

**Graduation Plan**

# **Super-Resolution for Enhanced Aerial Imagery**

Michalis Michalas  
6047378

July 2025

**Supervisors:** Dr.ir. Martijn Meijers  
Dr. Azarakhsh Rafiee  
**External Supervisor:** Sven Briels

# Contents

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Introduction</b>                                    | <b>1</b>  |
| 1.1      | Problem statement . . . . .                            | 1         |
| 1.2      | Scientific Relevance . . . . .                         | 2         |
| <b>2</b> | <b>Research Framework</b>                              | <b>3</b>  |
| 2.1      | Photogrammetry & True Ortho Images . . . . .           | 3         |
| 2.2      | Image Resolution & Enhancement . . . . .               | 4         |
| 2.3      | Super-Resolution in Remote Sensing . . . . .           | 4         |
| 2.3.1    | Concept . . . . .                                      | 5         |
| 2.3.2    | Mathematical Formulation . . . . .                     | 6         |
| 2.3.3    | Approaches . . . . .                                   | 7         |
| 2.3.4    | Deep-Learning-Based Approach . . . . .                 | 9         |
| 2.3.5    | Loss Functions . . . . .                               | 12        |
| 2.3.6    | Training and Test Datasets . . . . .                   | 13        |
| 2.3.7    | Quality evaluation of image super-resolution . . . . . | 14        |
| 2.4      | Research Questions . . . . .                           | 16        |
| 2.5      | Scope . . . . .  | 16        |
| <b>3</b> | <b>Methodology</b>                                     | <b>18</b> |
| 3.1      | Experimental Design . . . . .                          | 19        |
| 3.2      | General Hypotheses . . . . .                           | 20        |
| 3.3      | SR Model Hypotheses . . . . .                          | 21        |
| 3.4      | Experimental Setup Decisions . . . . .                 | 21        |
| 3.4.1    | Basic Model to be Adapted . . . . .                    | 21        |
| 3.4.2    | Data Selection . . . . .                               | 21        |
| 3.4.3    | Tile Size & Overlap . . . . .                          | 22        |
| 3.4.4    | Categorization of Urban Settings . . . . .             | 22        |
| 3.5      | Preliminary Results . . . . .                          | 22        |
| 3.5.1    | SRGAN . . . . .  | 22        |
| 3.5.2    | TransENet . . . . .                                    | 23        |
| 3.5.3    | WMCNN . . . . .  | 23        |
| 3.5.4    | Running Time . . . . .                                 | 26        |
| 3.5.5    | Observations . . . . .                                 | 27        |
| <b>4</b> | <b>Dataset &amp; Tools</b>                             | <b>28</b> |
| 4.1      | Aerial Imagery . . . . .                               | 28        |
| 4.1.1    | High-Resolution Imagery . . . . .                      | 28        |
| 4.1.2    | Low-Resolution Imagery . . . . .                       | 28        |
| 4.1.3    | Image Specifications . . . . .                         | 28        |
| 4.1.4    | Software Tools . . . . .                               | 30        |
| <b>5</b> | <b>Planning</b>  | <b>31</b> |



# 1 Introduction

High-resolution aerial imagery is a cornerstone of geospatial analysis, enabling the creation of datasets such as digital surface models (DSMs), TrueOrthos, solar irradiation maps, and point clouds. These products are crucial for applications in urban planning, environmental monitoring, and renewable energy. Factors such as sensor noise, optical distortion, and environmental interference can degrade the quality of remote sensing images [Wang et al., 2022a], while the high cost and infrequent capture of high-resolution imagery further complicate the ability to conduct detailed and continuous analyses. Single-image Super-Resolution (SR) is the process for obtaining high-resolution (HR) images from a single low resolution (LR) image. Although super-resolution remains an ill-posed and difficult problem meaning that for a single degraded image, there are multiple possible upscaled (HR) images. The challenge lies in predicting the most plausible HR reconstruction from incomplete data. Recent advances in neural networks and machine learning have enabled more robust SR algorithms that exhibit effective performance resulting to better reconstructed image. SR techniques have applications beyond geospatial fields, including medical diagnostics, object detection, and forensic analysis [Lepcha et al., 2023].

This thesis aligns with the goals of Readar B.V., a company specializing in high-quality geospatial datasets, by exploring SR methods to improve aerial imagery resolution. The research aims to support Readar’s mission of delivering accurate, consistent data products across industries such as government, utilities, and insurance.

## 1.1 Problem statement

The geospatial industry relies on high-resolution aerial imagery for generating precise datasets such as DSMs, TrueOrthos, and solar irradiation maps. Despite its importance, acquiring high-resolution imagery remains challenging due to the high costs, advanced equipment requirements, and limited acquisition frequency—often only two captures per year. This limits the availability of detailed data for applications requiring seasonal or continuous monitoring.

Low-resolution imagery is easier to access but lacks the detail needed for tasks like object detection and surface modeling. This issue is worsened by misalignment between low- and high-resolution images due to seasonal changes, vegetation growth, or moving shadows.

This thesis investigates single-image super-resolution (SISR) to address these challenges by enhancing low-resolution aerial images (e.g., 25 cm) into high-resolution outputs (e.g., 8 cm). Domain adaptation techniques will be explored to improve the robustness of SISR models across synthetic and real-world aerial images, ensuring compatibility with varying data environments and seasons.

It is important to note that while super-resolution (SR) has shown significant advancements in improving image quality, a research gap exists in addressing challenges specific to aerial imagery, particularly when datasets include HR and LR images captured during different periods. Temporal and seasonal variations, such as changes in vegetation, lighting conditions, and environmental factors, often result in misalignment and inconsistencies between images, further complicating the super-resolution process.



This research aims to bridge this gap by exploring methods to leverage both HR and LR datasets to produce accurate and consistent high-resolution outputs. By investigating how the weights trained on an SR model using one dataset can be adapted and utilized to enhance another SR model trained on temporally misaligned data, this study provides a novel solution to improve super-resolution performance in real-world scenarios. This approach ensures that even when HR and LR datasets are captured during different periods, they can still contribute effectively to generating reliable and high-quality results.

By addressing these challenges, this research advances the field of SR for aerial imagery and contributes to Readar B.V.'s pipeline by offering robust methodologies for integrating temporally inconsistent data into its geospatial analysis workflows.

## 1.2 Scientific Relevance

Super-resolution has mainly been explored for tasks involving real-world images, such as those depicting people or animals. High-resolution images provide more detailed information about locations and objects, which is essential for applications like high-definition TVs, computer screens, and portable devices. Reconstruction techniques for SR are also widely used in medical imaging, where improving resolution is crucial for accurate disease diagnosis and the identification of small anatomical features. Similarly, in the field of satellite imaging, super-resolution plays a key role in tasks such as image rectification, restoration, enhancement, and information extraction, improving clarity, reducing distortions, and enhancing geographic information [Lepcha et al., 2023].

This research focuses on aerial imagery captured over the Netherlands, examining the challenges of varying resolutions, distortions, and image characteristics unique to this domain. Aerial images are often used for tasks like military observation, environment monitoring, and weather forecast. However, aerial images captured by normal imaging devices usually have limited resolution that often does not satisfy requirements in operational tasks such as identification of small objects, analyzing texture variation or detect small scale environmental changes. Aerial images display a strong culture variability such that the image textures represent changes at different directions with various frequency characteristics [Wang et al., 2018].

Regarding the LR image as the degradation of its HR counterpart, SISR methods aim to reverse the degradation process that transforms high-resolution (HR) images into low-resolution (LR) images. Degradation modeling is central to establishing the HR-LR relationship and typically involves a combination of blurring, down-sampling, and noise [Su et al., 2024]. The degradation process provides the LR images needed to train the algorithm.

Normally, in real-world applications, obtaining HR-LR image pairs is often impractical or unattainable so solutions with SISR with unpaired images occur [Su et al., 2024]. In our approach, the HR and LR images cover exactly the same areas but are not captured simultaneously, meaning temporal changes (e.g., vegetation growth, shadow shifts, urban development) might create inconsistencies between the datasets. This setup allows the model to learn how to deal with temporal differences by focusing on the features that remain consistent across time. By incorporating data from different seasons, the robustness of the super-resolution model can be improved to handle seasonally affected features in the images. The availability of both low-resolution (LR) and high-resolution (HR) datasets, which are aligned, share the same reference system, and cover the exact same area, enables the model to focus entirely on learning the relationship between HR and LR imagery rather than addressing misalignment issues.

## 2 Research Framework

This chapter presents the research framework for the study, addressing both the practical programming tasks required and the theoretical considerations that underpin the methodology. By examining the specific types of data employed, it provides a deeper understanding of their roles within the overall pipeline.

The chapter begins with a discussion on photogrammetry and the advantages of True Ortho images compared to standard ortho images, offering essential context for the geospatial data utilized in this research. It then introduces the concept of image resolution and enhancement, highlighting their importance in improving spatial detail and visual quality. Following this, the chapter delves into super-resolution, covering its general concept, mathematical formulation, and an overview of existing approaches, with a focus on deep learning techniques. Together, these elements establish a robust foundation for the methodology outlined in subsequent sections.

### 2.1 Photogrammetry & True Ortho Images

Photogrammetry is the science and technology of extracting spatial information from images, with applications in mapping, surveying, and high-precision measurements [Förstner and Wrobel, 2016]. Aerial photogrammetry, in particular, uses overlapping images captured from above to ensure accurate alignment and rectification of datasets. These images form the foundation for generating geospatial products such as digital elevation models (DEMs) and True Ortho images.

Ortho images are created through orthorectification, a process that corrects distortions caused by terrain relief and perspective. While ortho images are geometrically accurate and can be used like maps, they do not correct distortions in elevated structures, such as buildings and bridges. True Ortho images address this limitation by incorporating detailed DEMs, which provide elevation values for each point above sea level, excluding vegetation and artificial objects. This results in an accurate 3D representation of the environment, ensuring that elevated real-world objects are rectified and aligned orthogonally to their bases.

True Ortho images are particularly valuable for tasks such as object detection, urban planning, and infrastructure analysis. However, their production is more complex than standard ortho imagery, requiring detailed DEMs and additional computational resources.

For this thesis, True Ortho images will be used as input data. Reader B.V.'s pipeline extends standard orthorectification by producing True Ortho images through precise correction of distortions in elevated objects [Kresse and Danko, 2012]. By leveraging DEMs to accurately represent the 3D topology, roofs and other elevated features are correctly positioned. This level of precision is essential for applications requiring accurate spatial alignment, such as object detection.

## 2.2 Image Resolution & Enhancement

Resolution can take on different meanings depending on the imaging application. Spatial resolution refers to the pixel density within an image and is typically measured as pixels per unit area. Radiometric resolution corresponds to the bit depth of the image, while temporal resolution relates to the number of frames captured per second. Spectral resolution, on the other hand, describes the number of color planes or spectral bands present in the image. In the context of this research, **Super-Resolution (SR)** specifically refers to achieving a higher **spatial resolution** than that originally captured by the camera sensor.

In imaging systems, image resolution can be improved either by decreasing pixel size through advancements in sensor manufacturing or by increasing the sensor's chip size. However, due to the physical constraints of imaging systems, employing algorithmic techniques offers a more cost-effective solution for enhancing image resolution [Vishnukumar et al. \[2014\]](#).

The resolution of an image refers to the density of pixels within it, which determines the amount of visual information the image can convey, often described as pixels per inch (PPI). In a low-resolution image, the pixels are fewer in number, and if those few pixels are too large, it can result in a blocky or pixelated appearance. Lower resolution can make small objects hard to distinguish, as they may overlap, be hidden, or blend together, making detection and recognition more difficult and less accurate.

High-resolution images, on the other hand, have more pixels per inch (PPI) and consist of a greater number of smaller pixels, allowing for finer detail and better visual quality. These images retain clarity even when enlarged or stretched, as the higher pixel density ensures that visual information is preserved. As a result, objects in high-resolution images are more visible and easier to recognize, improving their suitability for tasks requiring detailed image analysis.

The basic principle of image enhancement is to modify the information contribution of an image so that it is more suitable for a specific application [\[Singh and Mittal, 2014\]](#). Traditional image enhancement techniques typically fall into two categories: spatial domain and frequency domain processing. Spatial domain methods work directly with the pixels of an image, employing techniques like modified histogram approaches and improved unsharp masking methods. On the other hand, frequency domain methods transform the image into the frequency domain using mathematical functions such as Fourier Transform (FT), Discrete Cosine Transform (DCT), or Discrete Wavelet Transform (DWT). Image processing is then performed based on the characteristics of the frequency domain before converting the result back to the original image space [\[Qi et al., 2022\]](#). In the context of image super-resolution, these enhancement principles have been extended with advance machine learning techniques which are mentioned in the next section.

## 2.3 Super-Resolution in Remote Sensing

Remote sensing images differ significantly from natural images. Captured from high altitudes using aerial photography or satellites, they often depict large-scale scenes like forests, rivers, industrial zones, and airports, which contain small objects and diverse spatial distributions. These images are also affected by varying weather conditions, with factors such as sensor lighting, cloud cover, and fog influencing their clarity. In the context of super-resolution, reconstructing remote sensing images demands specialized approaches. For example, in forest and grassland scenes where object colors are similar, relying solely on color features can be ineffective. By leveraging texture characteristics, super-resolution methods

can distinguish between "rough" forests and "smooth" grass, improving classification and reconstruction [Wang et al., 2022a].

### 2.3.1 Concept

Super-resolution is a process that aims to reconstruct a high-resolution (HR) image from its low-resolution (LR) counterpart. In theory, any image has a *ground truth high-resolution version*, which may exist physically or be purely theoretical. For an image to be low resolution, it means that at some point, a *degradation function*  $D$  has acted on the high-resolution image. This degradation can include processes such as blurring, downsampling, or adding noise, with a factor  $\gamma$  representing the degree of change. This mapping from HR to LR is typically unknown and challenging to reverse as illustrated in Figure 2.1 (up left).

The reverse process, *super-resolution*, seeks to estimate the HR image from the LR input using a model  $F$ . The goal of  $F$  is to recover the lost information and approximate the ground truth as closely as possible (Figure 2.1 down left). However, this task is inherently ill-posed: because information is lost during degradation, there are infinite possible reconstructions of varying quality (Figure 2.1 down right). The key challenge in super-resolution is to develop a model capable of producing reconstructions that are both accurate and visually convincing, based on specific evaluation metrics. An overview of the super-resolution task is illustrated in Figure 2.2.

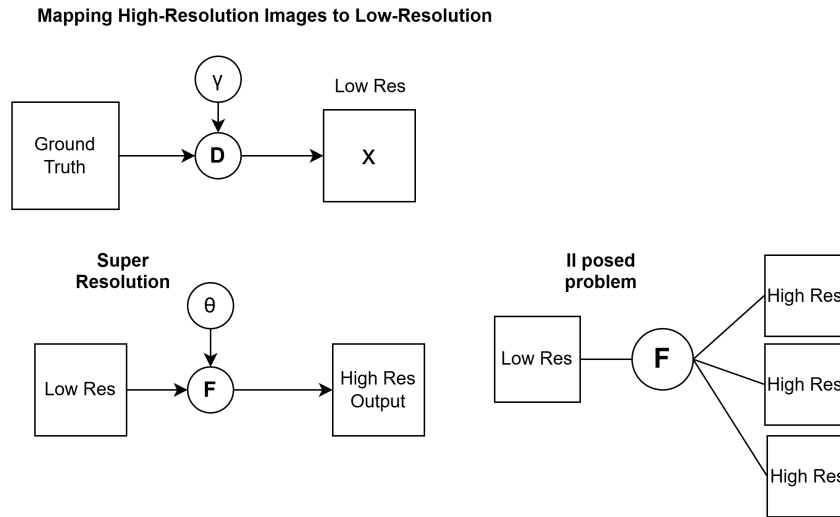


Figure 2.1: Concept of Super resolution

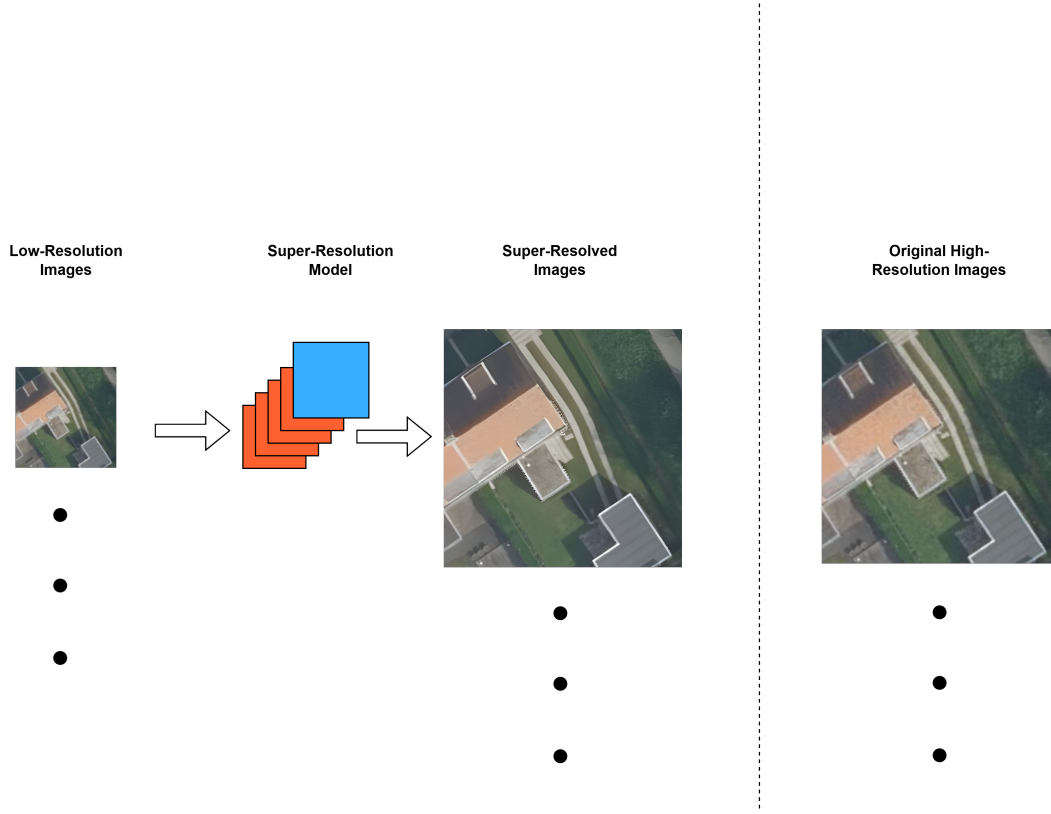


Figure 2.2: Overview of the super-resolution task: starting with low-resolution images, a super-resolution network is designed to enhance their quality, producing super-resolved versions of the input images.

### 2.3.2 Mathematical Formulation

According to [Kawulok et al., 2024], “SR poses an inherently ill-posed problem, wherein applying various degradation processes  $\mathbb{F}$  can yield many different low-resolution (LR) images  $I_{LR}$  from a single high-resolution (HR) image and vice versa”. To gain a deeper understanding of this challenge, we now delve into the mathematical formulation of the degradation and reconstruction processes, defining key functions and parameters that underpin super-resolution techniques. The degradation process and its mathematical modeling, are described in detail in [Anwar et al., 2020]. Let us denote a Low -Resolution (LR) image as  $y$  and the corresponding high-resolution (HR) image as  $x$ , then the degradation process is given as:

$$y = \Phi(x; \theta_\eta), \quad (2.1)$$

where  $\Phi$  is the degradation function, and  $\theta_\eta$  denotes the degradation parameters (such as the scaling factor, noise, etc.). In a real-world scenario, only  $y$  is available while no information about the degradation process or the degradation parameters  $\theta_\eta$ . SR aims to reverse this degradation and approximate the ground-truth image  $x$  by estimating image  $\hat{x}$  as:

$$\hat{x} = \Phi^{-1}(y, \theta_s), \quad (2.2)$$

where  $\theta_s$  are the parameters for the function  $\Phi^{-1}$ . The degradation process is unknown and can be quite complex. It can be affected by several factors, such as noise (sensor and speckle), compression, blur (defocus and motion), and other artifacts. To address this complexity, many studies adopt a more detailed degradation model instead of relying solely on Equation (1). This refined model is given as:

$$y = (x \otimes k) \downarrow_s + n, \quad (2.3)$$

where  $k$  is the blurring kernel and  $x \otimes k$  is the convolution operation between the HR image and the blur kernel,  $\downarrow_s$  is a downsampling operation with a scaling factor  $s$ . The variable  $n$  denotes the additive white Gaussian noise (AWGN) with a standard deviation of  $\sigma$  (noise level). In image super-resolution, the aim is to minimize the data fidelity term, which is its the degree to which data can be trusted to be accurate and reliable, associated with the model  $y = x \otimes k + n$ , as,

$$J(\hat{x}, \theta_s, k) = \underbrace{\|x \otimes k - y\|}_{\text{data fidelity term}} + \alpha \underbrace{\Psi(x, \theta_s)}_{\text{regularizer}}, \quad (2.4)$$

where  $\alpha$  serves as a balancing parameter between the data fidelity term and the image prior  $\Psi(\cdot)$ .

Super-resolution methods can be categorized based on how they utilize the image prior—that is, the pre-existing knowledge or assumptions about the image’s properties during the reconstruction process. An image prior represents a set of constraints or statistical properties believed to be true for the images being processed, guiding the super-resolution algorithm. These methods can be divided into categories such as prediction-based methods, interpolation-based methods, edge-based methods, statistical methods, patch-based methods, and deep learning methods [Yang et al., 2014]. For example, interpolation-based methods are non-adaptive and rely on local neighborhood information, making them computationally efficient but prone to issues such as aliasing and blurring. Statistical methods address the ill-posed nature of super-resolution by leveraging image priors to capture domain knowledge of natural images. These priors include Gaussian priors, Markov random field (MRF) priors, sparsity priors, and low-rank priors. However, due to the complex structure of real-world images, many of these priors struggle to accurately model image properties. Classical methods for single image super-resolution, such as linear interpolation or reconstruction-based approaches, often produce undesirable artifacts and over-smoothing in the reconstructed HR image, particularly around edges Vishnukumar et al. [2014].

This research focuses specifically on methods that employ deep neural networks to learn and apply the image prior. These methods have the ability to automatically learn hierarchical features directly from data, bypassing the need for manually engineered priors. Deep learning techniques have demonstrated excellent performance in handling large, complex datasets like aerial imagery. Their ability to effectively model high-frequency details, suppress noise, and preserve edges makes them particularly suitable for reconstructing detailed and accurate high-resolution representations from low-resolution aerial images. The preservation of edges is beneficial for processes like solar panel detection or green roof detection, where sharp and distinct boundaries are critical for accurate identification. This suitability, combined with their scalability and adaptability, underscores their relevance to the goals of this research.

### 2.3.3 Approaches

Super Resolution (SR) algorithms are designed to enhance the spatial resolution of digital images. Early SR approaches relied on techniques such as nonuniform interpolations, fre-



quency domain analysis, deterministic and stochastic regularization, and projection onto convex sets [Kawulok et al., 2024]. In specific remote sensing tasks such as pansharpening and hyperspectral/multispectral image fusion, classical methods have included component substitution (CS), multi-resolution analysis (MRA), variational optimization (VO), spectral unmixing, and Bayesian models. However, for this research, as a first approach, no additional spectral bands are used for the reconstruction tasks.

Over the past decade, advancements in computational power have led to the dominance of deep neural networks in state-of-the-art super-resolution (SR) systems. SR approaches based on the nature of the input data can be categorized as: single-image super-resolution (SISR), which enhances the resolution of a single low-resolution (LR) image and multi-image super-resolution (MISR), which reconstructs a higher-resolution output using multiple LR images. Our approach is aligned with SISR as we don't have multiple shifted LR observations of the same scene.

The use of convolutional neural networks (CNNs) for super-resolution (SR) began with the introduction of SRCNN in 2015 by Dong et al.. This architecture featured three convolutional layers designed for feature extraction, nonlinear mapping, and reconstruction. SRCNN required low-resolution (LR) images to be pre-upsampled to the target resolution using bicubic interpolation. Building on this, VDSR was introduced by Kim et al. [2016a], focusing on predicting the residual image. The reconstructed HR image is subsequently obtained by adding the residual image to the bicubically upsampled LR image.

As deeper neural networks are more challenging to train. Techniques such as ResNet [He et al., 2016] introduced a residual learning framework to facilitate the training of significantly deeper networks compared to earlier architectures. In ResNet, the layers are explicitly reformulated to learn residual functions with respect to the layer inputs, rather than attempting to learn unreferenced functions directly (directly mapping inputs to outputs).

Furthermore, recursive networks are based on the concept of parameter sharing among convolutional layers. This approach reduces both the number of trainable parameters and the computational complexity. DRCN, introduced by Kim et al., is a deeply recursive network that applies the same convolutional layer multiple times. The outputs from all intermediate shared convolutional blocks, along with the final output, are sent to the reconstruction layer, which generates the high-resolution image by utilizing all these inputs.

With high computational cost being the common problem for most of the SR techniques, postupsampling frameworks were suggested in order to replace the traditional upsampling methods with learnable upsampling layers. These frameworks construct an end-to-end architecture in which the whole feature extractions are implemented in a low-dimensional space [Lei et al., 2022]. However, a drawback of these techniques is that the HR images are directly reconstructed at the final stage, without intermediate enhancement of feature representation. This increases the difficulty of training and limits reconstruction accuracy.

This problem is addressed by transformer-based enhancement network (TransENet) described by Lei et al. which after the upsampling layers, both high-dimensional and low-dimensional features are utilized to improve the network's ability to represent fine details. This ensures that the network effectively learns from both detailed (high-dimensional) and broader (low-dimensional) feature contexts.

Generative adversarial networks (GANs), like the one described by Ledig et al. [2017] have delivered impressive results by focusing on perceptual quality, even though they still face challenges like hallucination artifacts and training instability. Their main goal is to generate images that are visually pleasing rather than strictly matching reference images pixel by pixel. GANs work with two components: a generator, which creates synthetic images, and a discriminator, which evaluates whether the images are real or fake. Since the generator is designed to "fool" the discriminator, it prioritizes creating plausible-looking

images rather than ensuring they are perfectly aligned with the ground truth. This can result in hallucination, where parts of the generated images look realistic but deviate from the actual content. While this can be useful for tasks like artistic rendering, it becomes problematic for applications that demand precise, ground truth-aligned outputs, such as medical imaging or geospatial analysis. Table 6.2 in the Appendix summarizes the super resolution approaches with a small description of how they operate.

In this research, the methodology begins with Single Image Super-Resolution (SISR) in the first iteration. Here, only high-resolution (HR) images are used to train the model, aiming to produce super-resolution results that closely approximate the ground truth. To avoid confusion, this output can be defined as the Generated High-Resolution (GHR) image.

The second iteration shifts towards an image fusion approach, incorporating the weights learned from the first iteration. This step utilizes both low-resolution (LR) and HR images to further refine the super-resolution outputs. The key challenge lies in the fact that the HR and LR datasets were captured during different time periods, potentially reflecting varying conditions. This aspect tests the adaptability and robustness of the model's learned weights. Specifically, it examines how well the weights, trained under different conditions, can enhance the accuracy of super-resolution outputs for images captured at different times.

### 2.3.4 Deep-Learning-Based Approach

Unlike traditional super-resolution approaches, deep learning relies on neural networks to automatically learn features, complex patterns and representations from the data, making the process more efficient and accurate. The goal of deep learning in super-resolution is to uncover the feature distribution within data by learning a hierarchical representation of its underlying characteristics [Wang et al., 2022a]. This is achieved through advanced network architectures, optimization techniques, and loss function designs, while addressing the challenges posed by the ill-posed nature of super-resolution.

Deep learning methods rely on learning mappings directly from paired low-resolution (LR) and high-resolution (HR) image datasets. The relationship between an LR input image  $I_{LR}$  and its corresponding HR output image  $I_{HR}$  is modeled by a neural network  $f_{\theta}$ , parameterized by weights  $\theta$ , as:

$$I_{HR} = f_{\theta}(I_{LR}), \quad (2.5)$$

where  $f_{\theta}$  learns to map  $I_{LR}$  to  $I_{HR}$  by minimizing a loss function  $\mathcal{L}$ . This loss function quantifies the difference between the predicted HR image  $\hat{I}_{HR}$  and the ground truth HR image  $I_{HR}$ :

$$\mathcal{L} = \|\hat{I}_{HR} - I_{HR}\|^2, \quad (2.6)$$

where  $\|\cdot\|^2$  represents the mean squared error (MSE) loss, commonly used in deep learning-based SR models. The optimization process adjusts  $\theta$  to minimize  $\mathcal{L}$ , improving the quality of the predicted HR image.

Deep learning methods excel in handling the ill-posed nature of super-resolution by leveraging data-driven learning to infer missing high-frequency details. This approach allows for a more effective and robust solution compared to traditional techniques, as it directly learns the complex mappings between LR and HR images.

An essential aspect of deep learning for super-resolution is the choice of appropriate loss functions. These functions guide the training process by evaluating and minimizing the errors between the reconstructed and ground truth images.



## Convolutional Neural Networks (CNNs)

Deep learning methods such as Convolutional Neural Networks (CNNs) are widely used for super-resolution tasks, leveraging convolutional layers to extract hierarchical features from low-resolution (LR) images and reconstruct high-resolution (HR) outputs. In CNN architectures lower layers capture low-level features and higher layers capture more complex and abstract information [Kawulok et al. \[2024\]](#). The fact that they can capture multi-level features makes them produce high-quality HR outputs. An outline of a the process of CNN based methods for Super Resolution is illustrated in Figure 2.3.

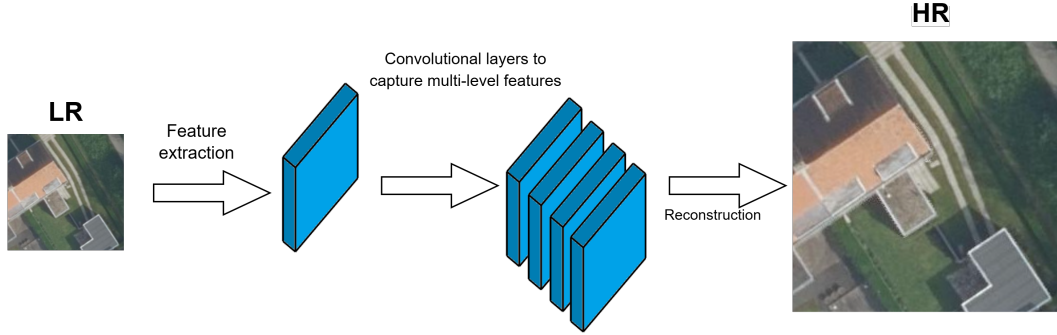


Figure 2.3: Process of CNN based methods for Super Resolution

## Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) introduce an adversarial framework for super-resolution, consisting of a generator that creates high-resolution images and a discriminator that evaluates their quality. This approach excels at producing perceptually realistic and visually pleasing results, addressing the over-smoothing issues often seen in CNN-based methods. GANs are particularly effective in scenarios requiring high perceptual quality, such as aerial and satellite imagery analysis. GANs also incorporating adversarial training to enhance the visual realism of the generated HR [Ledig et al. \[2017\]](#) which may not be suitable for certain use cases. An illustration of the process is shown in Figure 2.4

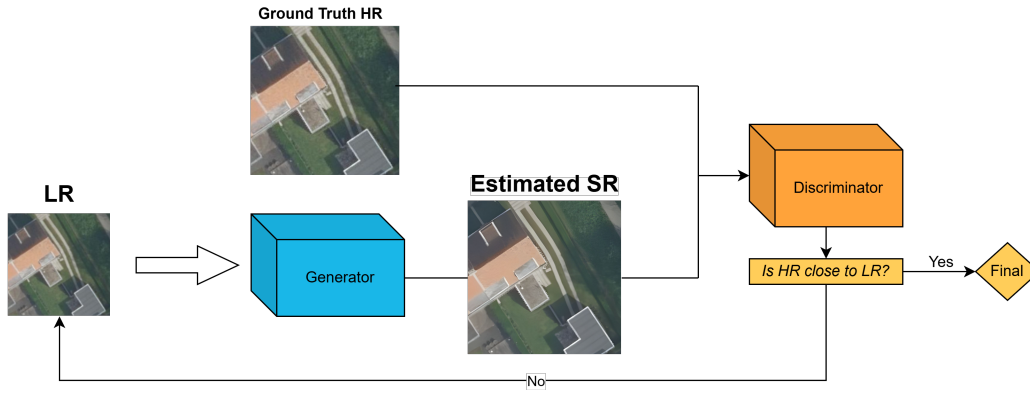


Figure 2.4: Process of GAN based methods for Super Resolution

### Transformer Based

Transformer-based super-resolution methods build upon the core principles of transformers, which were originally developed for natural language processing (NLP) tasks. The key advantage of transformers lies in their self-attention mechanism, which can model long-range dependencies in images and recover high-frequency information. This ability is critical for reconstructing texture details and improving image quality. While transformers were initially designed to model dependencies in sequential text, recent advancements have demonstrated their capacity to address limitations of convolutional neural networks (CNNs) by overcoming inductive bias through self-attention [Wang et al. \[2022b\]](#).

In recent years, a hybridization of deep learning and transformer models has emerged as a successful strategy for image super-resolution tasks. For example, [Lei et al. \[2022\]](#) introduced a transformer-based multi-stage enhancement structure, known as TransENet, which fuses multi-scale high- and low-dimensional features. In this architecture, encoders embed multi-level features during feature extraction, while decoders fuse these embeddings to reconstruct enhanced high-resolution images.

This hybrid approach combines the global context modeling capabilities of transformers with the local feature extraction strengths of CNNs. By leveraging these complementary methods, TransENet effectively captures long-range dependencies and detailed textures, making it particularly well-suited for the complex and diverse structures found in remote sensing images. Experimental results demonstrate that TransENet outperforms several state-of-the-art methods, achieving superior super-resolution results and improving overall image quality. Figure 2.5 illustrates the workflow of a transformer-based super-resolution method that also uses CNN. The low-resolution (LR) image is first processed through a CNN block to extract local features. These features are then passed through transformer blocks, where the self-attention mechanism models long-range dependencies and refines the features globally. The encoder is responsible for extracting features from the input data and gradually reducing its spatial dimensions while increasing its feature representation. The decoder fuses these multi-scale features and reconstructs the high-resolution (HR) image. This hybrid approach leverages the strengths of both CNNs and transformers to achieve superior super-resolution results.

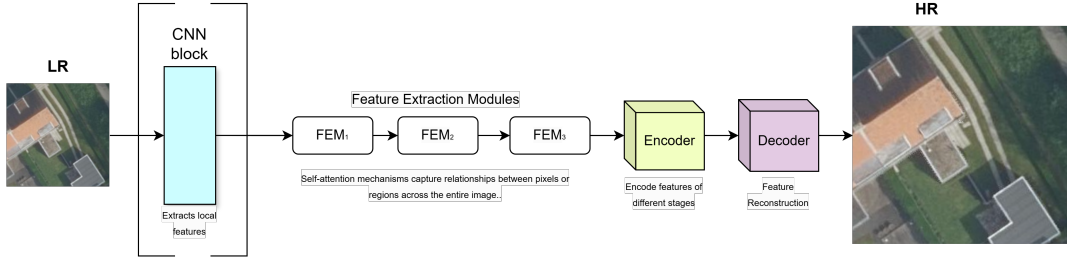


Figure 2.5: Hybrid Transformer Workflow

### 2.3.5 Loss Functions

Once the SR model generates the reconstructed images, various loss functions are used to calculate the error between them and the ground truth. A smaller loss function value indicates a more robust model. Changes in the loss function reflect the gap between the model's current training state and the expected outcome. These metrics evaluate the model's performance and simultaneously guide it during the training process. This section presents the commonly used loss functions.

**Pixel Loss** calculates the difference between reconstructed images and ground truth images using pixel values. Common loss functions for this purpose include Mean Squared Error (MSE), Mean Absolute Error (MAE), and Charbonnier Error. MSE is also known as L2 loss, while MAE and Charbonnier Error are referred to as L1 and improved L1, respectively. The mathematical expressions for these losses are as follows:

$$L_2 = \frac{1}{n} \sum_{i=1}^n \left( I_{SR}^i - I_{HR}^i \right)^2, \quad (2.7)$$

where  $n$  represents the number of training samples,  $I_{SR}^i$  denotes the reconstructed image, and  $I_{HR}^i$  corresponds to the ground truth high-resolution image.

The MSE loss function is characterized by its smoothness, continuity, and differentiability, which make it well-suited for gradient descent optimization. As the error decreases, the gradient magnitude reduces, enabling the algorithm to converge quickly. However, due to its squaring operation, MSE assigns higher weights to large errors, making it highly sensitive to outliers. This sensitivity can result in overly smoothed and blurred reconstruction outputs, particularly in regions with high-frequency details. To address this limitation, the L1 loss (MAE) is used, which is less sensitive to outliers and maintains a stable gradient for a wider range of input values and expressed as:

$$L_1 = \frac{1}{n} \sum_{i=1}^n \left| I_{SR}^i - I_{HR}^i \right|, \quad (2.8)$$

However, L1 loss has non-differentiable points, which can hinder convergence during training. To deal with this limitation the Charbonnier Loss was introduced by [Lai et al., 2017], which is a smooth approximation of L1 loss and is defined as:

$$L_{\text{char}} = \frac{1}{n} \sum_{i=1}^n \rho \left( I_{SR}^i - I_{HR}^i \right), \quad \text{where } \rho(x) = \sqrt{x^2 + \epsilon^2}, \quad (2.9)$$

Here,  $\epsilon$  is a small constant (typically  $10^{-3}$ ) added to improve numerical stability and ensure smoothness at zero gradients. However, pixel-based loss functions overlook the perceptual quality and texture of the reconstructed image, often resulting in the loss of high-frequency details. As a result, it becomes challenging to achieve high-quality reconstructed images.

**Perceptual Loss** addresses these limitations by averaging and often produce images with smoother textures that lack visual appeal. These loss functions became popular when GAN methods were introduced. Perceptual loss is optimized by minimizing the distance between extracted features, thereby enhancing the perceptual quality of the image. In SRGAN [Ledig et al., 2017], perceptual loss is represented as the weighted sum of content loss and adversarial loss, expressed in Equation 2.10.

$$L^{SR} = L_{\text{content}} + 10^{-3}L_{\text{Gen}}(I_{SR}) \quad (2.10)$$

where  $L^{SR}$  is the perceptual loss,  $L_{\text{content}}$  represents the content loss, and  $L_{\text{Gen}}(I_{SR})$  denotes the adversarial loss.

**Content Loss** focuses on evaluating the similarity between the reconstructed image and the reference image at a perceptual level, aligning with how the human eye perceives visual details. It is defined as:

$$L_{\text{content}} = \frac{1}{n_l} \sqrt{\sum_{i,j} \left( \Phi_{ij}^{(l)}(I) - \Phi_{ij}^{(l)}(\hat{I}) \right)^2} \quad (2.11)$$

Here,  $n_l$  denotes the number of pixels in the feature map of the  $l$ -th layer, and  $\Phi_{ij}^{(l)}(I)$  and  $\Phi_{ij}^{(l)}(\hat{I})$  represent the feature maps extracted from the  $j$ -th convolution before the  $i$ -th pooling layer in the  $l$ -th layer for the ground truth and reconstructed images, respectively.

Other loss functions, such as texture loss and adversarial loss, are also commonly used to enhance the quality of reconstructed images. However, these will not be described in detail here, as the final choice of loss functions will depend on the specific requirements of our methodology, which is still in the early stages of development.

### 2.3.6 Training and Test Datasets

The success of deep-learning-based SR methods relies heavily on high-quality training and testing datasets. Diverse datasets have been developed to address various SR tasks, ranging from natural to remote sensing images. Representative training datasets mostly including images from people, animal, scenery, decoration, plant, etc. include:

- **DIV2K**: Comprising 800 training images, 100 validation images, and 100 test images, this dataset is a standard for SR tasks.
- **BSDS300, BSDS500**: Widely used for benchmarking SR models.
- **Set5, Set14, Urban100**: Classic test datasets for evaluating SR performance.

Remote sensing datasets, tailored to specific geospatial tasks, often differ from natural image datasets. Some notable examples include:

- **AID**: Contains 10,000 images ( $600 \times 600$  pixels) featuring airports, beaches, deserts, etc.

- **RSSCN7**: Includes 2800 images categorized by season and scale, depicting farmland, residential areas, and industrial zones.
- **WHU-RS19**: Comprises 950 images representing 19 scene categories, such as ports and parking lots.
- **UC Merced**: Features 2100 images ( $256 \times 256$  pixels) across 21 categories, including forests, rivers, and agricultural land.

Our dataset consists of aerial images of the Netherlands captured at two different times and will be described in detail in Chapter 4. These are aerial images include multiple elements, similar to the datasets briefly described earlier. However, the final categories will need to be determined and refined based on our use case.

Table 6.1 provides an overview of commonly used SR datasets, including their size, resolution, and content description and can be found in Chapter 6.

### 2.3.7 Quality evaluation of image super-resolution

The evaluation index of image reconstruction quality can reflect the reconstruction accuracy of an SR model and in this section, the evaluation methods of image reconstruction quality and reconstruction efficiency will be discussed. Evaluating the quality of reconstructed images is crucial due to the widespread use of super-resolution (SR) techniques. Image quality refers to the visual attributes of an image, and evaluation methods can be broadly categorized into subjective and objective assessments. Subjective evaluation assesses image quality based on human perception, focusing on how natural or realistic the image looks. While it reflects human judgment, it is inefficient and challenging to scale. In contrast, objective evaluation relies on numerical algorithms to measure quality, making it more practical. Full-reference objective methods are commonly used for image quality assessment.

**Peak Signal-to-Noise Ratio (PSNR)** is one of the most commonly used objective metrics in SR [Wang et al., 2004]. For a ground truth image  $I_y$  with  $N$  pixels and a reconstructed image  $I_{SR}$ , PSNR is defined as:

$$\text{PSNR} = 10 \cdot \log_{10} \left( \frac{L^2}{\text{MSE}} \right),$$

where  $L = 255$  for an 8-bit grayscale image and the Mean Squared Error (MSE) is:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (I_y - I_{SR})^2.$$

PSNR is computationally simple and has a clear physical meaning, but it focuses purely on pixel-level differences and does not account for human visual perception.

**Structural Similarity Index (SSIM)** is designed to measure the perceptual similarity between two images in terms of brightness and contrast. SSIM is defined as:

$$\text{SSIM} = \left[ l(I_{SR}, I_y)^\alpha \cdot c(I_{SR}, I_y)^\beta \cdot s(I_{SR}, I_y)^\gamma \right],$$

where:

$$l(I_{SR}, I_y) = \frac{2\mu_{I_{SR}}\mu_{I_y} + C_1}{\mu_{I_{SR}}^2 + \mu_{I_y}^2 + C_1},$$

$$c(I_{SR}, I_y) = \frac{2\sigma_{I_{SR}}\sigma_{I_y} + C_2}{\sigma_{I_{SR}}^2 + \sigma_{I_y}^2 + C_2},$$

$$s(I_{SR}, I_y) = \frac{\sigma_{I_{SR}I_y} + C_3}{\sigma_{I_{SR}}\sigma_{I_y} + C_3}.$$

Here,  $\mu$  represents the mean,  $\sigma$  the variance, and  $\sigma_{I_{SR}I_y}$  the covariance of the images. Constants  $C_1$ ,  $C_2$ , and  $C_3$  prevent division by zero. SSIM values range from 0 to 1, with higher values indicating greater similarity.

Subjective evaluation metrics, like the **Mean Opinion Score (MOS)**, rely on human observers to rate image quality. While reflective of human perception, MOS is time-intensive, costly, and prone to biases, making it impractical for large-scale evaluations.

To overcome limitations of traditional metrics like PSNR and SSIM, alternative objective metrics have been introduced:

- **Natural Image Quality Evaluator (NIQE)**: A blind metric that predicts image quality using statistical features, independent of reference images or human input.
- **Learned Perceptual Image Patch Similarity (LPIPS)**: Focuses on comparing deep features between reconstructed and HR images, calculating L2 distances in feature space to better align with human perception.

The focus of this research is not solely on generating visually appealing images but also on producing images that are functional and suitable for downstream tasks, such as object detection. Metrics like PSNR and SSIM, while effective at evaluating perceptual quality, may not reflect the utility of images in object detection pipelines. To address this limitation, [Shermeyer and Van Etten](#) proposed the use of object detection metrics to evaluate the applicability of reconstructed images. Specifically, ground truth bounding boxes were compared to predicted bounding boxes for each test image. A true positive is defined as a prediction with an Intersection over Union (IoU) exceeding a predefined threshold. This threshold can be adjusted based on the size of the target objects, with lower thresholds applied for smaller objects to improve detection accuracy.

While significant advancements have been made in super-resolution techniques, key gaps remain, particularly in their application to aerial imagery. Most methods focus on natural or satellite images, with limited exploration of their effectiveness on aerial datasets containing diverse features such as urban environments and buildings. Additionally, there is a lack of research on how models trained on data captured during one time period perform when **applied to data captured under different temporal or seasonal conditions, where environmental changes may impact performance**. Evaluating how approaches handle critical details like building edges or high-frequency textures is essential, as these features are often crucial for downstream tasks like object detection. Furthermore, there is limited research on the influence of artifacts introduced by generative methods, particularly GANs, on the performance of object detection pipelines. This study seeks to bridge these gaps by **investigating metrics that effectively evaluate super-resolution performance in aerial imagery and examining the applicability of generative methods to this domain while considering their potential impact on subsequent object detection tasks**.

## 2.4 Research Questions

The primary aim of this research is to identify the most effective super-resolution techniques for enhancing 25 cm aerial images to 8 cm resolution, ensuring their applicability for object detection tasks in deep learning pipelines. To achieve this, the study is guided by the following main research question and sub-questions:

### Main Question

*Which super-resolution techniques are most effective in enhancing 25 cm aerial images to 8 cm resolution, ensuring their applicability for object detection tasks in deep learning pipelines?*

### Sub - Questions

- How accurately can super-resolution techniques improve spatial resolution from 25 cm to 8 cm for aerial images?
- What are the implications of domain adaptation on super-resolution performance when comparing synthetic and real aerial images under varying seasonal conditions?
- Which deep learning architectures and methods yield the best results for enhancing aerial images in terms of both perceptual quality and functional utility for object detection tasks?
- What metrics should be used to evaluate the suitability of super-resolved images for object detection in geospatial applications?

## 2.5 Scope

The focus of this research will be to evaluate the effectiveness of super-resolution techniques in enhancing aerial images from 25 cm to 8 cm resolution for object detection tasks. The study will concentrate on the applicability of super-resolution for geospatial analysis, with an emphasis on improving the functional utility of reconstructed images rather than purely enhancing perceptual quality. The research will primarily explore deep learning-based super-resolution methods and their integration with object detection pipelines. While both perceptual and functional metrics will be analyzed, the primary objective is to assess how well super-resolved images support object detection tasks, such as identifying specific features in aerial imagery such as solar panels in buildings roofs. Emphasis will be placed on methods that have been applied to remote sensing imagery, ensuring relevance to the domain of aerial data analysis.

The study will not delve deeply into advanced domain adaptation techniques or alternative super-resolution frameworks outside the scope of deep learning. Similarly, seasonal variations will be considered only to the extent they impact model performance for specific use cases. Finally, while multiple loss functions will be evaluated, only those relevant to the chosen methodology will be analyzed in detail.

This research takes advantage of the availability of both LR and HR datasets to evaluate super-resolution techniques. The methodology involves a two-step iterative process: first, downscaling HR images to create synthetic LR datasets for model training and initial super-resolution outputs, and second, applying the saved model weights to real-world LR datasets

to assess their performance. This approach allows for a comprehensive evaluation of super-resolution techniques, focusing on both synthetic and real-world data and will be described more in depth in the following chapter.



### 3 Methodology

The general methodology adopted for this research is presented in Figure 3.1. The workflow begins with preliminary research to explore and evaluate various algorithms suited for the specific use case. Then the data will be collected and pre-processed. The data will be drawn from different categories of urban settings. Following this, the most effective deep learning model will be selected based on initial experiments conducted on a subset of the data to ensure computational efficiency.

Once the optimal model is identified, it will be fully implemented and trained using the entire dataset to achieve comprehensive results. The next phase is about the optimization of the model and will focus on the validation and evaluation of the model's performance as well as fine tuning the parameters. Finally, the results will be obtained and if time is not limited, the super-resolution model will be integrated with the object detection pipeline provided by the company to assess its applicability in real-world tasks.

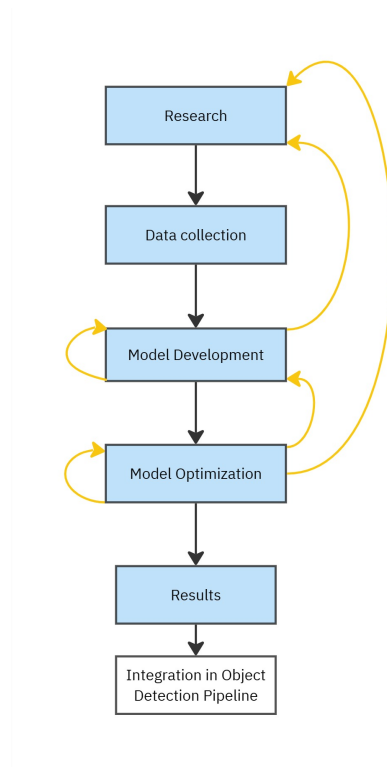


Figure 3.1: Methodology Flow Chart

### 3.1 Experimental Design

Initially, aerial images of the study will be selected from specific regions of the Netherlands. A grid will be applied to the selected areas, and tiles will be generated to create HR-LR pairs for analysis. Optionally, a script will be implemented to ensure consistent tile selection based on unique tile IDs. The tile size will be determined based on the requirements of the approach being followed, and adjustments, such as overlaps between tiles, may be considered if necessary. Additionally, there is a possibility of adopting different approaches tailored to regions with distinct characteristics, ensuring the methodology accommodates the diverse attributes of the study area. Once tiling is complete, files containing both HR and LR data will be generated, completing the pre-processing step for the data. Figure 3.2 illustrates an example of these steps for the Delft region.

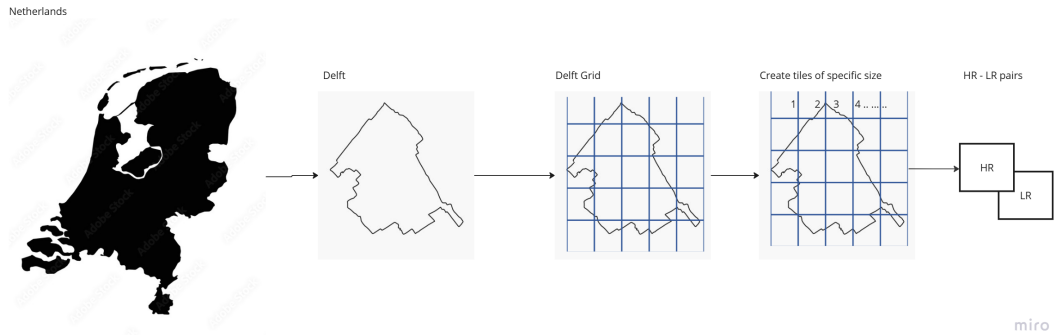


Figure 3.2: Pre-processing steps for Delft (example)

Next, the strategy, also shown in Figure 3.3, involves two iterations:

1. In the first iteration, HR images will be downsampled to create synthetic LR images. The model will then be trained on these pairs to generate super-resolution (SR) outputs.
2. In the second iteration, ground truth LR images will be used alongside the saved model weights from the first iteration to produce new SR images.

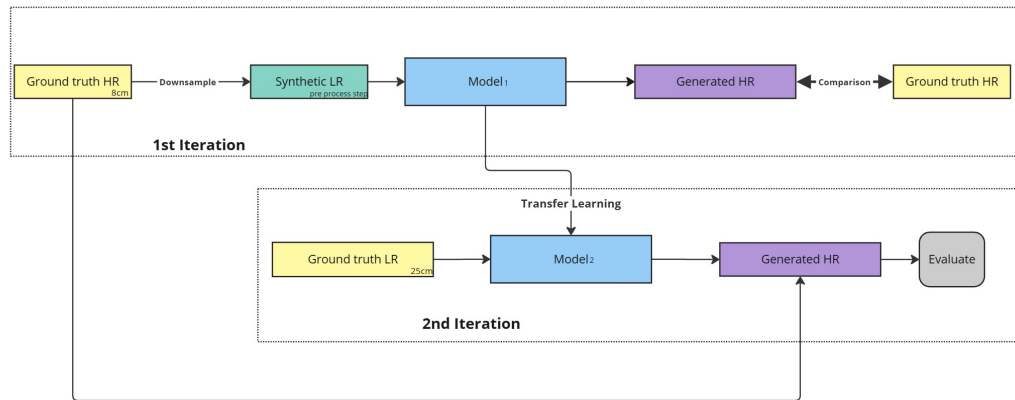


Figure 3.3: Strategy illustration

## 3.2 General Hypotheses

The hypotheses are as follows:

### For Iteration 1:

1. The SR model trained on synthetic LR-HR pairs will achieve high reconstruction quality on synthetic LR validation images due to consistent domain characteristics.

**Testing:** Evaluate model performance on synthetic LR-HR validation pairs. Compare the Generated HR with the Ground Truth HR and with the use of the metrics try to get as close to this.

### For Iteration 2:

1. The SR model trained on synthetic LR-HR pairs will show reduced performance on real LR images due to the domain gap between synthetic and real LR data.

**Testing:** Compare metrics for real LR-HR pairs with synthetic LR-HR pairs.

### Mutual Hypothesis:

- Fine-tuning the model trained on synthetic LR-HR pairs with a small subset of real LR-HR pairs will significantly improve SR performance on real LR images.

### Testing:

- Compare performance on real LR-HR pairs before and after fine-tuning with real LR-HR data.
- See the impact of having LR-HR pairs of different time periods.
- Focus more on the buildings and not vegetation, forests, fields etc.

### 3.3 SR Model Hypotheses

Here's a refined and more polished version of your paragraph:

As reviewed previously, there are numerous deep learning approaches for super-resolution, each utilizing distinct methodologies and learning techniques. For this research, three specific approaches were selected for implementation: transformer-based methods, CNN-based methods, and generative methods. The hypotheses for these methods are as follows:

1. **Transformer-Based Methods:** Transformer-based models are expected to perform particularly well in scenes with rich edges and complex contours, such as urban environments with high-rise buildings or intricate structures. Their ability to model long-range dependencies and recover high-frequency details enables them to effectively reconstruct fine textures and delineate edges, even in challenging settings with high variability.
2. **CNN-Based Methods:** Leveraging their hierarchical feature extraction capabilities, CNNs are hypothesized to produce robust results by capturing both low-level and high-level image features. Additionally, their ability to incorporate multiscale wavelet analysis allows them to extract multiple orientations and frequency representations, making edges scalable across various urban settings. This means that CNNs can adapt to features of small buildings as well as larger structures like skyscrapers, ensuring consistent edge preservation and detail reconstruction across diverse urban environments.
3. **Generative Methods:** With their emphasis on perceptual quality, generative approaches are hypothesized to excel in delivering visually realistic outputs. While these methods may occasionally introduce artifacts, the use of carefully designed loss functions can mitigate this risk, making them potentially suitable for tasks like building edge reconstruction and texture enhancement. Generative methods, when optimized, may be particularly effective for cases like this study, where both structural and perceptual qualities are critical.

### 3.4 Experimental Setup Decisions

#### 3.4.1 Basic Model to be Adapted

An important decision in this research involves selecting the appropriate model architecture for the super-resolution task. The choice of model will depend on the requirements and characteristics of the two iterations in the experimental setup. One possibility is to use a single model for both iterations, leveraging a shared architecture to handle both synthetic and real-world low-resolution data. Alternatively, different models may be employed for each iteration, with the first iteration focusing on synthetic LR-HR data and the second iteration addressing real-world LR-HR pairs.

This decision will be guided by the performance of candidate models during preliminary testing, with factors such as reconstruction accuracy, computational efficiency, and the ability to generalize to different environmental and temporal conditions being taken into account.

#### 3.4.2 Data Selection

To speed up preliminary testing and optimize code, the experiments will initially be conducted on a limited number of tiles rather than the entire NL dataset. If the approach and

results from the preliminary tests are satisfactory, the experiment will be scaled up to include more tiles covering the complete NL region. These tiles will be split appropriately from the entire dataset.

### 3.4.3 Tile Size & Overlap

Commonly used tile dimensions in similar projects are  $256 \times 256$  pixels or  $400 \times 400$  pixels. The final tile size will be determined based on the requirements of the selected super-resolution approach, balancing computational efficiency and output quality. Additionally, a decision will be made on whether the tiling process should include overlapping tiles. Incorporating overlap may help preserve contextual information at the edges of tiles, which can be critical for achieving better super-resolution results, but it may also increase computational overhead. For the context, the entire area of the Netherlands contains:

- Approximately **11,316 tiles of 256x256 pixels** at a resolution of 25 cm.
- Approximately **110,080 tiles of 256x256 pixels** at a resolution of 8 cm.

### 3.4.4 Categorization of Urban Settings

To enhance model generalization and ensure robust performance across various regions, the dataset will be categorized based on distinct urban settings. These categories are designed to capture diverse environmental characteristics and will serve as the classes for the model. Potential categories include areas with high-rise buildings, open fields or agricultural land, lakes and water bodies, low-density suburban regions, and industrial areas or ports. The final selection of categories will depend on the approach adopted and the relevance of these settings to the super-resolution problem. This categorization will guide the model in learning features specific to each type of urban environment, ultimately improving its adaptability and accuracy in diverse scenarios. Ideally we want the model to include samples of different regions of the Netherlands so it can learn from different set ups such as high raise buildings in the area of Rotterdam to houses near farmlands in areas like Limburg.

## 3.5 Preliminary Results

For this stage of the research, in order to provide preliminary results, the three methods described below were applied to evaluate their performance. The results align with the workflow of iteration 1, where the input consisted of HR (8 cm) photos that were down-scaled to create synthetic LR (25cm) photos. These synthetic LR photos were then processed through the models to generate HR (8cm) images.

### 3.5.1 SRGAN

SRGAN was introduced by [Ledig et al. \[2017\]](#) and involved applying GANs to address the super-resolution problem. This framework was the first model capable of generating realistic natural images at a  $4\times$  scale. It uses GANs for super-resolution reconstruction, introduces a perceptual loss function to replace the traditional MSE-based content loss, and proposes a novel image quality evaluation metric. The SRGAN architecture consists of a generative network trained using perceptual loss and a discriminative network. While SRGAN achieves effective reconstruction, it falls short in refining image texture details, leaving some artifacts

in the output. The results of this method across various categories are illustrated in Table 3.1.

### 3.5.2 TransENet

As described in Section 2.3.4, TransENet is a hybrid model combining transformers and CNNs to enhance remote sensing images. Introduced by [Lei et al. \[2022\]](#), this approach captures long-distance dependencies and effectively mines correlations between high- and low-dimensional features. It incorporates a transformer-based multistage enhancement structure composed of multiple encoders and decoders that leverage multilevel information. The results achieved by this method across various categories are shown in Table 3.2.

