties are detailed in Sección 2.1. This section outlines how these datasets are leveraged during the training process for SR models. Due to the high parameter count of the autoencoder and denoising U-Net, large quantities of training data are required. The original authors (Rombach et al., 2022) trained their models using the OpenImages dataset, which contains millions of images. For RGB super-resolution tasks, it is possible to fine-tune the well-optimized checkpoints from these authors using remote sensing imagery. Despite the limited availability of training images in this domain, fine-tuning can be sufficient for this task.

RGB-NIR SISR. For RGB-NIR SR, training from scratch is required, as no pre-trained models are available for 4-band input. Therefore, large datasets are needed. The following training strategy utilizes different datasets in a sequential manner to take full advantage of their strengths:

1. *CV natural images dataset*: This dataset, compiled for this project, contains approximately 280k images. Although it only includes RGB images, we generate the missing fourth band by appending the intensity level of the image as the 4th band. This method is not identical to the NIR band but provides useful spatial and spectral information, as the intensity level is normalized to the same range as the other datasets.

2. *Sentinel-2 degraded dataset*: Following initial training with natural images, remote sensing data is used. This dataset includes 240k images, although it lacks the spatial properties specific to the research question, as the LR version has a resolution of 40m and the HR version has 10m resolution. However, as it consists of remote sensing images with the same spectral properties as Sentinel-2 data, it helps train the model on relevant spectral and spatial features.

3. *NAIP dataset*: This dataset closely resembles both the spatial and spectral properties of Sentinel-2 data. The LR version is at 2.5 meters, matching the Sentinel-2 sensor's spectral characteristics, while the HR version is at 2.5m. With 250k images, this dataset alone is sufficient for training from scratch. However, pretraining on previous datasets makes the models more robust.

4. *Worldstrat dataset*: With only 3k images, this dataset is too small for standalone training or fine-tuning, as the model will overfit. Furthermore, it uses a cross-sensor approach, complicating verification of synthetic results. The final SISR models are not trained on this dataset.

20m-band Multispectral SISR. The 6-band model follows a similar training strategy. By generating synthetic 6-band S2NAIP datasets, the training process remains consistent through repetition and permutation strategies. The training proceeds as follows:

1. *CV natural images dataset*: This dataset serves as the initial training phase, leveraging its abundant data.

2. *Sentinel-2 degraded dataset*: After pretraining, the model is refined on Sentinel-2 data to focus on satellite-specific image features. However, the interpolation and SR from 80m to 20m introduce challenges with different degradation kernels.

3. *NAIP dataset*: The 250k images in this dataset serve as the best approximation for the degradation problem at hand, with spatial and spectral similarities making it ideal for training the SR models.

These training strategies enable the 4-band and 6-band models to evolve gradually, moving from a general approach to image reconstruction and SR towards the specific spatial and spectral requirements of the research question.

3.2.1 Cross-sensor vs synthetic data

Various deep learning techniques for super-resolution in earth observation have been proposed (H. Liu et al., 2021; Sdraka et al., 2022; X. Wang et al., 2022). These methods are broadly classified into cross-sensor and synthetic approaches. Cross-sensor algorithms require careful spatial alignment and spectral matching of HR and LR pairs, often constrained by the need for large-scale datasets. Synthetic methods, on the other hand, generate LR images by degrading HR images through a blur kernel and downsampling, though domain gaps between synthetic and real-world data remain a challenge (Dong et al., 2022; Qiu et al., 2023; J. Zhang et al., 2022).

4. OBJETIVES

4.1 General objective

The general aim of this project is to develop a super-resolution model capable of improving the spatial resolution of Sentinel-2 images from 10 meters and 20 meters to 2.5 meters using a combination of convolutional neural networks and multispectral data fusion techniques. The goal is to enhance both the spatial detail and spectral consistency of the images to support advanced geospatial analysis.

4.2 Specific objectives

1. **To explore and implement advanced image fusion techniques:** The aim is to combine multispectral bands from Sentinel-2 and other sources (such as NAIP) to enhance spatial and spectral resolution.

2. **To design and develop a deep learning-based super-resolution model:** The focus will be on creating a neural network model capable of processing the multispectral data to produce high-resolution outputs while preserving spectral integrity.

3. **To apply the Wald Protocol for validation:** This involves using the Wald Protocol to ensure that the super-resolved images maintain consistency with the original images in terms of spatial and spectral properties.

4. **To evaluate the performance of the proposed model:** The objective is to assess the model's performance through qualitative and quantitative metrics, comparing it with existing super-resolution techniques.

REFERENCES

Alparone, L., Aiazzi, B., Baronti, S., & Garzelli, A. (2015). *Remote sensing image fusion*. Crc Press.

Cornebise, J., Oršolić, I., & Kalaitzis, F. (2022). Open high-resolution satellite imagery: The worldstrat dataset–with application to super-resolution. *Advances in Neural Information Processing Systems*, *35*, 25979-25991.