### 3.5.3 WMCNN

The framework adopted by [Wang et al. \[2018\]](#) involves an image super-resolution method designed specifically for aerial imagery, utilizing wavelet analysis. Multiple CNNs are trained to approximate multiscale representations, which enable efficient image restoration. This framework combines the representational power of CNNs for learning specific features with the multiscale capability of wavelet analysis to capture multiple orientations and frequency representations. The results of this method across various categories are illustrated in Table 3.3.




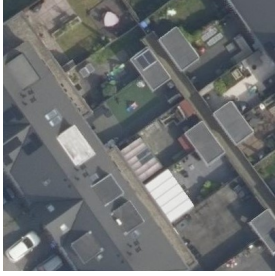
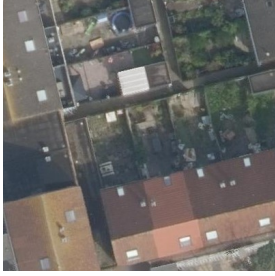




| Image 1   | Image 2  | Image 3   |
|---|--|---|
| <b>LR Input (100x100)</b><br>          | <b>LR Input (100x100)</b><br>          | <b>LR Input (100x100)</b><br>          |
| <b>Real World (400x400)</b><br>        | <b>Real World (400x400)</b><br>        | <b>Real World (400x400)</b><br>        |
| <b>Generated Output (400x400)</b><br> | <b>Generated Output (400x400)</b><br> | <b>Generated Output (400x400)</b><br> |

Table 3.1: Comparison of SRGAN Input, Real World, and Generated Output Images with Horizontally Aligned Labels


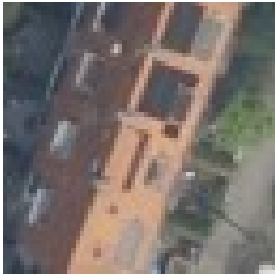


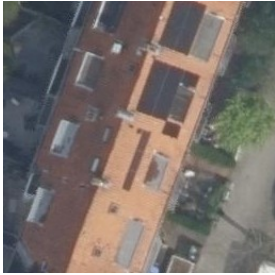
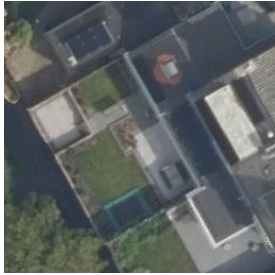
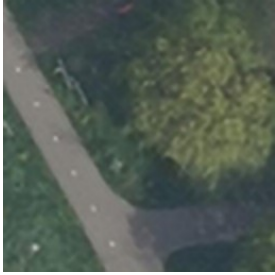
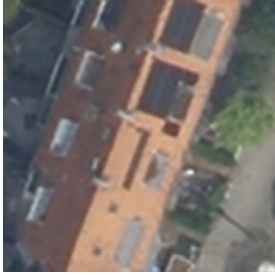

| Image 1  | Image 2   | Image 3  |
|--|---|--|
| LR Input (64x64)<br>            | LR Input (64x64)<br>            | LR Input (64x64)<br>            |
| Real World (256x256)<br>        | Real World (256x256)<br>        | Real World (256x256)<br>        |
| Generated Output (256x256)<br> | Generated Output (256x256)<br> | Generated Output (256x256)<br> |

Table 3.2: Comparison of TransENet Inputs, Real World, and Generated Outputs






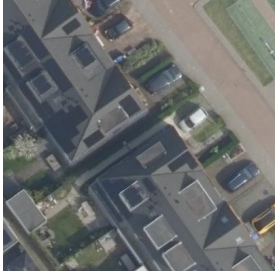
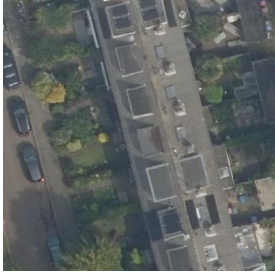
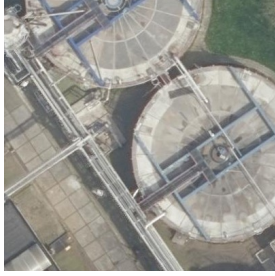


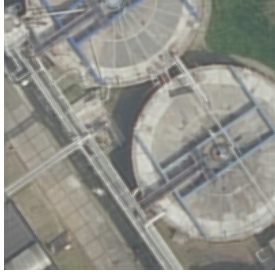
| Image 1  | Image 2   | Image 3  |
|--|---|--|
| LR Input (100x100)<br>          | LR Input (100x100)<br>          | LR Input (100x100)<br>          |
| Real World (400x400)<br>        | Real World (400x400)<br>        | Real World (400x400)<br>        |
| Generated Output (400x400)<br> | Generated Output (400x400)<br> | Generated Output (400x400)<br> |

Table 3.3: Comparison of WMCNN Inputs, Real World, and Generated Outputs

### 3.5.4 Running Time

The running time for the tiling and saving process as well as for the Super Resolution models are given below. These would play an important role to decide later the extend of the data that we will be using for the experiments.

| Resolution                | Speed (Tiles/s) | Speed (Tiles/s) |
|---------------------------|-----------------|-----------------|
| Low Resolution 25 cm (LR) | 2.5             | 3.5             |
| High Resolution 8 cm (HR) | 1.5             | 2.5             |

Table 3.4: Tiling and Saving Process Speeds for LR and HR

| Method    | Train | Test | Validation | Epochs | Scale | Running Time |
|-----------|-------|------|------------|--------|-------|--------------|
| TransENet | 400   | 100  | 60         | 500    | x4    | 40 minutes   |
| WMCNN     | 400   | 100  | 60         | 500    | x4    | 1 hour       |
| SRGAN     | 400   | 40   | 40         | 4000   | x4    | 7 hours      |

Table 3.5: Super-Resolution Process Details

### 3.5.5 Observations

Even though it is not yet possible to determine which model performed the best without fine-tuning specific parameters and loss functions, it can be seen that all models successfully produced satisfactory results. They managed to preserve the edges and avoided introducing artifacts into the generated images. While the results might appear slightly blurry, considering the limited amount of training data and the constraints on the number of epochs—since these were preliminary results—they are efficient. In the next stage, a deeper understanding of each method will be incorporated, alongside parameter optimization and process modifications, to achieve improved outcomes. Additionally, due to time constraints, metrics could not be gathered at this point but will be included and discussed during the P2 presentation.

## 4 Dataset & Tools

### 4.1 Aerial Imagery

The aerial imagery provided to Readar B.V. by [Beeldmateriaal](#) is captured using airplane-mounted cameras to accurately map the Netherlands. These flights produce a variety of products, with this research focusing on high-resolution (HR) and low-resolution (LR) aerial photographs. Both types of imagery are captured annually, ensuring up-to-date geospatial data.

#### 4.1.1 High-Resolution Imagery

High-resolution photographs (HR photos) are captured during the winter, also referred to as the leafless season, before April 23. These images are taken with a 60% longitudinal overlap and a 30% lateral overlap to enable stereoscopic viewing. The resulting images have a resolution of 7.5 cm. After processing, they are stitched together to form a nationwide ortho-photo mosaic with a ground pixel resolution of 8 cm.

#### 4.1.2 Low-Resolution Imagery






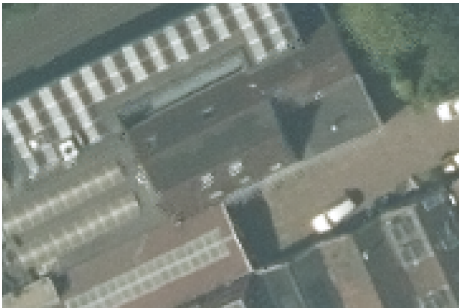
Low-resolution aerial photographs (LR photos) are taken during the summer, when trees are in full leaf. These images provide a nationwide color aerial photograph with a ground pixel resolution of 25 cm. The photographs are captured with an 80% longitudinal overlap and a 20% lateral overlap.

#### 4.1.3 Image Specifications

Both HR and LR photographs are stored in TIFF format and feature a 32-bit RGBI color palette, with 8 bits per color channel. Each individual aerial photograph is accompanied by an XML file containing metadata information. This metadata complies with the current Dutch metadata profile based on ISO 19115. For this research, the images will primarily be utilized in either **RD (Rijksdriehoek)**: EPSG: 28992 or **WGS-84 (World Geodetic System 1984)**: EPSG: 4326 projections.

For the initial phase of this research, a sample of the data is presented using images captured in the Delft region as an example. Figure 4.1 illustrates high-resolution (HR) and low-resolution (LR) aerial images of Delft at different zoom levels, highlighting the differences in detail and resolution.

Table 4.1: Aerial Imagery of Delft in High-Resolution (HR) and Low-Resolution (LR) at Different Zoom Levels

| High-Resolution (HR)  | Low-Resolution (LR)  |
|---|--|
|    |    |
| <i>Delft HR Image</i>   | <i>Delft LR Image</i>  |
|   |   |
| <i>Zoomed HR Image</i>  | <i>Zoomed LR Image</i>   |
|  |  |
| <i>More Zoomed HR Image</i>   | <i>More Zoomed LR Image</i>  |

#### **4.1.4 Software Tools**

The implementation of this thesis will involve several software tools. QGIS will be used for visualizing the data and analyzing the results. For programming, Python and MATLAB will be utilized to develop, edit, and run code, either through the Windows Command Prompt or the Windows Subsystem for Linux (WSL). For accessing the database, DBeaver will be used alongside SQL queries to retrieve data as needed. The thesis document will be written in Overleaf, a LaTeX editor and all figures will be created using Miro.com and drawi.io.

## 5 Planning

The planning and phasing aligned with the academic calendar are illustrated in Table 5.1 and Figure 5.1. Key dates and events are outlined below, based on the academic calendar. The exact dates will be finalized during the year after agreeing with the supervisors.

| Stage               | Timeline                 |
|---------------------|--------------------------|
| <b>Kick-Off</b>     | Week 5: 27 Jan - 02 Feb  |
| <b>Midterm</b>      | Week 16: 14 Apr - 20 Apr |
| <b>Green Light</b>  | Week 21: 19 May - 25 May |
| <b>Finalisation</b> | Week 26: 23 Jun - 29 Jun |

Table 5.1: Thesis Timeline

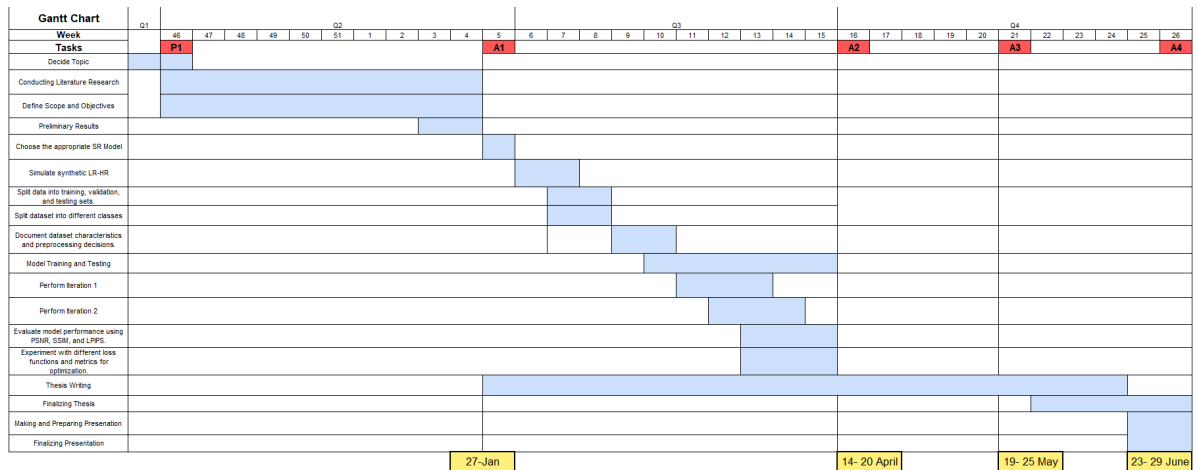


Figure 5.1: Gantt chart of activities

## 6 Meetings-Supervisors-Contact

Meetings with TU Delft supervisors will occur approximately every two weeks to provide academic guidance. Meetings with company supervisors at Readar will be held every 5–6 weeks, focusing on aligning research outcomes with practical objectives. Additionally, frequent updates on thesis progress will be shared to maintain communication.

The supervisors for this thesis project are as follows:

### **Delft University of Technology**

- **Main mentor:** Dr.ir. B.M. Meijers
- **Second mentor:** Dr. A. (Azarakhsh) Rafiee

### **Readar**

- **External supervisor:** Sven Briels

This thesis project is conducted in collaboration with Readar, with the work primarily taking place at their offices located at Utrecht University. The external supervisor from Readar will provide support for use cases and technical matters. At TU Delft, research-related guidance will be provided by both university supervisors. Feedback from both Readar and TU Delft will be received throughout the thesis to ensure steady progression and quality improvement.

Contact details:

Michalis Michalas  
Student number: 6047378  
M.MICHALAS@student.tudelft.nl

# Appendix

Table 6.1: Common datasets for image super-resolution (SR) and remote sensing image tasks.

| Dataset             | Resolution         | Description   |
|---------------------|--------------------|---|
| BSDS300/500         | 435×367<br>432×370 | Animal, scenery, decoration, plant, etc.  |
| DIV2K               | 1972×1437          | People, scenery, animal, decoration, etc.   |
| Set5/Set14/Urban100 | 256×256<br>512×512 | Includes categories such as baby, butterfly, bird, head, woman, baboon, bridge, coastguard, foreman, etc. |
| Urban100            | 984×797            | Construction, architecture, scenery, etc.   |
| AID                 | 600×600            | Airports, deserts, bare land, beach   |
| RSSCN7              | 400×400            | Farmlands, parking lots, residential areas, lakes, etc.   |
| WHU-RS19            | 600×600            | Bridge, forest, pond, port, etc.  |
| UC Merced           | 256×256            | 21 categories: rivers, forests, and agricultural zones, etc.  |
| NWHU-RESISC45       | 256×256            | 45 categories: airports, basketball courts, residential areas, ports, etc.                                |



Table 6.2: Summary of Super-Resolution (SR) Approaches

| Approach   | Description   |
|--|---|
| <b>Early Methods</b>                                     | Techniques such as nonuniform interpolation, frequency domain analysis, deterministic and stochastic regularization, and projection onto convex sets. Commonly used in the early stages of SR development.                                    |
| <b>Pansharpening and Multispectral Fusion</b>            | Classical methods for remote sensing tasks, using techniques like component substitution (CS), multi-resolution analysis (MRA), variational optimization (VO), spectral unmixing, and Bayesian models. These methods require multi-band data. |
| <b>Single-Image SR (SISR)</b>                            | Enhances the resolution of a single low-resolution (LR) image. Simple to implement and widely applicable.   |
| <b>Multi-Image SR (MISR)</b>                             | Uses multiple LR images of the same scene to reconstruct a higher-resolution (HR) output by aligning and fusing the images.   |
| <b>LR-HR Image Fusion</b>                                | Combines low-resolution and high-resolution images to produce enhanced outputs by leveraging HR details.  |
| <b>SRCNN (2015)</b>                                      | A CNN-based SR method with three layers designed for feature extraction, nonlinear mapping, and reconstruction. Requires pre-upsampling using bicubic interpolation.  |
| <b>VDSR (2016)</b>                                       | Introduced residual learning to predict the residuals between the bicubically upsampled LR image and the HR output, improving accuracy and training efficiency.   |
| <b>ResNet (2016)</b>                                     | Uses a residual learning framework to simplify the training of very deep networks by focusing on residuals relative to layer inputs.  |
| <b>Recursive Networks (e.g., DRCN)</b>                   | Based on parameter sharing, these networks repeatedly use the same convolutional layer, reducing the number of trainable parameters and computational complexity.   |
| <b>Post-Upsampling Frameworks</b>                        | Replaces traditional upsampling methods with learnable upsampling layers, working entirely in low-dimensional space until the final reconstruction stage. Efficient but struggles with intermediate feature enhancement.                      |
| <b>TransENet (Transformer-Based Enhancement Network)</b> | Utilizes both high-dimensional and low-dimensional features after upsampling layers to enhance representation. Incorporates transformers to capture long-range dependencies and correlate features from different stages.                     |
| <b>GANs (Generative Adversarial Networks)</b>            | Generates visually realistic HR images by using a generator-discriminator framework. Focuses on perceptual quality but faces challenges like hallucination artifacts and training instability.  |

# Bibliography

- Anwar, S., Khan, S., and Barnes, N. (2020). A Deep Journey into Super-resolution: A Survey. *ACM Comput. Surv.*, 53(3):60:1–60:34.
- Beeldmateriaal (2023). Beeldmateriaal: Aerial and satellite imagery services.
- Dong, C., Loy, C. C., He, K., and Tang, X. (2015). Image Super-Resolution Using Deep Convolutional Networks.
- Förstner, W. and Wrobel, B. P. (2016). *Photogrammetric Computer Vision*, volume 11 of *Geometry and Computing*. Springer International Publishing, Cham.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. ISSN: 1063-6919.
- Kawulok, M., Kawulok, J., Smolka, B., and Celebi, M. E., editors (2024). *Super-Resolution for Remote Sensing*. Unsupervised and Semi-Supervised Learning. Springer Nature Switzerland, Cham.
- Kim, J., Lee, J. K., and Lee, K. M. (2016a). Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654. ISSN: 1063-6919.
- Kim, J., Lee, J. K., and Lee, K. M. (2016b). Deeply-Recursive Convolutional Network for Image Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 1063-6919.
- Kresse, W. and Danko, D. M., editors (2012). *Springer Handbook of Geographic Information*. Springer Handbooks. Springer, Berlin, Heidelberg.
- Lai, W.-S., Huang, J.-B., Ahuja, N., and Yang, M.-H. (2017). Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. (2017). Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. arXiv:1609.04802 [cs].
- Lei, S., Shi, Z., and Mo, W. (2022). Transformer-Based Multistage Enhancement for Remote Sensing Image Super-Resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60.
- Lepcha, D. C., Goyal, B., Dogra, A., and Goyal, V. (2023). Image super-resolution: A comprehensive review, recent trends, challenges and applications. *Information Fusion*, 91.
- Qi, Y., Yang, Z., Sun, W., Lou, M., Lian, J., Zhao, W., Deng, X., and Ma, Y. (2022). A Comprehensive Overview of Image Enhancement Techniques. *Archives of Computational Methods in Engineering*, 29:583–607.

- Shermeyer, J. and Van Etten, A. (2019). The effects of super-resolution on object detection performance in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- Singh, G. and Mittal, A. (2014). Various Image Enhancement Techniques - A Critical Review. *International Journal of Innovation and Scientific Research*, 10(2):267–274.
- Su, H., Li, Y., Xu, Y., Fu, X., and Liu, S. (2024). A Review of Deep-Learning-Based Super-Resolution: From Methods to Applications.
- Vishnukumar, S., Nair, M. S., and Wilsby, M. (2014). Edge preserving single image super-resolution with improved visual quality. *Signal Processing*, 105:283–297.
- Wang, J., Chen, H., Zhu, Y., Li, X., and Gong, M. (2022a). Enhanced super-resolution for remote sensing images based on dual-branch convolutional neural networks. *Remote Sensing*, 14(21):5423.
- Wang, T., Sun, W., Qi, H., and Ren, P. (2018). Aerial Image Super Resolution via Wavelet Multiscale Convolutional Neural Networks. *IEEE Geoscience and Remote Sensing Letters*, 15(5):769–773.
- Wang, X., Yi, J., Guo, J., Song, Y., Lyu, J., Xu, J., Yan, W., Zhao, J., Cai, Q., and Min, H. (2022b). A Review of Image Super-Resolution Approaches Based on Deep Learning and Applications in Remote Sensing. *Remote Sensing*, 14.
- Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.
- Yang, C.-Y., Ma, C., and Yang, M.-H. (2014). Single-Image Super-Resolution: A Benchmark. In *Computer Vision – ECCV 2014*, pages 372–386, Cham. Springer International Publishing.