Croitoru, F.-A., Hondru, V., Ionescu, R. T., & Shah, M. (2023). Diffusion Models in Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1-20. https://doi.org/10.1109/TPAMI.2023.3261988

Deshmukh, R., Roros, C. J., Kashyap, A., & Kak, A. C. (2023). An aligned multi-temporal multi-resolution satellite image dataset for change detection research. *arXiv preprint arXiv:2302.12301*.

Dong, R., Mou, L., Zhang, L., Fu, H., & Zhu, X. X. (2022). Real-world remote sensing image super-resolution via a practical degradation model and a kernel-aware network. *ISPRS Journal of Photogrammetry and Remote Sensing*, *191*, 155-170.

Duan, Y., Liu, Z., & Li, M. (2023). Super-resolution reconstruction of sea surface pollutant diffusion images based on deep learning models: a case study of thermal discharge from a coastal power plant. *Frontiers in Marine Science*, *10*. https://doi.org/10.3389/fmars.2023.1211981

ESA. (2019). Sentinel-2 User Handbook, ESA Standard Document, Issue I, Rev. 2.

Esser, P., Rombach, R., & Ommer, B. (2021). Taming Transformers for High-Resolution Image Synthesis.

Gargiulo, M., Mazza, A., Gaetano, R., Ruello, G., & Scarpa, G. (2019). Fast super-resolution of 20 m Sentinel-2 bands using convolutional neural networks. *Remote Sensing*, *11*(22), 2635.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. En Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 27). Curran Associates, Inc. https://encr.pw/wh8wY

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. En H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (pp. 6840-6851, Vol. 33). Curran Associates, Inc. https://l1nq.com/jq19G

Jia, S., Zhu, S., Wang, Z., Xu, M., Wang, W., & Guo, Y. (2023). Diffused Convolutional Neural Network for Hyperspectral Image Super-Resolution. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1-15. https://doi.org/10.1109/TGRS.2023.3250640

Lanaras, C., Bioucas-Dias, J., Galliani, S., Baltsavias, E., & Schindler, K. (2018). Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, *146*, 305-319.

Lillesand, T., Kiefer, R. W., & Chipman, J. (2015). *Remote sensing and image interpretation*. John Wiley & Sons.

Liu, H., Qian, Y., Zhong, X., Chen, L., & Yang, G. (2021). Research on super-resolution reconstruction of remote sensing images: A comprehensive review. *Optical Engineering*, *60*(10), 100901-100901.

Liu, J., Yuan, Z., Pan, Z., Fu, Y., Liu, L., & Lu, B. (2022). Diffusion Model with Detail Complement for Super-Resolution of Remote Sensing. *Remote Sensing*, *14*(19). https://doi.org/10.3390/rs14194834

Michel, J., Vinasco-Salinas, J., Inglada, J., & Hagolle, O. (2022). Sen2venμs, a dataset for the training of sentinel-2 super-resolution algorithms. *Data*, *7*(7), 96.

Navarro, M. A., & Sánchez, J. (2020). Sharp estimates of semistable radial solutions of k-Hessian equations. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, *150*(4), 2083-2115. https://doi.org/10.1017/prm.2019.14

Qiu, Z., Shen, H., Yue, L., & Zheng, G. (2023). Cross-sensor remote sensing imagery super-resolution via an edge-guided attention-based network. *ISPRS Journal of Photogrammetry and Remote Sensing*, *199*, 226-241.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., & Norouzi, M. (2021). Image Super-Resolution via Iterative Refinement.

Salgueiro Romero, L., Marcello, J., & Vilaplana, V. (2020). Super-resolution of sentinel-2 imagery using generative adversarial networks. *Remote Sensing*, *12*(15), 2424.

Sdraka, M., Papoutsis, I., Psomas, B., Vlachos, K., Ioannidis, K., Karantzalos, K., Gialampoukidis, I., & Vrochidis, S. (2022). Deep learning for downscaling remote sensing images: Fusion and super-resolution. *IEEE Geoscience and Remote Sensing Magazine*, *10*(3), 202-255.

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition.

Wang, X., Yi, J., Guo, J., Song, Y., Lyu, J., Xu, J., Yan, W., Zhao, J., Cai, Q., & Min, H. (2022). A Review of Image Super-Resolution Approaches Based on Deep Learning and Applications in Remote Sensing. *Remote Sensing*, *14*(21), 5423.

Wang, Z., Chen, J., & Hoi, S. C. H. (2021). Deep Learning for Image Super-Resolution: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(10), 3365-3387. https://doi.org/10.1109/TPAMI.2020.2982166

Zhang, J., Xu, T., Li, J., Jiang, S., & Zhang, Y. (2022). Single-Image Super Resolution of Remote Sensing Images with Real-World Degradation Modeling. *Remote Sensing*, *14*(12), 2895.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric.