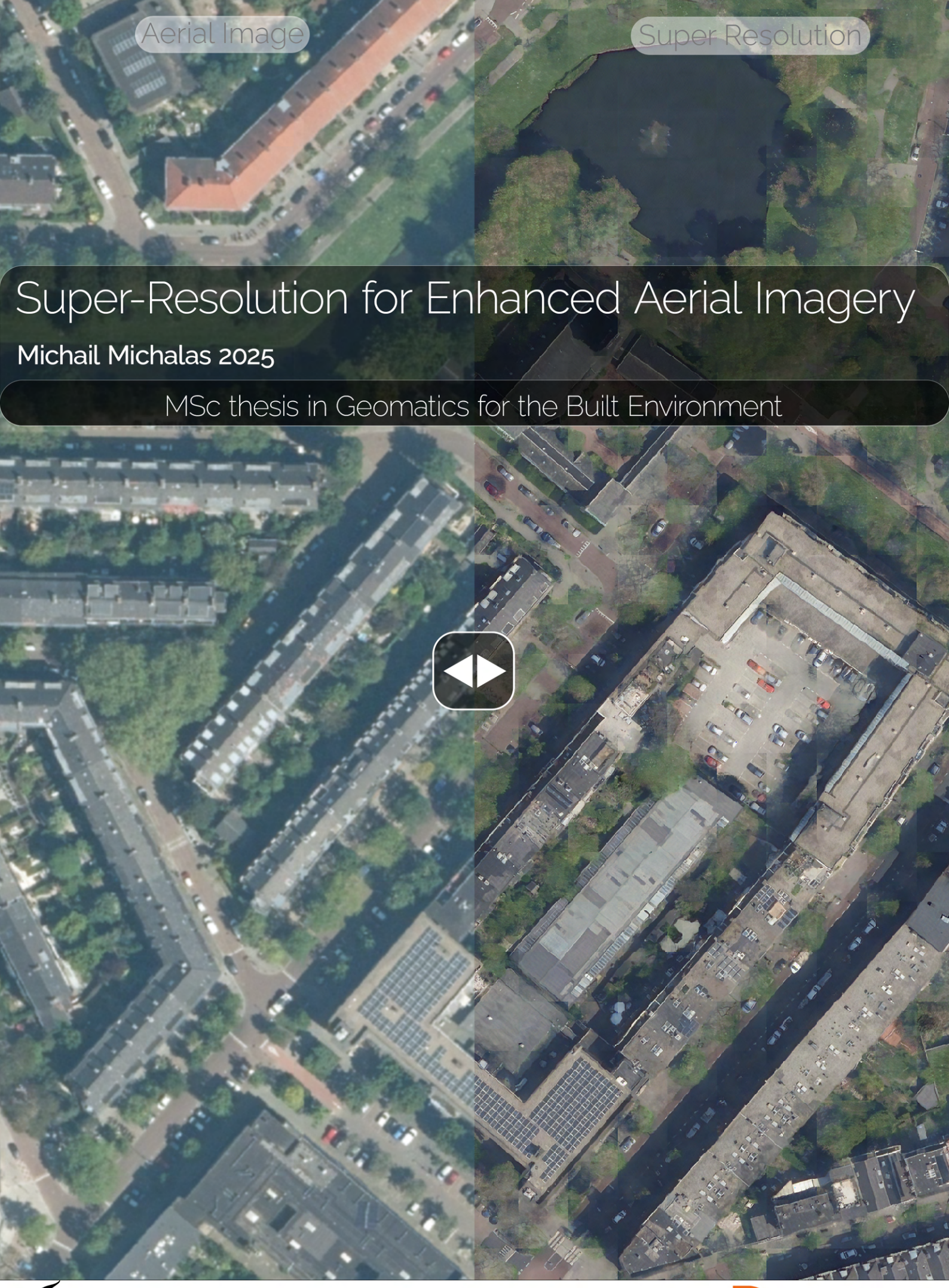

Super-Resolution for Enhanced Aerial Imagery

Michail Michalas 2025

MSc thesis in Geomatics for the Built Environment

TU Delft

READAR
real estate radar

Cover illustrations:

Left: True ortho imagery of Rotterdam in 25cm resolution, generated by READAR.

Right: Super-resolution generated imagery of Rotterdam in 8cm resolution.

MSc thesis Geomatics for the Built Environment

Super-Resolution for Enhanced Aerial Imagery

Michalis Michalas
6047378

June 2025

A thesis submitted to the Delft University of Technology in partial fulfillment of the requirements for the degree of Master of Science in Geomatics for the Built Environment

Michail Michalas: *Super-Resolution for Enhanced Aerial Imagery*

An electronic version of this thesis is available at
http://repository.tudelft.nl/

The work in this thesis was carried out with:

Geo-Database Management Center
Faculty of the Built Environment & Architecture
Delft University of Technology
https://www.gdmc.nl

READAR
Princetonlaan 6
3584 CB, Utrecht
https://readar.com

Supervisors:	Dr.ir. Martijn Meijers
	Dr. Azarakhsh Rafiee
Co-reader:	Nail Ibrahimli
External Supervisors:	Sven Briels
	Camilo Caceres

Abstract

High-resolution aerial imagery plays a critical role in urban planning, energy mapping, and land-use classification. However, many datasets remain limited to lower resolutions due to acquisition costs or legacy data sources. Super-resolution (SR) techniques offer a means to enhance 25 cm aerial imagery to 8 cm, making it more suitable for object-level analysis. This thesis investigates the capability of a modified SRGAN architecture to enhance the visual and structural fidelity of aerial images, thereby improving the representation of urban features such as rooftops, dormers, and solar panels. The architecture incorporates an EdgeMaskBlock to improve edge awareness and preserve sharp contours in reconstructed imagery.

To address the challenges of spatial complexity and temporal misalignment, a two-phase training strategy is implemented. First, the model is trained on synthetically downsampled HR-LR pairs to establish a robust initialization. This is followed by fine-tuning on real-world 25 cm inputs misaligned with their 8 cm HR counterparts, enabling the model to generalize under realistic and variable acquisition conditions.

Evaluation is conducted across both training iterations using standard image quality metrics (PSNR, SSIM, LPIPS), along with downstream segmentation benchmarks. For Iteration 2, the generalization capability of the model is assessed across new cities and seasonal conditions. Two segmentation pipelines are used: the Segment Anything Model (SAM) and the operational semantic segmentation system developed by Readar B.V., which detects buildings, dormers, and PV panels using both RGB and DSM data. Metrics such as precision, recall, and F1-score demonstrate that super-resolved outputs significantly outperform bicubic upsampling, particularly for fine-scale rooftop objects.

The results show that the proposed SRGAN model improves perceptual quality while enabling effective domain transfer across seasons. These enhancements contribute to more reliable segmentation outputs, reinforcing the potential of GAN-based super-resolution as a practical tool in geospatial workflows that require fine-grained object recognition.

Acknowledgements

Within this section I would like to take the opportunity to thank a few people that had an impact during the execution of this research.

First, I would also like to thank my academic supervisors at TU Delft, *Martijn Meijers* and *Azarakhsh Rafiee*, for their valuable guidance, constructive feedback, and continuous support during the research. Their insights helped me navigate both the technical and methodological aspects of the thesis with more clarity.

A special thanks goes to my company advisors at READAR, *Sven Briels* and *Camilo Caceres*. Their technical supervision, willingness to brainstorm, and the many times they helped me get unstuck when things got complicated were essential to this work. Their input was particularly valuable when it came to framing the machine learning strategy and structuring the workflow, offering clarity at moments when the direction of the project was uncertain. This work would not have been possible without their collective input.

I would also like to thank my *parents* and my *sister* for their unconditional support throughout this project — their patience and encouragement helped me stay grounded through every phase of the thesis.

Finally, I would like to thank my *close friends* for always being there to help me unwind and recharge during busy times. Their presence made the intense periods of work more manageable and reminded me to keep perspective throughout the journey.

Contents

List of Figures

List of Tables

Acronyms

1. Introduction

1.1. Motivation & Problem Statement

High-resolution aerial imagery is a cornerstone of geospatial analysis, enabling the creation of datasets such as (Digital Surface Model (DSM)), 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, while the high cost and infrequent capture of high-resolution imagery further complicate the ability to conduct detailed and continuous analyses [Wang et al., 2022a]. 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 [Kawulok et al., 2024] . 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 in better reconstructed image. SR techniques have applications beyond geospatial fields, including medical diagnostics, object detection, and forensic analysis [Lepcha et al., 2023]. Most Single Image Super-Resolution (SISR) research has focused on natural images (faces, animals), with far less attention paid to aerial data that contains both semantic and geometric complexity. In particular, the presence of buildings, shadows, vegetation, and seasonal changes introduces unique challenges that are underexplored.

Among existing approaches, Generative Adversarial Networks (GAN) have proven effective in synthesizing visually convincing textures. This makes them suitable for aerial imagery where perceptual quality is often as important as pixel fidelity. In particular, this work builds upon SRGAN, a model proposed by Ledig et al. (2017) and designed to balance high-level realism with low-level accuracy through a perceptual loss strategy. The SRGAN architecture is adapted specifically for aerial imagery, with architectural modifications proposed to better handle domain-specific challenges such as temporal changes and texture regularity in urban structures. An edge-aware refinement mechanism is introduced through a custom Edge Mask Block, designed to guide the network's attention to structural boundaries such as rooftops and building contours.

This research focuses on enhancing the resolution of aerial imagery over the Netherlands, with the goal of improving its suitability for downstream applications such as object detection. In particular, the study targets the detection of rooftop elements such as solar panels, which requires imagery with sufficient spatial detail to resolve fine-grained structures. While high-resolution images are generally expected to support products like DSMs, TrueOrthos, and solar irradiation maps, these outputs are not directly evaluated in this study. Existing imagery is often limited in resolution and affected by distortions arising from sensor characteristics and environmental conditions. Furthermore, temporal variability—such as seasonal changes, vegetation growth, and shadow shifts—introduces misalignment between available high-resolution (HR) and low-resolution (LR) datasets, adding complexity to the super-resolution task.

In the dataset used for this study, HR and LR images are available for the same areas but are captured at different times of the year. While this provides broad coverage, it complicates the SR task, as differences in lighting, and urban or vegetation development must be accounted for. The objective is to develop a method that remains robust to these temporal inconsistencies. This challenge is addressed by first training a super-resolution model on synthetic pairs generated from HR images. The learned weights are then transferred and fine-tuned using real-world LR images aligned with HR references that were not captured simultaneously. This two-phase process supports domain adaptation and allows the model to generalize under real-world noise and variability. By leveraging both HR and LR datasets that are spatially aligned and share the same reference system, this work enables a practical solution for enhancing imagery in real-world geospatial pipelines.

This thesis aligns with the objectives of Readar B.V.,[1] a company specializing in high-quality geospatial datasets. The study employs 8 cm and 25 cm resolution True Ortho imagery of Rotterdam, Delft, and Utrecht for training and validation, while generalization performance is evaluated on additional cities, including The Hague and Zwolle.

To achieve the goals described and keep the research aligned with practical applications, a set of research questions and hypotheses was developed.

1.2. Research Questions & Hypothesis

Main Research Question:

> *To what extent can GAN-based super-resolution enhance 25 cm aerial imagery to 8 cm, ensuring its applicability for object detection tasks?*

To address this main question, the study investigates how accurately super-resolution reconstructs high-resolution content, what limitations arise in realistic conditions, and how the enhanced imagery performs in downstream object detection tasks. This requires evaluating both perceptual and functional quality, accounting for challenges such as temporal misalignment, structural fidelity, and artifact introduction. The following sub-questions guide this investigation.

Sub-Questions:

1. How accurately can a GAN reconstruct 8 cm HR images from 25 cm LR aerial inputs in high-density urban settings, especially at edges and rooftop details?
2. How do seasonal differences between HR and LR images (e.g., winter HR vs. summer LR) affect GAN performance, and can domain adaptation via pre-training on synthetic data mitigate these effects?
3. What are the limitations of GANs in preserving geometric fidelity (e.g., artifacts, hallucinations) for geospatial use cases?
4. What metrics best assess SR image quality for downstream object detection tasks?

[1] https://www.readar.com

Hypotheses:

1. Integrating edge-aware architectural components will improve the model's ability to preserve urban geometry particularly in high-density settings and minimize common GAN-related issues such as hallucinated textures or distorted contours. Edge supervision is expected to help the model prioritize spatial consistency over perceptual realism alone.

2. GAN-based super-resolution can enhance 25 cm aerial imagery to approximate the quality of native 8 cm imagery, particularly in terms of visual clarity and structural definition of rooftop features. This is expected to outperform interpolation methods such as bicubic upsampling, which lack the ability to recover high-frequency details and geometric structure.

3. A two-phase training strategy consisting of pre-training on synthetic HR–LR image pairs and fine-tuning on real LR–HR pairs captured in different seasons can improve robustness to seasonal appearance differences. This approach is hypothesized to support generalization beyond ideal conditions by enabling the model to learn domain-invariant structural features.

4. A multi-metric evaluation approach that combines perceptual (LPIPS), structural (SSIM), pixel-based (PSNR), and task-specific assessments (semantic segmentation accuracy) will provide a comprehensive measure of practical utility. It is hypothesized that perceptual gains alone do not guarantee suitability for downstream tasks, so functional benchmarks are essential. Task-specific assessments include both the Segment Anything Model (SAM) and the semantic segmentation pipeline developed by Readar B.V.

1.3. 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 building 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 which is described more in depth in the following chapter.

1.4. Thesis Outline

The remainder of this thesis is structured as follows:

- Chapter 2 introduces the theoretical background, covering aerial imagery, spatial resolution, image enhancement, and super-resolution techniques.
- Chapter 3 reviews prior work on classical and deep learning-based SR methods, with a focus on challenges in remote sensing and the identified research gap.
- Chapter 4 presents the proposed modifications to the SRGAN architecture, including the edge-aware refinement mechanism tailored to urban structures.
- Chapter 5 describes the dataset, implementation pipeline, and training procedure, including both synthetic pre-training and fine-tuning on real LR data.
- Chapter 6 provides quantitative and qualitative evaluation results, including performance on segmentation tasks and error analysis.
- Chapter 7 concludes with key findings, discusses limitations, and outlines directions for future research.

2. Theoretical Background

This chapter provides the theoretical background relevant to the research. It begins by introducing fundamental concepts in photogrammetry and remote sensing to establish the geospatial context in which this work is situated. Following that, the term spatial resolution is addressed, and traditional image enhancement techniques are discussed, with special attention to bicubic interpolation, which serves as the baseline for comparison. The core principles of super-resolution are then introduced, including the mathematical formulation of the problem and degradation models. This leads into a detailed overview of deep learning-based approaches to super-resolution, with a focus on GAN-based methods. Finally, the architectural structure and loss functions of the baseline model, SRGAN, are explained, and the evaluation metrics used throughout are defined.

2.1. Remote Sensing and Aerial Imagery Products

Remote sensing is the science of acquiring information about objects or areas from a distance, usually using airborne or satellite sensors that capture electromagnetic radiation [Campbell and Wynne, 2011]. Photogrammetry is the science of making accurate measurements from photographs, its considered a branch within remote sensing that focuses on the geometric reconstruction of scenes, having applications in mapping, surveying, and high-precision measurements [Förstner and Wrobel, 2016]. It applies principles of optics and camera geometry to estimate dimensions and spatial positions of features within images. While some single-image approaches exist, photogrammetry most commonly relies on pairs or sets of overlapping images, especially in aerial applications, where stereo analysis is used to derive topographic elevation. In aerial photogrammetry, overlapping photographs are captured from above to reconstruct three-dimensional surfaces. This approach enables the generation of key geospatial products such as digital elevation models (DEMs), 3D models, point clouds, orthomosaics, and True Ortho images.

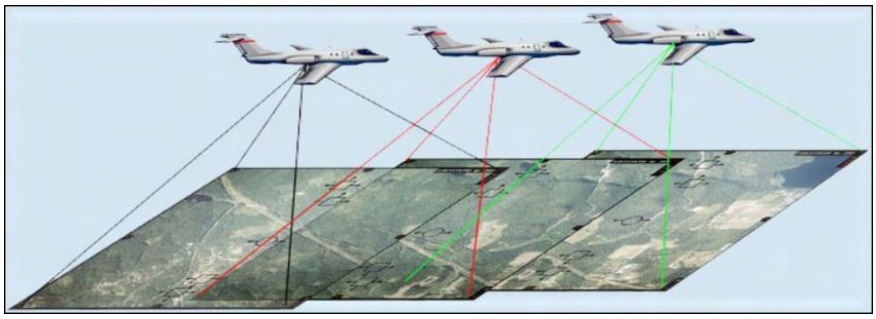

Figure 2.1.: Representation of aerial photogrammetry. Adapted from Jebur et al. [2017]

Ortho images are produced through a process called orthorectification, which corrects distortions caused by terrain relief and sensor perspective. While these images are geo-

metrically accurate and suitable for use as maps, they do not correct for displacements in elevated structures such as buildings and bridges.

True Ortho images address this limitation by incorporating detailed DEMs that provide elevation values for each point above sea level, excluding vegetation and artificial objects. This results in an accurate 3D representation of the earth's surface, ensuring that elevated features are rectified and aligned orthogonally to their bases. This level of geometric accuracy is critical for applications requiring precise spatial alignment.

As shown in Figure 2.2, unlike standard ortho images, True Ortho imagery removes distortions in elevated objects by integrating elevation data directly into the rectification process. The figure highlights how elevated structures such as buildings suffer from geometric misalignment in standard ortho images, whereas True Ortho imagery preserves correct alignment by accounting for elevation. In Figure 2.2a, the building outlines shown in orange are visibly misaligned due to perspective distortions, causing features like roofs and facades to appear shifted or tilted. In contrast, Figure 2.2b shows that these same outlines are properly aligned to the building bases, demonstrating the geometric accuracy provided by True Ortho processing.

(a) Standard ortho image showing leaning building edges due to perspective distortion.

(b) True Ortho image with vertically corrected structures aligned to their bases.

Figure 2.2.: Comparison between standard ortho and True Ortho imagery. True Orthos eliminate distortion in elevated features, improving spatial accuracy for urban analysis. Adapted from Readar B.V. [2024].

True Ortho images are increasingly used in applications where high spatial accuracy is essential, particularly in dense urban environments. Unlike standard orthophotos, which suffer from perspective displacements and occlusions, True Orthos provide a consistent geometric reference across the entire image. In conventional orthophotos, it is often impossible to accurately superimpose vector data for tasks such as change detection and quality control, as parts of the image content may be geometrically incorrect or incomplete. The need for a complete and geometrically reliable image dataset can only be fulfilled by True Ortho imagery [Amhar et al., 1998]. By leveraging detailed elevation data during the rectification process, these images ensure that elevated structures such as roofs and building facades are accurately positioned and vertically aligned. This makes them especially suitable for downstream tasks like object detection, infrastructure monitoring, and urban planning. The production of True Ortho imagery, however, requires high-quality DEMs and additional processing, making it more computationally intensive than standard orthophotos [Kresse and Danko, 2012].

As previously discussed, the geometric accuracy of True Ortho images makes them ideal

for solar panel detection; that's why they are used as inputs in this study. Using standard ortho photos would require additional orthorectification, introducing alignment issues, potential artifacts, and an unnecessary preprocessing step that falls outside the scope of this work.

2.2. Spatial Resolution

Even when aerial images are geometrically rectified, the level of detail they contain is ultimately governed by their **spatial resolution**, which describes the amount of spatial information represented per pixel. In aerial imagery, this is typically expressed as the ground sampling distance (GSD), measured in centimeters or meters per pixel. Spatial resolution plays a critical role in determining the visibility and separability of small-scale features and directly impacts the effectiveness of downstream tasks such as object detection, segmentation, and classification.

The term "resolution" can refer to several properties in remote sensing, including *radiometric*, *temporal* , and *spectral resolution*, this thesis specifically focuses on spatial resolution. **Super-Resolution (SR)** in this context aims to increase spatial resolution—that is, to reconstruct finer detail than what was originally captured by the sensor.

Figure 2.3 shows the same urban scene captured at two different spatial resolutions: 25 cm on the left and 8 cm on the right. Each side includes a zoomed-in view of a central region to highlight how resolution affects the visibility of details. In the low-resolution tile (left), larger pixels cause blurring, making rooftop structures and boundaries hard to see. In the high-resolution tile (right), smaller pixels preserve sharp edges and individual features. This shows why higher resolution is important when clear outlines and small objects need to be recognized, such as rooftops or roads in cities.

Figure 2.3.: Comparison of aerial image resolution. Left: low-resolution tile at 25 cm/pixel with zoomed inset; right: high-resolution tile at 8 cm/pixel.

2.3. Image Enhancement Techniques and Bicubic Interpolation

In the context of aerial imagery, 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 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].

While traditional image enhancement focuses on improving visual quality, interpolation-based upsampling methods like bicubic interpolation aim to increase the spatial resolution of images. Although these methods do not generate new structural details, they serve as a baseline for evaluating more complex super-resolution techniques. These methods estimate unknown pixel values by analyzing known neighboring pixels. Common interpolation methods include nearest-neighbor, bilinear, and bicubic interpolation, with the latter being the most commonly used due to its balance between computational cost and output quality.

Bicubic interpolation for resizing images provides smoother results than bilinear interpolation and introduces fewer artifacts than nearest-neighbor, as it considers a 4×4 neighborhood of known pixels. In image processing, it is commonly implemented using cubic convolution.

Mathematically, the interpolated intensity at (x, y) is computed as:

$$I(x,y) = \sum_{i=-1}^{2} \sum_{j=-1}^{2} w(i,j) \cdot I(x+i, y+j) \tag{2.1}$$

where $I(x+i, y+j)$ are the neighboring pixel values and $w(i,j)$ are weights from the bicubic kernel, typically based on Keys' cubic convolution formula [Keys, 1981].

Upscaling with bicubic interpolation generates new pixels by analyzing local structures such as edges and textures—to produce higher-resolution images with reduced pixelation and improved visual continuity. **Downscaling**, on the other hand, often introduces aliasing artifacts, such as jagged edges or distortions. Bicubic interpolation mitigates these issues by incorporating neighboring pixels to create smoothed and more accurate low-resolution representations.

Figure 2.4 below shows how an image changes when it is downscaled and then upscaled using bicubic interpolation. The first image is the original high-resolution version. The second image is a lower-resolution version created by downscaling it by a factor of four. The third image is the result of upscaling that low-resolution version back to the original size. While bicubic interpolation helps smooth out the image and reduce blocky artifacts, it cannot fully restore the fine details lost during downscaling.

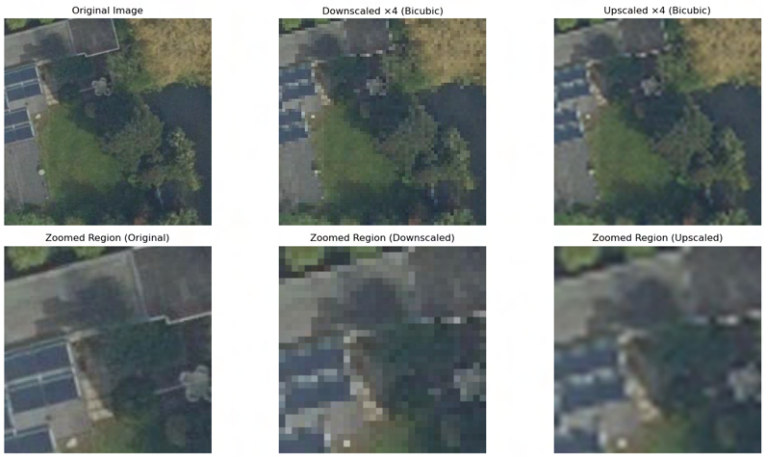

Figure 2.4.: Example of bicubic downscaling and upscaling by scale factor of 4

Bicubic interpolation is commonly used in image enhancement studies both as a baseline method and as a way to synthetically generate low-resolution images from high-resolution inputs. It serves as a conventional, non-learning-based approach to upscale low-resolution aerial imagery and is frequently used in Super Resolution literature for performance comparison. While bicubic interpolation preserves general structure and offers smoother results than simpler techniques, it lacks the ability to recover high-frequency details such as rooftop edges or fine textures—features that are critical in geospatial analysis. More importantly, bicubic downsampling does not reflect the complex degradations present in real-world imagery, which may include noise, compression artifacts, atmospheric distortions, and motion blur [Kawulok et al., 2024]. This mismatch limits the generalization of methods trained solely on bicubic data. To address this, the experimental design of this study includes a second training iteration using real aerial low-resolution imagery. This provides a more realistic degradation model and motivates the introduction of super-resolution techniques, which are discussed in the next section.

2.4. Super-Resolution Fundamentals

Given the limitations of interpolation-based methods like bicubic interpolation in capturing high-frequency detail and modeling real-world degradations, more advanced learning-based approaches such as super-resolution have emerged to overcome these challenges. **Super-resolution** is a process that aims to reconstruct a high-resolution (HR) image from its low-resolution (LR) counterpart. The framework of a single-image super-resolution (SISR) system typically consists of two key components: a nonlinear mapping module and an upsampling module. The nonlinear mapping module is responsible for learning the transformation from LR to HR images, guided by a loss function during training. The upsampling module performs the actual enlargement of the image to the desired resolution [Yu et al., 2024].

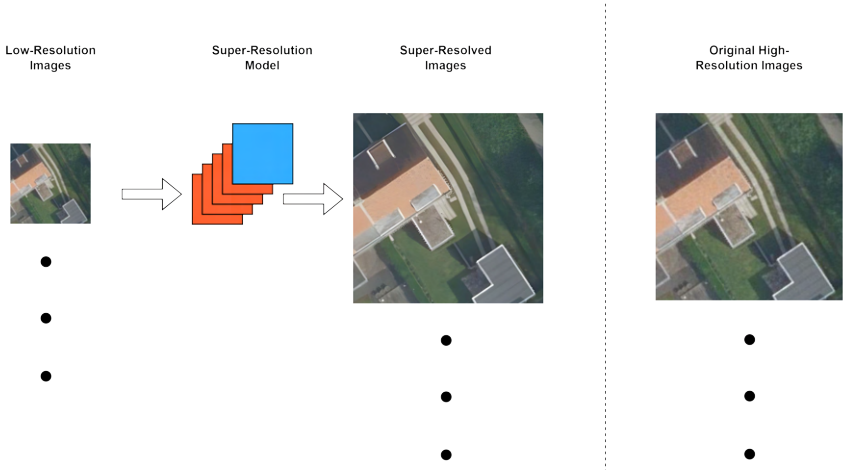

Figure 2.5.: Overview of the super-resolution task: starting with low-resolution images, a super-resolution network is designed to enhance their quality, producing super-resolved versions of the input images.

2.4.1. Concept

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 may involve processes such as blurring, downsampling, or the introduction of noise. The severity of degradation can be expressed by a factor γ, as illustrated in Figure 2.6.

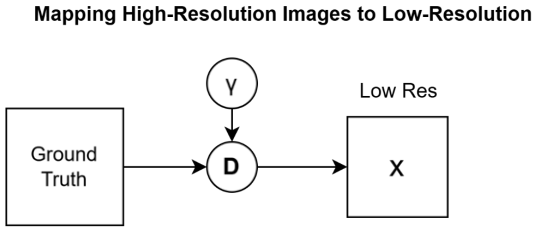

Figure 2.6.: The forward process of degradation: A ground truth high-resolution image undergoes transformation via a degradation function D, yielding a low-resolution version.

The reverse process, *super-resolution*, attempts to reconstruct a high-resolution approximation from the low-resolution input using a model F. This model learns to estimate the lost details and upscale the image spatially (e.g., from 64×64 to 256×256). Figure 2.7 illustrates this upsampling pathway.

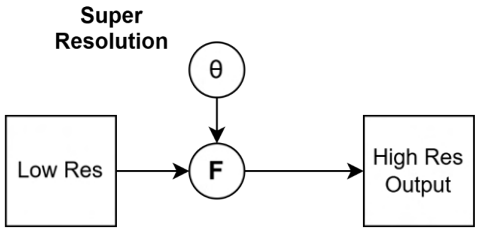

Figure 2.7.: The super-resolution task: A model F takes a low-resolution image and produces a higher-resolution version, ideally recovering details lost during degradation.

However, this inverse task is fundamentally *ill-posed* — because multiple high-resolution images could correspond to the same low-resolution input. As Figure 2.8 shows, there are infinite plausible HR reconstructions that differ in texture, edges, or object boundaries, all mapping to the same LR image.

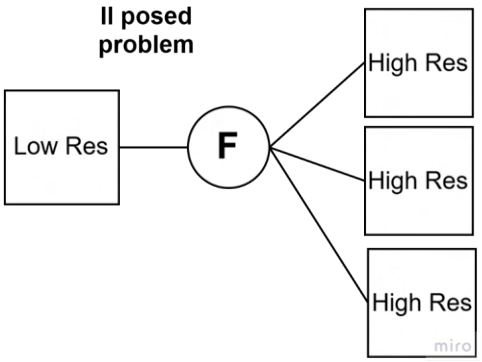

Figure 2.8.: The ill-posed nature of super-resolution: multiple plausible high-resolution outputs can exist for a single low-resolution input.

The key challenge is to train a model that reconstructs HR outputs that are not only visually plausible but also functionally useful — especially for downstream tasks like object detection or segmentation. An overview of the overall super-resolution process is provided in Figure 2.5.

2.4.2. Mathematical Formulation

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 [Kawulok et al., 2024]. To better understand this challenge, this subsection presents 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]. Denoting a Low -Resolution (LR) image as y and the corresponding high-resolution (HR) image as x, then the degradation process is given as:

$$y = \Phi(x; \theta_\eta), \tag{2.2}$$

where Φ is the degradation function, and θ_η 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 θ_η. SR aims to reverse this degradation and approximate the ground-truth image x by estimating image $\hat{x}$ as:

$$\hat{x} = \Phi^{-1}(y, \theta_s), \tag{2.3}$$

where θ_s are the parameters for the function Φ^{-1}. The degradation process is unknown and can be quite complex. To address this complexity, many studies adopt a more detailed degradation model instead of relying solely on Equation 2.2. This refined model is given as:

$$y = (x \otimes k) \downarrow_s + n, \tag{2.4}$$

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 σ (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,

$$J(\hat{x}, \theta_s, k) = \underbrace{\|x \otimes k - y\|}_{\text{data fidelity term}} + \alpha \underbrace{\Psi(x, \theta_s)}_{\text{regularizer}}, \tag{2.5}$$

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

In practice, the super-resolution process involves two distinct yet interconnected phases: a *training phase* and an *inference phase*. During training, the model learns a mapping function f_θ that approximates the inverse degradation process by minimizing a defined loss over known (I_{LR}, I_{HR}) pairs. This learning phase adjusts the network parameters θ to enable accurate reconstruction. Once trained, the inference phase uses the learned parameters to perform *reconstruction*—that is, to generate high-resolution outputs from unseen low-resolution inputs without further weight updates. While these two phases are conceptually separate, most SR literature presents them together since the reconstruction capability is entirely determined by what was learned during training.

2.4.3. Taxonomy of Super Resolution Approaches

Super-resolution methods can be categorized based on how they incorporate 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 prop-

erties. 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.

2.4.4. Blind vs Non-Blind Super-Resolution

While much of the previous discussion assumes a controlled setup where the degradation function is predefined, this is often not the case in real-world applications. The next section desrcibes the difference between scenarios where the degradation model is known versus unknown. When the degradation function is known and explicitly defined—such as bicubic downsampling—the task is referred to as *non-blind super-resolution*. In this case, the model is trained on paired LR-HR images where the LR input has been synthetically generated. This controlled setup allows the model to learn an accurate mapping for a specific and consistent degradation process. In contrast, *blind super-resolution* deals with scenarios where the degradation process is unknown or varies across samples. The LR images in such cases may have undergone multiple forms of degradation, including sensor noise, motion blur, compression artifacts, or arbitrary downsampling kernels. The model must not only reconstruct the HR image but also implicitly learn or infer the degradation process itself during training.

Blind SR is particularly relevant for real-world applications, such as satellite, medical, or remote sensing imagery, where the conditions under which LR images are captured are not controlled or standardized. It presents a more challenging but realistic problem setting compared to non-blind SR. Some approaches to blind SR include kernel estimation to model the blur kernel, adversarial training to encourage natural outputs without explicit degradation supervision, and self-supervised learning techniques that rely on unpaired real-world data. For instance, Zhang et al. [2021] introduced a practical degradation framework that simulates real-world conditions by randomly applying blur, downsampling, and noise in shuffled sequences during training. This distinction is important for framing the approach taken in this research as will be described in Chapter 5.

2.5. Deep Learning in Super Resolution

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_θ, parameterized by weights θ, as:

$$I_{HR} = f_\theta(I_{LR}), \tag{2.6}$$

where f_θ 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}:

$$\mathcal{L} = \|\hat{I}_{HR} - I_{HR}\|^2, \tag{2.7}$$

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

2.5.1. Generative Adversarial Networks

Generative Adversarial Network (GAN) is a deep learning model and one of the most promising methods for unsupervised learning on complex distributions in recent years [Qi et al., 2022]. GANs introduce an adversarial framework for super-resolution, consisting of a *generator* (G) that creates high-resolution images and a *discriminator* (D) that evaluates their quality. It is a game-theoretic approach where the two component of the model fight against each other with the first trying to fool the later. The generator creates SR images that a discriminator cannot distinguish as a real HR image or an artificially super-resolved output [Anwar et al., 2020].

This method is effective at generating perceptually realistic and visually pleasing results, addressing the over-smoothing issues often seen in Convolutional Neural Network (CNN)-based methods. Such models 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 which may not be suitable for certain use cases [Ledig et al., 2017].

A more detailed illustration of the process is shown in Figure 2.9. In this framework, a generator G takes in random noise z sampled from a distribution similar to that of the real data and produces a synthetic image $G(z)$. The discriminator D then takes as input either a real image x or a generated image $G(z)$ and outputs a probability $D(x)$ indicating whether the input is real (closer to 1) or fake (closer to 0).

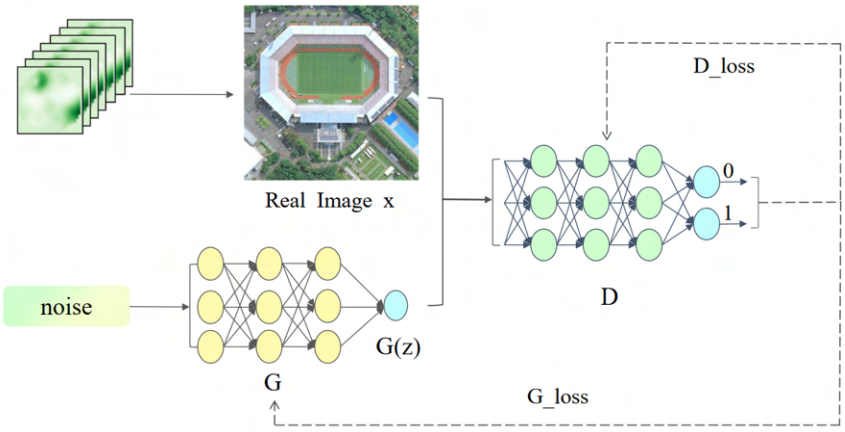

Figure 2.9.: Interaction of Generator and Disciminator adapted from [Anwar et al., 2020]

.

The idea behind GANs can be also explained using the following three equations.

The generator tries to create fake images that look real. Its goal is to fool the discriminator, with the objective function being the following:

$$\min_G \left(D(x) - D(G(z))\right), \tag{2.8}$$

where x is a real image, and $G(z)$ is a generated one based on random input z.

At the same time, the discriminator is trying to do the opposite: tell real images apart from fake ones. So its objective is:

$$\max_D \left(D(x) - D(G(z))\right). \tag{2.9}$$

Putting it all together, the full training process becomes a game between the generator and discriminator:

$$\min_G \max_D \left(D(x) - D(G(z))\right). \tag{2.10}$$

The closer the distribution of generated images $D(G(z))$ gets to the real one $D(x)$, the more realistic the results will be [Wang et al., 2022b].

2.5.2. SRGAN Architecture and Losses

Building upon the general GAN framework and aiming to overcome the limitations of traditional super-resolution methods, SRGAN was introduced by Ledig et al. [2017]. SRGAN combines adversarial training with perceptual loss to produce photo-realistic high-resolution images. It incorporates dedicated architectural designs for both the generator and discriminator, alongside a loss formulation that balances pixel-wise accuracy with perceptual fidelity. The adversarial objective function guides the generator to produce outputs that reside near the manifold of natural images [Anwar et al., 2020]

Method

This section describes the architectural components of SRGAN and the loss functions used to guide its training process. The goal of SRGAN is to reconstruct a high-resolution (HR) image $\mathbf{I}SR$ from a corresponding low-resolution (LR) input $\mathbf{I}LR$, with the model being capable of inferring photo-realistic natural images for 4× upscaling factors.

During training phase, the HR image $\mathbf{I}_{HR}$ is known, and $\mathbf{I}_{LR}$ is synthetically generated by applying Gaussian blur to $\mathbf{I}_{HR}$, followed by downsampling with a fixed scale factor r. The resulting tensor shapes are $W \times H \times C$ for $\mathbf{I}_{LR}$ and $rW \times rH \times C$ for both $\mathbf{I}_{HR}$ and $\mathbf{I}_{SR}$, where C denotes the number of channels.

Then a generative function G is defined and implemented as a feedforward convolutional neural network (CNN) G_{θ_G} with learnable parameters $\theta_G = \{W_{1:L}, b_{1:L}\}$ representing the weights and biases of an L-layer network. The objective is to optimize these parameters such that the generated super-resolved image $G_{\theta_G}(\mathbf{I}_{LR})$ closely approximates the ground truth HR image $\mathbf{I}_{HR}$.

Training involves minimizing a super-resolution-specific loss function $\mathcal{L}_{SR}$ over a dataset of N image pairs. The optimization objective is mathematically defined as:

$$\hat{\theta}_G = \arg\min_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{SR}\left(G_{\theta_G}(\mathbf{I}_{LR}^{(n)}), \mathbf{I}_{HR}^{(n)}\right) \tag{2.11}$$

With this way, the total perceptual loss $\mathcal{L}_{SR}$ is constructed as a weighted combination of multiple components, each designed to enforce specific characteristics in the reconstructed image, such as pixel accuracy, structural consistency, and perceptual fidelity. These components are described in more detail in the next subsection.

Adversarial network architecture

The discriminator network D_{θ_D} of SRGAN is based on Goodfellow et al. [2014] and optimized in different way together with G_{θ_G} in order to solve the adversarial min-max optimization problem:

$$\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{\mathbf{I}_{HR} \sim p_{\text{train}}(\mathbf{I}_{HR})} \left[\log D_{\theta_D}(\mathbf{I}_{HR})\right] + \mathbb{E}_{\mathbf{I}_{LR} \sim p_G(\mathbf{I}_{LR})} \left[\log(1 - D_{\theta_D}(G_{\theta_G}(\mathbf{I}_{LR})))\right] \tag{2.12}$$

This approach helps the generator create solutions that are very similar to the real images and therefore difficult for the discriminator to distinguish. Compared to other SR approaches, this architecture encourages perceptual better solutions residing in the subspace, the manifold, of natural images.

More in detail, the generator and discriminator networks of SRGAN can be seen in Figure 2.10. The generator consists of B residual blocks with block layout proposed by Gross and Wilber [2016], and two convolutional layers with 3×3 kernels and 64 feature maps responsible for extracting low-level features and preparing the feature space for residual learning.

Then, batch normalization layers [Ioffe and Szegedy, 2015] apply a transformation that maintains the mean output close to 0 and the output standard deviation close to 1, followed by a Parametric ReLU [He et al., 2015] as the activation function, described mathematically as:

$$f(y_i) = \begin{cases} y_i, & \text{if } y_i > 0 \\ \alpha_i y_i, & \text{if } y_i \leq 0 \end{cases}$$

This modification of ReLU addresses the problem of returning zero for any negative input value, also known as the vanishing gradient problem. The image resolution is then increased using two trained sub-pixel convolution layers proposed by Shi et al. [2016]

The architecture base for the discriminator was adapted from Radford et al. [2016] and use LeakyReLU activation avoiding max-pooling. With the discriminator trained to solve Equation 2.12 and containing 8 convolutional layers with an increasing number of 3×3 filter kernels, increasing by a factor of 2 from 64 to 512 kernels. The image resolution is reduced each time by strided convolutions doubling the number of features. The probability is then obtained by the resulting 512 feature maps followed by two dense layers and a final sigmoid activation.

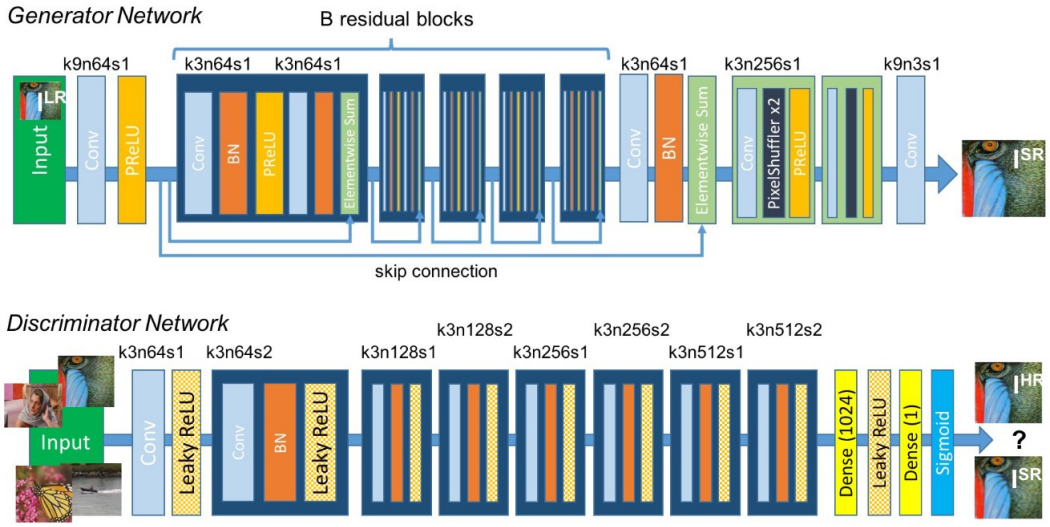

Figure 2.10.: Architecture of Generator and Discriminator Network with corresponding kernel size (k), number of feature maps (n) and stride (s) indicated for each convolutional layer. Image taken from Ledig et al. [2017].

Once the SR model generates the reconstructed images, loss functions are used to measure their difference from the ground truth. These guide the training process and evaluate how well the model learns over time.

Pixel Loss calculates the difference between predicted and ground truth images at the pixel level. Examples include Mean Squared Error (MSE), Mean Absolute Error (MAE), and Charbonnier Loss. MSE (L2) is smooth and easy to optimize but tends to produce overly smoothed results, especially in textured or high-frequency areas. MAE (L1) is less sensitive to outliers but has non-differentiable points. Charbonnier Loss offers a differentiable alternative to L1 with added stability [Lai et al., 2017].

However, all pixel-based losses focus on numerical similarity rather than perceptual quality. They often lead to blurred outputs that lack sharp edges or realistic textures.

Perceptual Loss addresses this by comparing high-level features extracted from pretrained networks rather than raw pixels. In SRGAN [Ledig et al., 2017], it is formulated as a

weighted sum of content loss and adversarial loss:

$$L^{SR} = L_{\text{content}} + 10^{-3} L_{\text{Gen}}(I_{SR}) \tag{2.13}$$

The L_{content} evaluates perceptual similarity using features from a VGG network, and L_{Gen} encourages outputs that appear realistic through adversarial training. This combined loss pushes the network to generate visually convincing results beyond pixel fidelity.

Content Loss focuses on evaluating the similarity between the reconstructed image and the reference image at a perceptual level, aligning more closely with how humans perceive visual details. In the SRGAN paper, after introducing the option of using pixel-wise MSE loss, the authors propose the use of **VGG loss**, which better preserves visual features.

The VGG loss is defined as the Euclidean distance between the feature representations of a reconstructed image $G_{\theta_G}(I^{LR})$ and the reference image I^{HR}, both extracted from a pre-trained VGG network. This is expressed mathematically as:

$$l^{SR}_{\text{VGG}/i,j} = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \left(\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}(G_{\theta_G}(I^{LR}))_{x,y} \right)^2 \tag{2.14}$$

where $\phi_{i,j}(\cdot)$ denotes the activation (feature map) from the j-th convolution (after activation) before the i-th maxpooling layer of the VGG network, and $W_{i,j}$, $H_{i,j}$ represent the spatial dimensions of the feature map at layer (i, j).

In addition to the content loss, SRGAN adds **Adversarial Loss**, a generative component that encourages the network to produce outputs aligned with the manifold of natural images. This is achieved by training the generator to fool the discriminator. The adversarial loss is computed based on the discriminator's output probabilities D_{θ_D}, evaluated on the generator's predictions $G_{\theta_G}(I^{LR})$. It is formally defined as:

$$L^{SR}_{\text{Gen}} = \sum_{n=1}^{N} -\log D_{\theta_D}(G_{\theta_G}(I^{LR}_n)) \tag{2.15}$$

Here, $D_{\theta_D}(G_{\theta_G}(I^{LR}_n))$ represents the likelihood assigned by the discriminator that the generated image is a real high-resolution sample. This formulation improves gradient flow during training and promotes perceptual realism in the generated outputs.

2.5.3. Image Quality Metrics

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 numerical 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 perceptual metric designed to measure the 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, μ represents the mean, σ 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.

Learned Perceptual Image Patch Similarity (LPIPS) is a deep learning based perceptual metric focused on comparing deep features between reconstructed and HR images, calculating L2 distances in feature space to better align with human perception. The (LPIPS) metric is associated with image similarity—lower values indicate greater resemblance between two images, whereas higher values signify more substantial differences. LPIPS is calculated using a model trained on a dataset with human-judged perceptual similarity labels. It measures similarity by comparing the activations of the model for two given images. Although LPIPS goes beyond pixel-wise differences, it is vulnerable to adversarial attacks that produce result that are not aligned with human visual similarity judgment. the LPIPS distance between a reference image x and a distorted image x_0 is defined as:

$$d(x, x_0) = \sum_{l} \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot \left(\hat{y}^l_{hw} - \hat{y}^l_{0hw} \right) \right\|_2^2 \tag{2.16}$$

Where:

- $\mathcal{F}$ is a deep feature extractor (e.g., VGG or AlexNet).
- $\hat{y}^l$ and $\hat{y}^l_0 \in \mathbb{R}^{H_l \times W_l \times C_l}$ are unit-normalized feature maps from layer l, computed from input images x and x_0, respectively.
- $w_l \in \mathbb{R}^{C_l}$ is a learned per-channel weight vector.

- $\odot$ denotes element-wise multiplication.
- H_l and W_l represent the spatial dimensions of the feature map at layer l.

The total distance is obtained by computing the weighted squared ℓ_2 distance between the normalized feature activations, averaging over spatial locations, and summing across all selected layers. Notably, when $w_l = 1$ for all channels and layers, the LPIPS metric simplifies to a cosine distance between features [Zhang et al., 2018a].

Figure 2.11 illustrates how increasing levels of blurring degradation influence the values of PSNR, SSIM, and LPIPS. As the degradation increases, the **PSNR** value decreases, showing a decline in pixel-wise similarity to the reference image; higher PSNR indicates greater fidelity. The **SSIM** value also drops with stronger blur, as this metric is sensitive to structural and perceptual changes; values closer to 1 suggest higher visual similarity. The **LPIPS** score increases with degradation, since it captures perceptual dissimilarity in feature space; lower LPIPS values imply that the distorted image remains perceptually closer to the original.

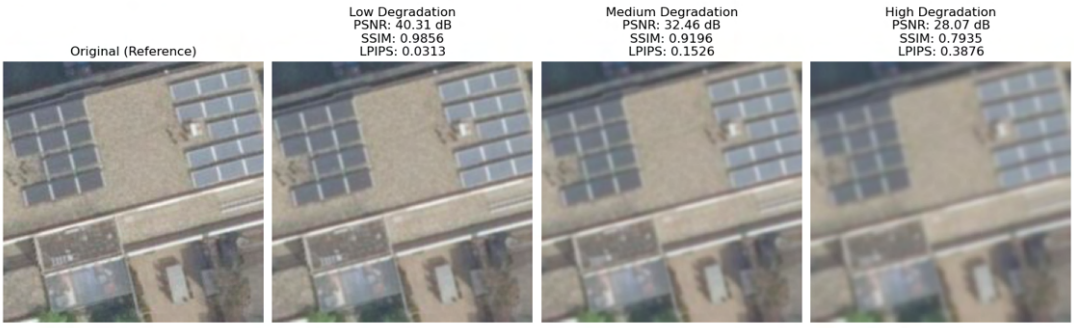

Figure 2.11.: Influence of blurring degradation in PSNR, SSIM and LPIPS

3. Related Work

This chapter presents an overview of super-resolution techniques as they have been proposed in the literature, particularly focusing on methods relevant to enhancing aerial imagery. The methods are commonly categorized into four main groups: interpolation-based methods, reconstruction-based algorithms, learning-based approaches (including deep learning), and transformer-based architectures [Lepcha et al., 2023]. The chapter connects traditional techniques with deep learning approaches, addressing the specific problems that the latter aim to solve in SR. It then focuses on literature concerning methods designed to provide results with well-defined edges and preserved geometry. Following this, studies specifically addressing super-resolution in the context of remote sensing are discussed, covering their applications and the unique challenges encountered in this domain. Lastly, the chapter describes examples of available benchmark datasets and relevant evaluation metrics utilized in these studies, expanding beyond the metrics presented in the theoretical background. Finally, a summary of the reviewed literature is provided, highlighting the existing research gaps that are addressed by this thesis.

3.1. Classical SR Methods

Early SR techniques aimed at enhancing the spatial resolution of digital images are generally divided into three groups: interpolation-based methods, frequency-domain methods, and reconstruction-based methods [Kawulok et al., 2024]. In remote sensing tasks such as pansharpening and hyperspectral/multispectral image fusion, classical methods have also included component substitution (CS), multi-resolution analysis (MRA), variational optimization (VO), spectral unmixing, and Bayesian modeling. However, this thesis focuses solely on RGB reconstruction and does not incorporate spectral fusion.

3.1.1. Interpolation-based methods

These approaches enhance resolution by estimating pixel intensities on an upsampled grid, typically relying either on fixed kernel functions or adaptive local structures. For example, Keys [1981] established a foundational method using bicubic interpolation based on predefined kernels. In contrast, more recent works such as Zhang and Wu [2006] and Li and Orchard [2001] introduced edge-guided interpolation methods that adapt to local structural features for improved sharpness and detail preservation. These methods were applied to grayscale images and color reconstructions from CCD samples. Their primary advantage lies in simplicity and computational efficiency. However, they often introduce visual artifacts, especially at higher scaling factors, resulting in noticeable blurring and jagged edges [Lepcha et al., 2023].

In an effort to address such limitations, Zhang et al. [2018b] introduced a bivariate rational fractal interpolation model that combines the benefits of both rational and fractal functions. Their method divides the LR image into textured and non-textured regions and applies structure-aware interpolation tailored to each region.

Despite their simplicity, speed, and ease of implementation, interpolation-based methods are inherently non-adaptive; they apply fixed mathematical models to all inputs without learning from data. While they perform reasonably well in low-frequency regions, they often suffer from aliasing, edge blurring, and loss of fine details—particularly in high-frequency areas where accurate reconstruction is most challenging [Deshpande and Patavardhan, 2019].

3.1.2. Reconstruction-based methods

Reconstruction based or regularization methods, rely on the assumption that LR images are generated by a degradation process involving blurring, warping, and downsampling. These methods focus on reversing that degradation by incorporating prior knowledge or constraints during the reconstruction step. They are generally divided into two categories: deterministic and stochastic. Deterministic methods encode assumptions about what the HR image should look like and regularize the solution using constrained least squares techniques [Deshpande and Patavardhan, 2019].

Sun et al. [2010] proposed a method that integrates contextual constraints to reconstruct high-frequency details by leveraging a patch-based model guided by neighboring contextual features. Xu et al. [2013] focused on enhancing detail using local fractal analysis of image gradients. Meanwhile, Wang et al. [2013] introduced an edge-directed prior to preserve sharp transitions and improve structural clarity, by adapting gradient magnitudes during left interpolation.

These methods differ primarily in the types of priors they use. For example, Zhang et al. [2012] proposed a regularization approach that combines both local and non-local priors learned from the LR image itself. The non-local prior leverages repeated structures (patch redundancy) across the image, while the local prior assumes each pixel can be predicted as a weighted average of its neighbors. While these methods often yield sharper results, they may introduce unwanted ringing artifacts around prominent edges or lead to discontinuities and pixel loss near fine structures.

3.1.3. Frequency-domain-based methods

These methods transform LR images into the frequency domain and reconstruct HR versions by estimating missing high-frequency components. These methods typically use either Fourier or Wavelet transformations. A foundational study by Tsai and Huang [1984] introduced the use of subpixel shifts between multiple LR images and exploited the shift property of the Fourier Transform to reconstruct HR images. This approach assumed band-limited input signals and demonstrated how aliasing in discrete Fourier Transforms (DFT) could be used advantageously. Later, Rhee and Kang [1999] proposed a model based on the Discrete Cosine Transform (DCT), offering lower memory consumption and computational load compared to traditional DFT-based methods. The main advantage of frequency-domain approaches lies in their low computational complexity and the ability to reconstruct high-frequency details by extrapolating existing spectral components.

However, these techniques require strong assumptions—such as the presence of only global translational motion and space-invariant blur during image capture—which rarely hold in real-world applications. As a result, they are generally not suitable for dynamic or unstructured environments [Deshpande and Patavardhan, 2019].

3.2. Deep Learning-Based SR Methods

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. The present study aligns with the SISR paradigm, as only single, non-redundant LR images are available. While numerous deep learning-based SR models exist, this study focuses on three main categories CNN-based, GAN-based, and Transformer-based architectures. This decision was made based on practical constraints, including limited time and computational resources, as well as the need to narrow down the scope to representative models from distinct architectural families. Preliminary experiments were conducted with selected models from each category to understand their behavior on aerial imagery; however, these early results served only as internal guidance and are not reported, as the goal was not to benchmark every method but to determine a suitable direction for further development.

Among these, GAN-based methods were chosen for this study due to their ability to recover perceptual detail and structural sharpness in urban environments—both critical for downstream object detection tasks. Moreover, the generative nature of GANs, which involves learning to synthesize realistic image textures through adversarial training, made them a more compelling and exploratory focus for this research.

3.2.1. CNN-Based

The use of convolutional neural networks (CNNs) for super-resolution (SR) began with the introduction of SRCNN by Dong et al. [2015]. 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. The network structure of SRCNN is shown in Figure 3.1.

However, the need to upsample LR images prior to processing increased computational complexity. To address this, Shi et al. [2016] proposed a sub-pixel convolutional neural network named ESPCN, which performs upsampling using convolutional layers at the sub-pixel level instead of conventional deconvolution layers. Although CNN-based methods generally rely on stacked convolutional layers and relatively simple network designs, they laid the foundation for the development of more advanced deep learning-based SR techniques.

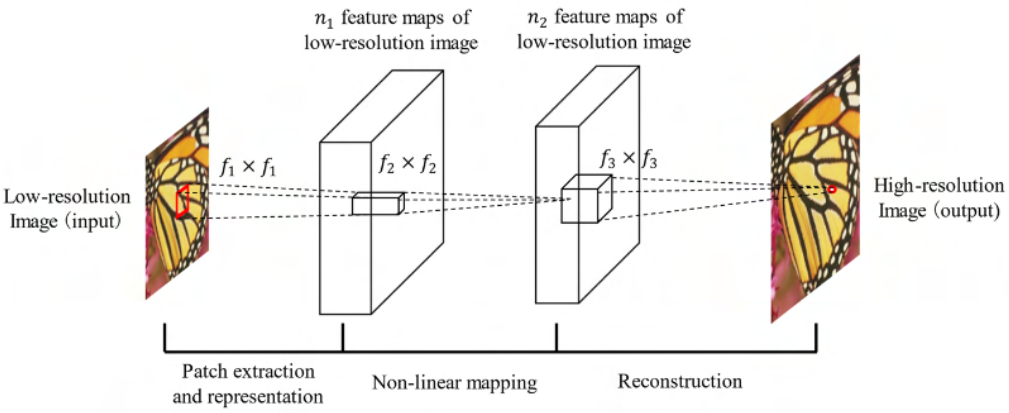

Figure 3.1.: Architecture of SRCNN adapted Dong et al. [2015].

Building on this foundation, VDSR was introduced by Kim et al. [2016a], with a focus on predicting the residual image. The final high-resolution output is obtained by adding this residual to the bicubically upsampled LR image. These architectures are referred to as residual networks, and they reflect a growing interest in increasing model depth to improve performance.

However, simply increasing the depth and width of a network introduces several challenges, including vanishing or exploding gradients and the risk of network degradation [Yu et al., 2024]. To mitigate these issues, He et al. [2016] proposed the ResNet architecture, which introduced a residual learning framework. This approach reformulates layers to learn residual functions with respect to their inputs, rather than attempting to learn unreferenced functions directly.

While deeper networks can lead to improved performance, they also introduce significantly more parameters, increasing the risk of overfitting and imposing greater memory and computational demands. To address this, recursive networks offer a more efficient alternative by reusing convolutional layers multiple times. This reduces the number of trainable parameters and the overall complexity of the model.

A prominent example is DRCN, introduced by Kim et al. [2016b], which employs a single convolutional layer recursively across the network. Outputs from all intermediate recursive steps, along with the final output, are passed to the reconstruction layer, which synthesizes the high-resolution image by leveraging all collected feature representations.

Conventional CNN-based super-resolution frameworks typically rely on synthetic low-resolution (LR) images generated by downsampling high-resolution (HR) data using fixed kernels. Although this approach provides controlled training conditions, it limits the model's ability to generalize to real-world degradations, which often differ significantly from such idealized downsampling schemes.

3.2.2. Transformer-Based

While CNN-based architectures have demonstrated strong capabilities in capturing local spatial patterns and hierarchical features, they are inherently limited in modeling long-range dependencies due to their localized receptive fields. Since super-resolution tasks require both fine-detail reconstruction and contextual understanding across larger spatial extents, researchers have explored architectures capable of modeling global interactions

more effectively. This exploration has led to the adoption of transformer-based models for super-resolution, leveraging self-attention mechanisms to overcome the limitations of convolutional and recurrent designs.

One key limitation of earlier SR frameworks was the computational overhead associated with processing high-resolution features throughout the network. To address this, post-upsampling architectures were proposed, in which most operations are carried out in the low-resolution feature space, and upsampling is deferred until the final stages. Although this strategy reduces memory consumption and computation time, it restricts the model's ability to refine high-frequency details during intermediate processing.

Transformer-based models such as TransENet [Lei et al., 2022] mitigate this issue by incorporating enhancement modules that fuse both low- and high-resolution feature representations after upsampling. This design strengthens the network's capacity to capture both coarse and fine-grained information, resulting in more detailed reconstructions. Yang et al. [2020] proposed a texture transformer that transfers relevant texture features from a reference image to the target high-resolution image through a learnable attention mechanism. Later, hybrid approaches were developed, such as the lightweight design by Lu et al. [2022], which combines CNNs for deep feature extraction with transformers that model long-range dependencies across similar image patches.

Figure 3.2 illustrates a typical transformer-based architecture for SR. The model uses a learnable texture extractor along with soft and hard attention mechanisms to transfer relevant features from auxiliary reference images to the low-resolution input. It applies a query-key-value (QKV) attention scheme followed by relevance embedding and soft attention layers. These enriched features are then fused with the upsampled LR image to produce the final output. This structure allows the model to preserve both local textures and global structures.

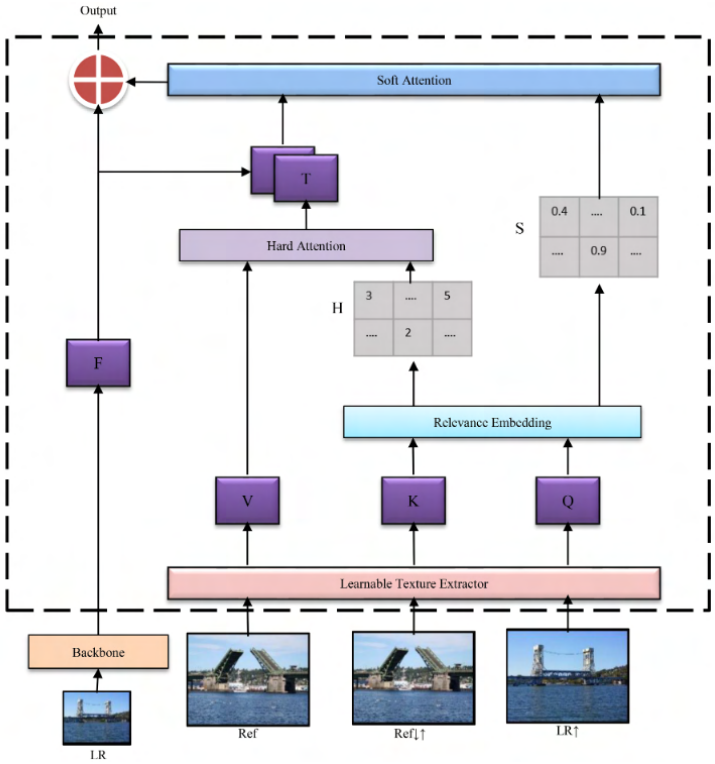

Figure 3.2.: Transformer-based architecture for image super-resolution adapted by Conde et al. [2022].

Compared to CNNs, which expand their receptive fields through stacked convolutional layers, and Recurrent Neural Neworks, which capture sequential dependencies recursively, transformers provide a more efficient mechanism for modeling global interactions. Their self-attention layers support parallel computation, reduce training time, and effectively learn long-range dependencies, making them well-suited for high-fidelity image super-resolution. However, despite their strong performance, transformer-based SR models still face challenges related to large parameter sizes and high computational demands, which remain areas of active research.

3.2.3. GAN-Based

While advancements in convolutional and transformer-based models have yielded significant improvements in the pixel-wise accuracy of super-resolved images, these methods often fall short in reconstructing fine textural details that are perceived as realistic by human observers.

Generative Adversarial Networks (GANs) present a distinct paradigm to address this limitation. By employing an adversarial training mechanism, GAN-based approaches shift the focus from strict pixel-level reconstruction towards generating outputs that reside on the manifold of natural images. This prioritizes perceptual fidelity, compelling the models to learn the distribution of natural image characteristics and produce results more congruent with human visual expectations.

Following the introduction of SRGAN in 2017, which aimed to obtain textural details

more aligned with human perception, Wang et al. (2018b) proposed Enhanced SRGAN (ESRGAN). This work improved upon SRGAN by introducing a Residual-in-Residual Dense Block (RRDB) as the basic network unit, adopting concepts from multi-level residual networks and dense connections simultaneously. However, despite improving perceptual quality, GAN-based approaches can sometimes introduce artifacts like checkerboard patterns and undesirable high-frequency components [Yu et al., 2024]. Consequently, further research has explored alternative architectures, such as the Multi-Perspective Discriminator GAN (MPDGAN) introduced by Lee et al. (2019). Numerous other architectures have also been built upon the ESRGAN foundation, aiming to enhance perceptual quality further, often by incorporating residual learning into dense blocks to increase network capacity. Incorporating Generative Adversarial Networks (GANs) into the SR process can produce high-quality images with superior perceptual characteristics. Jain et al. (2023) proposed ISRGAN, using SRGAN as the base model but featuring a modified loss function and network architecture, although it still resulted in blurry outputs. Subsequently, Xu et al. (2022) used a unified loss function to address imperfections with a model named TE-SAGAN. Although image SR algorithms based on GANs have improved visual perception quality, the inherent adversarial dynamic—the constant "fight" between the generator and discriminator—can result in an unstable training process, simultaneously increasing computational cost and memory consumption.

3.2.4. Edge-Preserving and Structure-Aware SR Models

Photo-realistic algorithms, such as those covered in the previous section, often yield outputs with low Peak Signal-to-Noise Ratio (PSNR) values and noticeable visual artifacts. In these methods, the reconstructed high-frequency details, particularly image edges, may be inconsistent with the ground truth, resulting in conspicuous artificial artifacts [Jiang et al., 2019]. Furthermore, conventional GAN methods tend to amplify noise or introduce spurious high-frequency details that are irrelevant to the actual content of the low-resolution input image. These limitations negatively impact subsequent computer vision tasks, such as object detection or land cover classification.

For this reason, researchers have investigated new ways to preserve these critical edge features. Mao et al. (2018), for instance, introduced EP-GAN, which utilizes edge structures from high-quality images as prior labels to guide the network. Later, Yang et al. (2017) based his method on the premise that a low-resolution image and its corresponding edge map can collectively inform the inference of sharp edge details in the high-resolution output during the reconstruction process. The input for this process was satellite imagery. Using an edge-guided recurrent residual network and assuming that the prior edge map label could be easily inferred, the method attempted to model the edge prior and retain high-frequency details. However, this was not completely achieved. The complex degradation processes affecting satellite imagery, including factors leading to areas of significant detail loss (described potentially as 'over-erosion' in the source context), caused reconstruction failures in these parts, leading to noisy results and false image edges.

Jiang et al. (2019) addressed this issue with Edge-Enhanced GAN (EEGAN), featuring a generator that consisted of two sub-networks: an ultra-dense sub-network and an edge enhancement sub-network. This ultra-dense sub-network consists of several dense blocks to ensure deep feature extraction. It also deals with artifacts and noise by utilizing a mask branch.

The most recent advancement in this domain was proposed by Ren et al. (2024, who introduced the Context-Aware Edge-Enhanced Generative Adversarial Network (CEEGAN). This super-resolution framework is specifically designed to reconstruct visually coherent

images suitable for real-world applications. The generator architecture features an edge feature enhancement module that fuses edge and contextual information, thereby improving the preservation of structural boundaries. Furthermore, the model incorporates a dedicated sub-network that exploits multi-scale edge features to generate a refined edge map. To ensure consistency between the predicted and actual edges, an edge loss term is introduced during training. Compared to prior GAN-based methods that attempted to integrate edge enhancement, CEEGAN achieves improved reconstruction quality while maintaining sharper and more semantically aligned edges.

3.3. Super-Resolution for Remote Sensing Images

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 remote sensing, including applications like object detection, classification, land surveying, and disaster monitoring, high-resolution imagery is a crucial component that contributes to their success. Consequently, reconstructing remote sensing images for super-resolution demands specialized approaches, and researchers have shown great interest in high-resolution remote sensing images [Wang et al., 2022a]. 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.

3.3.1. Existing Approaches in Remote Sensing SR

Throughout the years, many approaches have emerged, with researchers exploring methods that utilize extra bands or hyperspectral images to create outputs with rich detail. Most current remote sensing image super-resolution methods employ supervised learning. This involves using pairs of low-resolution (LR) and high-resolution (HR) remote sensing images to train models to learn the mapping from LR inputs to HR outputs.

Babu and Dubey (2021) utilized a double discriminator GAN named CDGAN, which processes both real HR images and SR images to improve the network's ability to discriminate low-frequency regions of remote sensing images. They also introduced a coupled adversarial loss function to fine-tune the network. Another approach extensively explored by researchers is modifying the attention mechanism within the networks. Yu et al. (2020) adapted an Enhanced Residual Channel Attention Module, allowing the network to concentrate on the most significant portions of the remote sensing images, thus extracting features that are more helpful for super-resolution. Using aerial images as input, Guo et al. (2022) proposed a novel dense generative adversarial network for real aerial imagery super-resolution reconstruction (NDSRGAN) designed to handle distorted details during reconstruction. In this approach, features are extracted from the LR images and then fed into a dense network.

Shermeyer and Van Etten (2019) also studied the effects of super-resolution on object detection performance in satellite imagery. After first addressing the challenges associated with detecting small objects—such as cars—that often span as few as 10 pixels, appear densely clustered, and exhibit complete rotational invariance, the authors also highlighted further difficulties such as the limited availability of labeled datasets for satellite images.

Most satellite imaging sensors cover broad areas and produce images with hundreds of megapixels, effectively resulting in ultra-high resolution images, but labeling such datasets remains costly and time-consuming. In their study, they quantified the effect of super-resolution on object detection performance across multiple input resolutions by calculating mean average precision (mAP) scores. Their results showed that applying super-resolution as a pre-processing step consistently improved object detection performance at most tested resolutions, highlighting the potential of SR techniques to enhance downstream tasks in remote sensing applications.

Other researchers, such as Zhang et al. (2023), extended beyond using super-resolution merely as a pre-processing step by developing a Super-Resolution Assisted Object Detection framework. Their model fuses multimodal data and performs high-resolution object detection on multiscale objects by incorporating assisted super-resolution learning while considering both detection accuracy and computational cost. The performance and inference speed of their proposed model, SuperYOLO, compared to the standard YOLOv5x detector, highlight the value of SR techniques in remote sensing tasks and pave the way for future research on multimodal object detection.

Although outside the direct scope of this thesis, research efforts have also investigated the application of super-resolution techniques to hyperspectral imagery. Hyperspectral images not only capture two-dimensional spatial information but also record detailed spectral signals across continuous bands. This rich spectral information provides critical insights about the material composition of objects, aiding in their accurate identification and classification within the observed scene.

To conclude, Super-resolution has a wide range of applications in remote sensing, which involves using sensors on platforms such as satellites and aircraft to gather geospatial data about the Earth's surface. Several examples of remote sensing-based super-resolution applications include Feature Classification and Object Detection, where enhancing image spatial detail improves accuracy for tasks like identifying buildings, pools, and vehicles. Agricultural Management also benefits, as SR technology aids crop and land use monitoring; converting LR to HR allows accurate distinction of crop species, detection of infestations and diseases, and precise fertilizer application, enabling efficient management. Disaster Monitoring and Emergency Response rely on SR technology as well, with HR imagery crucial for accurately assessing damage from natural disasters like floods and forest fires, facilitating prompt rescue and recovery. Environmental Monitoring utilizes HR remote sensing imagery for tasks such as monitoring water quality, tracking harmful algal blooms, and assessing coral reef health. Finally, Urban Planning and Land Management benefit from SR methodologies that help urban planners better understand urban environments; HR imagery allows for more accurate assessment of structures, transportation systems, vegetation coverage, and other factors informing urban expansion and land governance [Wang et al., 2022b].

3.3.2. Challenges in Remote Sensing SR

Although deep neural networks have significantly advanced remote sensing SR, several challenges remain. Data acquisition limitations exist because a single satellite sensor often cannot acquire images with both high spatial and high temporal resolution simultaneously due to technical and budget constraints, necessitating effective leveraging of time, space, and spectral correlations. Remote sensing images also possess Information Richness and Complexity; as long-distance observations encoding extensive sensor and scene information, they are richer and more complex than natural images, making the extraction of high-quality images and valuable information fundamental yet challenging for applications like classification.

Environmental Factors further complicate matters, as remote sensing is susceptible to conditions like atmosphere and weather, which can impede valuable information extraction. Lastly, despite deep learning offering high performance, Computational Demands increase with network complexity. This can lead to prohibitive hardware requirements for practical large-scale applications, especially considering the variability in sampling methods, image resolutions, and available datasets.

In conclusion, various types of SR methods have been explored for remote sensing images, consistently focusing on preserving or enhancing edges and fine details. Many techniques have been applied to satellite imagery, with fewer specifically tailored for aerial images. Different approaches have attempted to prioritize edges, but limitations persist. A common thread among these methods is the goal of surpassing the limitations of visual perception and enhancing image utility for analysis. Many techniques have relied on LR-HR image pairs with diverse training methodologies. However, obtaining perfectly aligned LR-HR image pairs in real-world scenarios presents a significant challenge, which limits the applicability of some existing methods [Wang et al., 2022b]. Given the critical need for high-resolution remote sensing imagery across numerous applications discussed, the topic of SR in remote sensing remains a very interesting and active area of research. The challenges outlined, particularly regarding data availability, computational cost, and the need for methods robust to real-world conditions, highlight existing gaps in the field. These gaps, such as the limited evaluation on temporally misaligned aerial imagery, the lack of edge-aware or structure-preserving GANs specifically for real-world SR, and the need for more robust evaluation strategies per category (e.g., land-use), motivate the exploration of alternative approaches and evaluation strategies, which will be discussed in the following chapter.

3.3.3. Benchmark 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 [Agustsson and Timofte, 2017].
- **BSDS300, BSDS500:** Widely used for benchmarking SR models [Martin et al., 2001].
- **Set5, Set14, Urban100:** Classic test datasets for evaluating SR performance [Bevilacqua et al., 2012; Zeyde et al., 2012; Huang et al., 2015].

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

- **AID:** Contains 10,000 images (600 × 600 pixels) featuring airports, beaches, deserts and 27 more classes. [Xia et al., 2017].
- **RSSCN7:** Includes 2800 images (400 × 400 pixels) categorized by season and scale, depicting farmland, residential areas, and industrial zones [Zou et al., 2015].
- **WHU-RS19:** Comprises 950 images (600 × 600 pixels) representing 19 scene categories, such as ports and parking lots [Dai and Yang, 2010].
- **UC Merced:** Features 2100 images (256 × 256 pixels) across 21 categories, including forests, rivers, and agricultural land [Yang and Newsam, 2010].

The dataset used in this study comprises aerial images of the Netherlands captured at two distinct time points and is described in detail in Chapter 5. These aerial images contain a variety of elements, similar to the benchmark datasets previously discussed. However, the final category definitions will be tailored and refined based on the specific requirements of the research use case.

3.4. Additional Evaluation Metrics

While the theoretical background previously introduced common evaluation metrics , this section focuses on additional evaluation strategies encountered throughout the literature review. These include both alternative perceptual metrics and task-specific metrics tailored to the practical use of SR outputs in downstream applications such as object detection.

To overcome the limitations of traditional metrics like PSNR and SSIM, alternative objective metrics have been developed. Prominent among these is the Natural Image Quality Evaluator (NIQE) [Mittal et al., 2012b]. NIQE operates as a blind image quality metric, meaning it does not necessitate a reference image. It predicts image quality by analyzing statistical features derived from characteristics observed in natural images, thereby evaluating test images based on quality-aware statistical models Wang et al. [2022b].

While objective metrics provide quantifiable results, the perceived quality by human observers remains a significant factor. Subjective evaluation metrics, such as the Mean Opinion Score (MOS) [Mittal et al., 2012a], rely on human participants to rate image quality. Although MOS offers valuable insights into human visual perception, its practical application is often constrained in large-scale evaluations due to its time-intensive nature, associated financial costs, and susceptibility to observer biases.

Within the scope of this research, the objective extends beyond merely generating visually appealing images. A primary goal is the production of SR images that possess sufficient functionality and suitability for subsequent downstream tasks, particularly object detection. Traditional metrics such as PSNR and SSIM, while valuable for evaluating perceptual fidelity, may not comprehensively capture the utility of the reconstructed images within an object detection pipeline. Recognizing this limitation, Shermeyer and Van Etten (2019) proposed the incorporation of object detection metrics to assess the applicability of reconstructed images for such tasks. This approach typically involves comparing ground truth bounding boxes with predicted bounding boxes on the SR outputs. A prediction is conventionally considered a true positive if its Intersection over Union (IoU) with a corresponding ground truth box exceeds a predefined threshold. This threshold can be adjusted based on the characteristics of the target objects, often set lower for smaller objects to enhance detection performance.

3.5. Summary and Research Gap

The literature review highlights the potential of Single Image Super-Resolution (SISR) for enhancing remote sensing imagery, which is crucial for various applications. However, several challenges remain. A key issue is the domain gap between synthetic training data and real-world remote sensing images, which often exhibit complex and unknown degradations. While significant advancements have been made in super-resolution techniques, especially those aiming to preserve edge information through edge-aware and structure-preserving GAN variants, their application to aerial imagery remains underexplored. Multiple studies

have attempted to improve the sharpness of reconstructed edges through attention mechanisms, edge priors, or edge-guided loss terms. In contrast to these strategies, the present study introduces a simpler architectural modification: additional convolutional layers are incorporated into the generator network to strengthen its capacity to reconstruct fine details. Although minimal, this adjustment targets a persistent weakness in existing models—their difficulty in reconstructing sharp structural boundaries in urban environments.

Standard SISR techniques often assume perfect temporal alignment, which is rarely the case in real-world scenarios where images are acquired at different times, leading to artifacts and reduced performance. 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. While advancements have been made in edge-aware and structure-preserving GANs for SISR, their effectiveness on real-world remote sensing data, particularly with temporal inconsistencies, is still limited.

Finally, the evaluation of SISR performance in remote sensing requires more robust strategies beyond generic metrics. 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. Task-specific and land-use dependent evaluation is essential to accurately assess the utility of super-resolved images for applications like land cover classification and object detection.

The research gap lies in developing SISR methods that can effectively bridge the domain gap, handle the complexity of remote sensing scenes, and specifically address the challenges posed by temporally misaligned aerial imagery. Further research is also needed to develop and benchmark evaluation metrics tailored to different land-use categories and downstream remote sensing tasks. 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.

4. Methodology

This chapter outlines the methodology followed in this research, beginning with the motivation for modifying the original SRGAN architecture. The baseline model is introduced, followed by a description of the architectural enhancements proposed to improve the reconstruction of edge structures commonly found in aerial imagery. In particular, the study explores the integration of a lightweight EdgeMaskBlock to the SRGAN generator to better preserve rooflines and building contours in super-resolved outputs. Finally, the chapter presents the training configuration, detailing the parameters used during both the supervised pre-training and adversarial fine-tuning phases.

4.1. Motivation for Architecture Modifications

Although SRGAN is capable of generating high-quality images, it tends to struggle with preserving fine texture details and often introduces artifacts. Previous research has proposed improvements to enhance SRGAN's visual output, such as the work by Wang et al. [2018a], which focuses on improving image quality and detail reconstruction. Notably, SRGAN was primarily evaluated on natural image datasets like Set5, Set14, and BSD100, which consist of photographs of animals, landscapes, and various objects.

In contrast, the objective of this study is to apply super-resolution to remote sensing imagery, where the accuracy of edges and structures—such as rooftops and building outlines—is particularly important. Therefore, visual realism alone is not sufficient; geometric fidelity must also be preserved in the reconstructed images.

To address these challenges, slight modifications to the original SRGAN architecture were introduced. The decision was inspired by edge-aware GAN models like EEGAN, proposed by Jiang et al. [2019], which integrate specialized subnetworks to improve edge sharpness and reduce noise. Given the practical constraints related to implementation complexity and programming experience, only lightweight adjustments were made to the SRGAN generator to better handle the structural characteristics of urban aerial images.

4.2. Baseline SRGAN Overview

SRGAN was selected as the baseline architecture for this study due to its simplicity, stability, and continued relevance in the field. Although introduced in 2017, SRGAN remains widely used and serves as the foundation for many later, more advanced GAN-based models. Its relatively straightforward design made it accessible for implementation, adaptation, and debugging within the limited timeframe of this research. More recent GAN architectures often feature complex components and dense connections that increase performance but also introduce significant implementation overhead and interpretability challenges. Since understanding the network behavior was essential for interpreting results and guiding architectural modifications, SRGAN provided a practical and well-documented starting point for exploration.

As discussed in Chapter 2, most super-resolution architectures aim to produce high-quality images using pixel-based accuracy metrics. SRGAN extends these approaches by introducing perceptual loss, enabling the generation of outputs that are more visually convincing to human observers. This shift in focus allows SRGAN to better capture high-frequency textures and photo-realistic detail.

The generator in the original SRGAN is based on a fully convolutional SRResNet model [He et al., 2016], consisting of an initial convolution layer, a series of residual blocks, and upsampling layers. This architecture is responsible for reconstructing the super-resolved image from the low-resolution input. The discriminator, which acts as a binary classifier, evaluates the authenticity of the generated image. Since the proposed modifications in this work target the generator, the structure of the discriminator remains unchanged.

4.3. Proposed Architecture Enhancements

To better suit the characteristics of aerial imagery—particularly the need for sharper edges and geometric precision—the generator was extended with two lightweight modules: an edge-aware refinement path and a mask-guided filtering mechanism.

The base generator processes the input through an initial 9×9 convolution layer with 64 feature maps, followed by a series of 16 residual blocks. These features are passed through additional convolution and upsampling layers to produce a preliminary super-resolved image.

In parallel, the input is also fed into a custom EdgeMaskBlock, which predicts both an edge map and a corresponding mask. The edge map captures structural features such as building contours, while the mask determines where these features should be emphasized in the final image. Both outputs are upsampled using bilinear interpolation to match the spatial dimensions of the main output. The edge features are then fused with the preliminary super-resolved image in a weighted manner, using the structural mask as guidance. The modified structure can be seen in Figure 4.1.

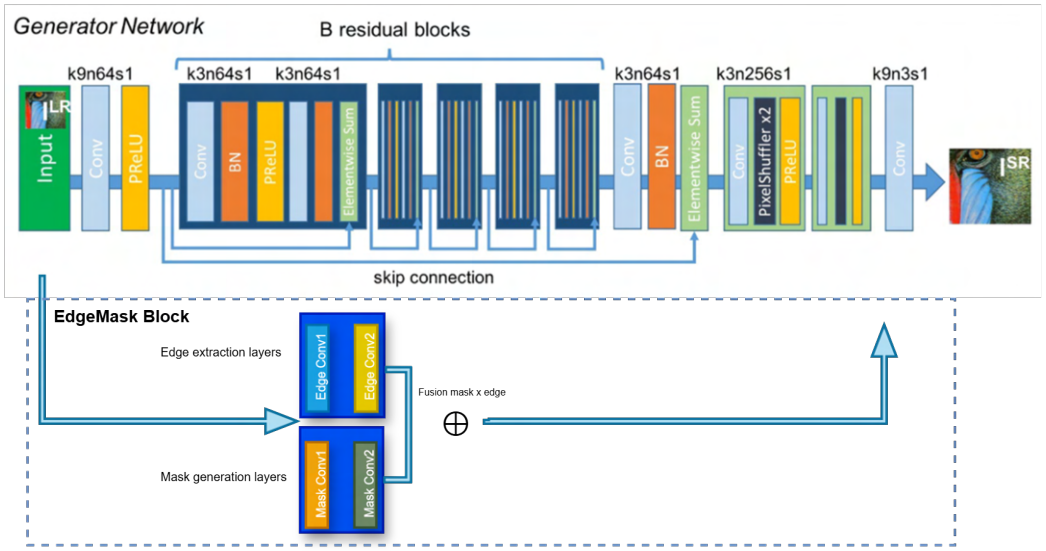

Figure 4.1.: Modified SRGAN architecture with the added `EdgeMaskBlock`. The EdgeMaskBlock is shown inside a blue, dotted-outline rectangle to highlight its integration into the main generator path.

This mask-guided fusion allows the generator to sharpen structural details selectively, without amplifying noise in flat regions. Figure 4.2 illustrates the working principle of the EdgeMaskBlock. For visualization purposes, simplified edge detection and masking techniques were used to demonstrate the role of each component. The demonstration was performed on a low-resolution input tile to illustrate how the EdgeMaskBlock enhances edges and preserves structure. In the actual SRGAN implementation, these operations are performed using convolutional layers within the network.

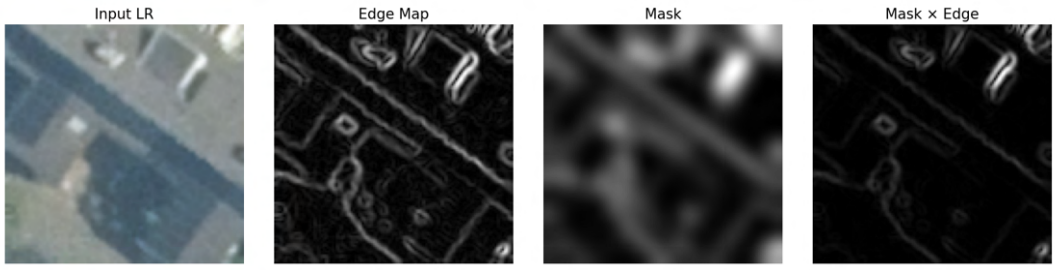

Figure 4.2.: Illustration of the Edge-Mask Block

The architectural adjustments are particularly relevant to the nature of urban aerial imagery. Buildings often present strong geometric patterns—rooflines, facades, and sharp object boundaries—which tend to become blurred in standard SRGAN outputs. By incorporating edge-aware refinement and spatially guided masking, the enhanced generator is better equipped to preserve these features. This leads to improved reconstruction of rooftops, clearer building contours, and more reliable spatial consistency in densely built environments.

4.4. Training Parameters

The SRGAN model was trained using the parameter settings described below. It was specifically trained to upscale images from 64×64 to 256×256, corresponding to a scale factor of $4\times$.

- **Patch size:** 24

  During training, the generator receives 24×24 low-resolution patches and produces super-resolved outputs of 96×96, using a scale factor of 4. These outputs are passed to the discriminator, which is configured to process 96×96 patches, matching the spatial resolution of the ground truth high-resolution tiles.

- **Upscaling factor:** $\times 4$ (resulting in 256×256 high-resolution patches)
  This scale was chosen to match the target resolution difference between input and output images in the aerial dataset.

- **Number of residual blocks:** 16
  The generator architecture includes 16 residual blocks, a design choice inspired by SRResNet and SRGAN literature. Residual connections help stabilize deep network training and improve convergence without significantly increasing complexity.

- **Batch size:** 25–50
  The batch size was adjusted based on the available hardware. Smaller batch sizes are sometimes preferred in GAN training to prevent the discriminator from overpowering the generator, which can destabilize learning. On the other hand, larger batch sizes typically improve convergence speed and overall SR performance. A balance was maintained to ensure training stability while fully utilizing GPU memory.

- **Pre-training epochs (MSE-only):** Variable

- **Fine-tuning epochs (perceptual + adversarial):** Variable
  The number of fine-tuning epochs was adjusted based on the experiment. In some cases, only pre-training was performed to test purely pixel-based super-resolution. In others, adversarial fine-tuning was enabled to improve perceptual quality.

- **Pixel loss (MSE) coefficient:** 1.0

- **Adversarial loss coefficient:** 10^{-3}

- **Total variation loss coefficient:** 0.0 (disabled)

- **VGG feature layer for perceptual loss:** `relu5_4`
  This refers to the specific activation layer used in the pretrained VGG19 network to compute the content (perceptual) loss. In this case, features are extracted after the fourth convolutional block in the fifth VGG stage, denoted as `relu5_4`. Using deeper feature layers like `relu5_4` allows the model to focus more on high-level semantic information and global structure, rather than low-level pixel similarity.

- **VGG rescale coefficient:** 0.006

- **Fine-tuning enabled:** True or False, depending on experiment goal

Ground truth (HR) and low-resolution (LR) image patches were extracted from the training datasets. The LR data was either synthetically generated using bicubic downsampling or taken directly from lower-resolution aerial inputs, depending on the experimental stage. Further details are provided in Chapter 5.

5. Implementation and Experiments

This chapter describes the complete experimental workflow used to train, fine-tune, and evaluate the proposed SRGAN-based approach for aerial image super-resolution. It begins by detailing the dataset characteristics and the preparation steps involved in generating high- and low-resolution tiles. The dual-iteration training strategy is then explained, distinguishing between the synthetic degradation setup and the real-world domain adaptation scenario. Following this, the chapter outlines how land use categories were leveraged for structured evaluation, and presents the training configurations and metrics used to monitor model performance. Finally, implementation details—including runtime characteristics and software tools—are summarized to ensure reproducibility and transparency of the overall pipeline.

5.1. Dataset Description

The dataset consists of TrueOrtho aerial imagery that were output from Readar B.V. pipeline. The initial aerial imagery before the orthorectification process was provided to Readar B.V. by Beeldmateriaal and 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.

5.1.1. Image Specifications and Metadata

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 RD (Rijksdriehoek): EPSG: 28992.

5.1.2. High-Resolution and Low-Resolution Imagery

High-resolution photographs 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 dataset with a ground pixel resolution of 8 cm.

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

Table 5.1.: TrueOrtho images of Delft in High-Resolution (HR) and Low-Resolution (LR) at Different Zoom Levels

High-Resolution (HR 8cm)	**Low-Resolution (LR 25cm)**
	
Delft HR Image	*Delft LR Image*
	
HR Image (1)	*LR Image (1)*
	
HR Image (2)	*LR Image (2)*

For this thesis, True Ortho images generated by Readar B.V.'s pipeline are used as the primary input data. Their orthorectification process precisely corrects distortions in elevated

objects by leveraging DEMs to accurately represent the 3D topology. As a result, roofs and other raised features are correctly positioned—providing the level of geometric precision required for tasks such as object detection [Kresse and Danko, 2012].

5.2. Strategy Overview

In this research, the methodology begins with Single Image Super-Resolution (SISR) in the first iteration. Here, only high-resolution (HR) images and synthetically downsampled low-resolution (LR) images (at 32 cm) are used to train the model, aiming to produce super-resolution results that closely approximate the ground truth. To avoid confusion, this output is referred to as the Super-Resolved (SR) 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.

The overall strategy, illustrated in Figure 5.1, is structured around two distinct training iterations. This approach was developed in collaboration with Readar B.V. and the TU Delft supervisory team to explore whether combining aerial images of the same area taken in different seasons could help improve the generalization performance of a GAN-based super-resolution model.

1. **Iteration 1** utilizes high-resolution (HR) aerial imagery with a spatial resolution of 8 cm, from which 256×256 pixel tiles are extracted. These HR tiles are synthetically degraded through bicubic downsampling to 64×64 pixels, simulating a low-resolution (LR) version with a spatial resolution of approximately 32 cm. A scale factor of 4 is used to train the SRGAN model to reconstruct super-resolved outputs at the original 256×256 resolution. The dataset is split into training and testing subsets with an 80–20 ratio to validate model performance within this synthetic setting.

2. **Iteration 2** introduces real LR aerial imagery with a spatial resolution of 25 cm. These images are tiled into 64×64 patches and used as input to the pre-trained model from Iteration 1, which performs super-resolution at a scale factor of 4. This yields output tiles with a spatial resolution of approximately 6 cm (256×256 pixels). In order to enable direct spatial comparison with the HR ground truth from Iteration 1, these super-resolved outputs are resampled into 8 cm and cropped to 200×200 pixels. The tiling configurations for both iterations are aligned to ensure that each tile corresponds to the same real-world spatial extent. This tiling alignment process is described in the following sections.

The complete strategy can be shown in Figure 5.1 below:

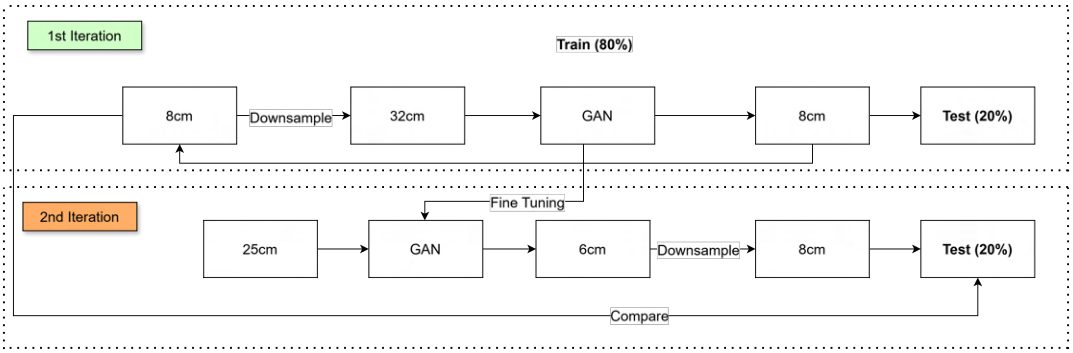

Figure 5.1.: Strategy Illustration

5.3. Data Preparation

To train and evaluate the SRGAN model effectively, two separate data preparation workflows were designed—one using synthetically degraded data and the other using real low-resolution (LR) imagery. These workflows are referred to as Iteration 1 and Iteration 2, respectively.

5.3.1. Iteration 1: Synthetic Data Preparation

The first iteration uses high-resolution (HR) aerial images at 8 cm resolution, which are synthetically downsampled to generate corresponding low-resolution inputs.

Tiling of High-Resolution Images

The input consists of HR orthophotos from Delft and Rotterdam. These images are divided into patches of 256×256 pixels, with a 10% overlap between adjacent tiles. This overlap was chosen to facilitate potential future mosaicking of the tiles, although the stitching of tiles into seamless outputs falls outside the scope of this thesis. These tiles serve as the ground truth for model training.

Synthetic Degradation via Bicubic Downsampling

To simulate low-resolution inputs, each HR tile is downsampled using bicubic interpolation by a factor of four, resulting in 64×64 LR tiles. These synthetic LR tiles act as inputs to the SRGAN model during training, while the original HR tiles serve as ground truth targets.

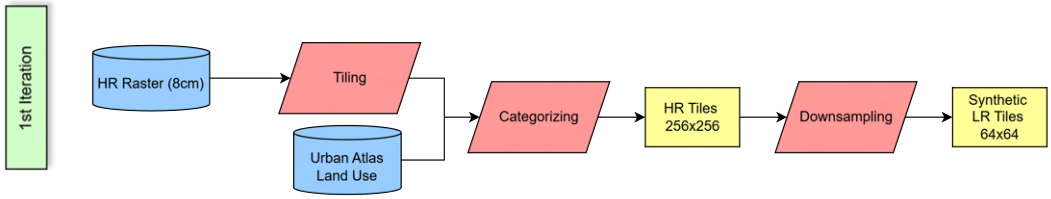

Figure 5.2.: Pre-processing pipeline for Iteration 1: HR tiles are downsampled to generate paired LR inputs.

5.3.2. Iteration 2: Real Low-Resolution Data Preparation

In the second iteration, real-world 25 cm resolution aerial imagery is used as LR input, paired with higher-resolution 6cm references to simulate a more realistic super-resolution scenario. In this case, orthophotos were taken from Delft, Rotterdam and Utrecht.

Tiling of Real LR and HR Images

The LR raster, captured at 25 cm resolution, is tiled into 64×64 patches with 25% overlap. At the same time, the HR raster—originally at 8 cm resolution—is first resampled to 6.25 cm and then tiled into 256×256 patches, also with 25% overlap. The increased overlap in both HR and LR tiling ensures better spatial correspondence across the datasets, enabling fair comparisons during evaluation.

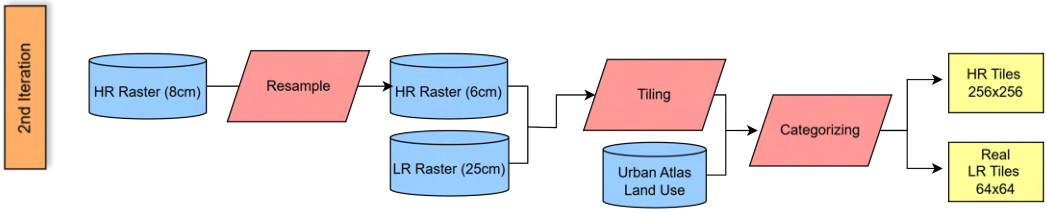

Figure 5.3.: Pre-processing pipeline for Iteration 2: Real LR tiles are paired with resampled HR tiles.

Matching HR Reference Tiles

To ensure consistency between the outputs of Iteration 2 and the high-resolution (HR) tiles from Iteration 1, both the spatial resolution and the physical ground area covered by each tile must be carefully aligned.

In Iteration 1, HR tiles were extracted from 8 cm orthophotos using a tiling window of 256×256 pixels with 10% overlap. These were downsampled by a factor of four via bicubic interpolation to generate synthetic low-resolution (LR) tiles of 64×64 pixels at 32 cm resolution. The SRGAN model was pretrained on these synthetic LR–HR pairs.

For Iteration 2, real-world 25 cm orthophotos were tiled into 64×64 LR patches with 25% overlap. Given that the pretrained model was designed to perform a $\times 4$ upscaling, the corresponding HR target resolution must be 6.25 cm. Thus, the original 8 cm imagery was

resampled to 6.25 cm and then tiled into 256 × 256 patches with 25% overlap to serve as HR references for fine-tuning.

After super-resolution inference in Iteration 2, the resulting 256 × 256 tiles at 6.25 cm resolution cover a larger ground area than the HR tiles from Iteration 1. This discrepancy arises both from the higher resolution and from the differing tiling strategies (25% vs. 10% overlap).

To enable a fair and spatially consistent comparison with the HR tiles from Iteration 1, each SR output from Iteration 2 is first downsampled to 8 cm resolution. Then, a central 200 × 200 crop is applied to match the physical ground area covered by the original 256 × 256 HR tiles from Iteration 1.

This two-step adjustment—resampling followed by cropping—ensures that evaluation metrics such as PSNR, SSIM, and LPIPS are computed over the same geographic footprint across both iterations. Importantly, consistent tile sizing was not only critical for metric evaluation but also for enabling the two-phase training strategy. Both iterations required inputs of 64 × 64 LR and 256 × 256 HR patches to align with the architecture of the SRGAN and its upscaling factor. Maintaining this consistency allowed the pretrained weights from Iteration 1 to be transferred and fine-tuned effectively in Iteration 2.

An overview of the tile dimensions used throughout the experiment—including the resolutions, overlap strategies, and how ground coverage was matched—is visualized in Figure 5.4. This alignment underpins both model training and evaluation and is essential to ensure comparability across iterations.

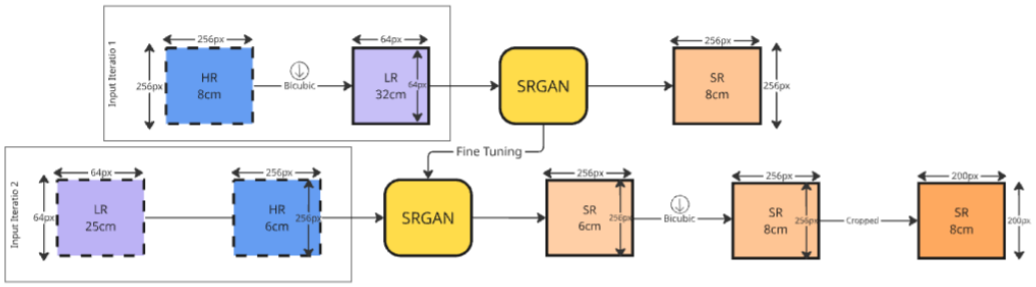

Figure 5.4.: Illustration of tile sizing throughout the experiment

5.4. Tile Categorization for Evaluation

Although the initial plan was to use land use categories to guide training data selection, the final approach utilizes categorization exclusively for evaluation purposes. This decision allows the performance of the super-resolution model to be assessed across different urban and non-urban contexts, without introducing bias or imbalance during training.

To support this evaluation, **Urban Atlas** land use shapefiles were employed [Copernicus Land Monitoring Service, 2024]. These shapefiles provide detailed geospatial information on land cover and use across Europe, including features such as impervious surfaces, vegetation, and water bodies.

Each image tile was spatially intersected with the land use polygons, and the dominant category was determined based on the largest area of overlap. This method ensures that each tile is consistently labeled according to its most representative land use type.

The resulting categorization enables performance analysis of the SRGAN model across diverse environments. Categories include densely built urban areas, agricultural fields, water bodies, suburban neighborhoods, and industrial or infrastructural regions. This classification system facilitates structured comparisons and helps quantify how well the model preserves features like rooftops or sharp boundaries in different settings.

By retaining a geographically diverse sample—from city centers to rural areas—the evaluation captures how well the model generalizes across varying spatial textures and land use conditions.

Initially, the land use categories were defined in the Urban Atlas dataset as follows:

Table 5.2.: Original land use categories and their merged classifications

Original Land Use Category	**Merged Category**
Continuous urban fabric (S.L.: >80%)	High-Density Urban
Discontinuous dense urban fabric (S.L.: 50%-80%)	High-Density Urban
Discontinuous medium-density urban fabric (S.L.: 30%-50%)	Low-Density Urban
Discontinuous low-density urban fabric (S.L.: 10%-30%)	Low-Density Urban
Discontinuous very low-density urban fabric (S.L.: <10%)	Low-Density Urban
Isolated structures	Low-Density Urban
Water	Non-Urban/Green
Sports and leisure facilities	Non-Urban/Green
Land without current use	Non-Urban/Green
Forests	Non-Urban/Green
Green urban areas	Non-Urban/Green
Pastures	Non-Urban/Green
Herbaceous vegetation associations (natural grassland, moors)	Non-Urban/Green
Open spaces with little or no vegetation (beaches, dunes, bare rocks, glaciers)	Non-Urban/Green
Wetlands	Non-Urban/Green
Arable land (annual crops)	Non-Urban/Green
Permanent crops (vineyards, fruit trees, olive groves)	Non-Urban/Green
Industrial, commercial, public, military, and private units	Industrial & Infrastructure
Port areas	Industrial & Infrastructure
Railways and associated land	Industrial & Infrastructure
Mineral extraction and dump sites	Industrial & Infrastructure
Other roads and associated land	Industrial & Infrastructure
Construction sites	Industrial & Infrastructure
Fast transit roads and associated land	Industrial & Infrastructure
Airports	Industrial & Infrastructure

To simplify the classification while maintaining the relevant information for the study, the categories were consolidated into four main groups: **High-Density Urban, Low-Density Urban, Non-Urban/Green, and Industrial & Infrastructure**. This restructuring ensures that

the dataset is optimized for assessing the impact of super-resolution on urban structures while avoiding excessive complexity.

Figure 5.5 presents an example of tile categorization, providing a visual reference for the typical content found within each category. Two sets of four tiles are shown from different geographic areas, both at high resolution (HR) and low resolution (LR), to illustrate the variety of textures and structures associated with each land-use type.

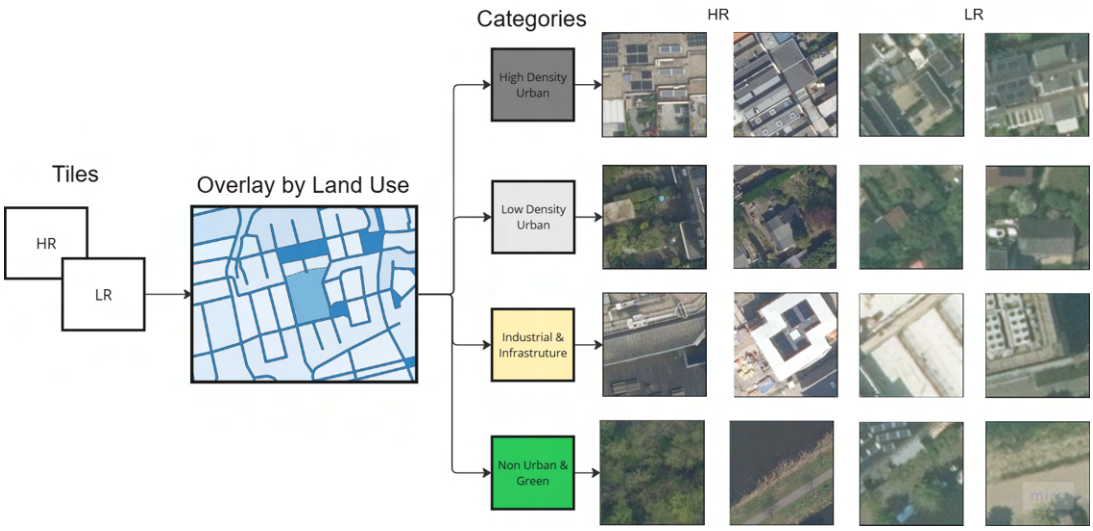

Figure 5.5.: Examples of HR-LR tiles in different categories

5.5. Training Setup

This section outlines the training configuration for both iterations of the SRGAN model. The training was divided into two distinct phases to accommodate the use of synthetic and real low-resolution imagery, respectively.

5.5.1. Iteration 1: Synthetic Super-Resolution

The first iteration focuses on training the SRGAN model using synthetic low-resolution images generated through bicubic downsampling of high-resolution aerial tiles.

- **Number of training samples:** 2,300, intended to provide a reliable baseline model.
- **Batch size:** 46
- **Pre-training epochs (MSE loss only):** 4,000
- **Fine-tuning epochs (perceptual + adversarial losses):** 2,000
- **Checkpoints:** Model weights during pre-training were saved every 800 epochs, resulting in five total checkpoints. During fine-tuning, generator and discriminator models were saved every 500 epochs, enabling visualization of intermediate outputs across training stages.

5.5.2. Iteration 2: Domain Adaptation on Real LR

In the second iteration, the model is fine-tuned using real-world low-resolution imagery captured at 25 cm resolution. The goal is to adapt the generator to realistic degradation patterns observed in operational aerial datasets.

- **Number of training samples:** 23,800, reflecting the primary focus of this research—capturing the full capacity of the model to fuse high-resolution features into realistic low-resolution inputs.
- **Batch size:** 46
- **Fine-tuning:** Enabled
- **Fine-tuning epochs:** 2,000
- **Pre-trained model:** `pre_trained_model_4000` (from Iteration 1)
- **Checkpoints:** Generator and discriminator models were saved every 500 epochs, as in Iteration 1, allowing comparative visual assessment of model evolution.

5.5.3. Evaluation Strategy

A variety of evaluation metrics were initially tested to assess how well they reflect the perceptual and structural quality of the super-resolved images. After careful analysis and comparison, three primary metrics were selected for final use: **PSNR**, **SSIM**, and **LPIPS**. These were chosen for their complementary characteristics and widespread use in the literature, as also discussed in the Related Work chapter.

PSNR (Peak Signal-to-Noise Ratio) was selected due to its sensitivity to pixel-level noise. Since noise suppression is an important factor in super-resolution, PSNR helps quantify fidelity, although it is known to penalize images that are perceptually sharp but deviate slightly from the reference.

SSIM (Structural Similarity Index) focuses on preserving structural information, which is particularly important for aerial imagery where features such as rooftops, road edges, and sharp transitions need to remain intact. SSIM evaluates luminance, contrast, and structural consistency, making it suitable for analyzing urban and non-urban textures.

LPIPS (Learned Perceptual Image Patch Similarity) is a deep learning-based metric that extracts feature embeddings from pre-trained networks and evaluates the perceptual distance between image patches. It has been shown to correlate well with human perception and captures high-level visual artifacts missed by pixel-based metrics.

To support the numerical evaluation, a custom script was developed to compute and aggregate all three metrics across both Iteration 1 and Iteration 2 datasets. This script also includes automated plotting routines that display central cropped regions of selected tiles, allowing for visual inspection of texture recovery and edge sharpness.

Each SRGAN output is compared to its respective ground truth and a baseline **bicubic interpolation output**. Bicubic interpolation serves as a classic example of an analytical, non-learning-based method for image enhancement and is commonly used as a baseline in super-resolution literature.

Segment Anything Evaluation

In addition to conventional quantitative metrics, the *Segment Anything Model (SAM)* model [Kirillov et al., 2023] was employed to evaluate how well the super-resolved outputs support downstream segmentation tasks. SAM is a promptable segmentation model trained on over

one billion masks from 11 million images. It incorporates a pre-trained Vision Transformer (ViT) encoder and supports zero-shot segmentation with strong generalization capabilities. All experiments were conducted using the `sam_vit_h_4b8939` checkpoint.

Technically, the SAM model (Figure 5.6 consists of three core modules: (1) a Vision Transformer-based **Image Encoder**, which computes patch-level embeddings with relative positional encodings; (2) a **Prompt Encoder**, which encodes sparse inputs like points and boxes using positional embeddings and dense prompts (e.g., masks) using convolutional layers; and (3) a **Mask Decoder**, which applies a modified two-way Transformer to perform cross-attention between image and prompt embeddings in both directions. The output is passed to a lightweight MLP that predicts mask logits across the image. This architecture enables SAM to produce valid masks quickly and robustly, even for ambiguous or overlapping prompts.

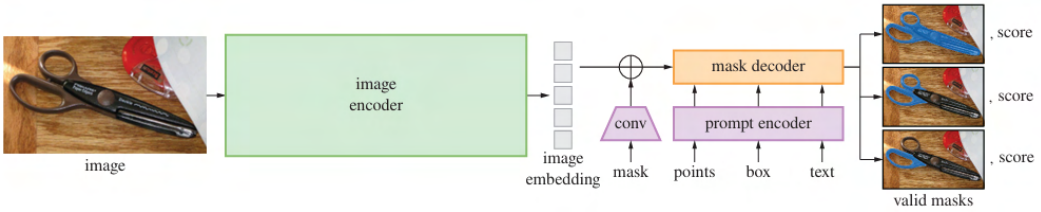

Figure 5.6.: Segment Anything Model (SAM) overview

Masks were generated for four image variants in each iteration:

- Iteration 1: GT-HR (8 cm), synthetic LR (32 cm), SRGAN (8 cm), Bicubic (8 cm)
- Iteration 2: GT-HR (8 cm), real LR (25 cm), SRGAN (8 cm), Bicubic (8 cm)

To improve perceptual clarity, masks are rendered with smoothed boundaries. When multiple masks of similar area are present, they are visualized using consistent colors to aid visual comparison. While the **number of detected masks** is displayed for reference, it is *not used as a standalone evaluation metric*, since a higher mask count does not necessarily indicate better segmentation quality. Instead, it serves as a contextual cue in combination with visual inspection, helping illustrate whether finer structures are preserved or overly merged. This qualitative visualization supports the assessment of how well each method maintains semantic detail in practice.

In addition to evaluating Iteration 1 and Iteration 2, PSNR was also applied to compare the baseline SRGAN and the modified edge-aware SRGAN introduced in Chapter 5. However, because PSNR alone does not fully showcase improvements along fine edges, we additionally calculated edge maps to quantify how well each model preserved structural boundaries. This metric complements SSIM and LPIPS by focusing specifically on high-frequency details, which are especially critical in aerial imagery analysis.

The edge maps were generated using the Canny edge detection algorithm proposed by Canny [1983]. It is a multi-stage algorithm that begins with noise reduction using a Gaussian filter, followed by computing the image's intensity gradient using Sobel kernels in both horizontal and vertical directions. After obtaining the gradient magnitude and direction, a full image scan is performed to suppress non-maximum pixels, retaining only the most significant edges. As a result, a binary image with thin, well-localized edges is produced.

These results are discussed in the opening section of Chapter 6, prior to the evaluation of each training iteration. This architectural comparison was conducted on a held-out valida-

tion subset of Iteration 1 tiles to ensure that the observed improvements were attributable to the network structure itself, independent of real-world degradation artifacts.

Semantic Segmentation Evaluation (Readar B.V.)

As part of the evaluation strategy, it was important to assess the performance of Readar B.V.'s semantic segmentation model on super-resolved imagery. The model produces four semantic classes: buildings, dormers, PV panels, and other. For this study, the evaluation focused on PV panels and dormers, as these classes are more sensitive to resolution quality and geometric clarity.

To compare the effectiveness of super-resolution, precision, recall, and F1 score were calculated for both the SRGAN output and the bicubic upsampled result. These metrics are based on the following components:

- **True Positive (TP):** A predicted object that correctly overlaps with a ground truth object of the same class.
- **False Positive (FP):** A predicted object that does not match any ground truth object (a hallucination).
- **False Negative (FN):** A ground truth object that was not detected by the model.

Precision is the proportion of predicted positives that are actually correct:

$$\text{Precision} = \frac{TP}{TP + FP}$$

High precision indicates that the model rarely produces false alarms. Low precision suggests a high number of false positives, where objects are detected but do not actually exist.

Recall is the proportion of actual positives that are correctly predicted:

$$\text{Recall} = \frac{TP}{TP + FN}$$

High recall means most relevant objects were detected, while low recall implies many true objects were missed.

The **F1 score** is the harmonic mean of precision and recall:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$$

This metric is especially useful for imbalanced datasets and reflects the balance between minimizing false positives and false negatives [Google Developers, 2023].

In addition to object-level metrics, building footprint segmentation was evaluated using the **Intersection over union (IoU)** metric. IoU is commonly used to assess the overlap between a predicted region and the ground truth region:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{TP}{TP + FP + FN}$$

High IoU values indicate strong alignment between predicted and actual building boundaries. A low IoU score suggests poor spatial consistency or inaccurate footprint delineation.

5.5.4. Training Monitoring via Intermediate Outputs

To qualitatively assess the training progression of SRGAN, generator model checkpoints were saved at regular intervals of 500 epochs. These checkpoints were used to generate super-resolved outputs on a fixed test tile, enabling direct visual comparison across training stages.

Figure 5.7 shows outputs from Epochs 500, 1000, 1500, and 2000 for a representative tile from Iteration 1, alongside their corresponding PSNR values. The top row displays the ground truth image at 8 cm resolution and the synthetic LR image at 32 cm used as input. The bottom row presents the outputs generated by the generator at each epoch. A clear qualitative improvement can be observed: at Epoch 500, the image is heavily distorted with hallucinated textures and blurry regions. At Epoch 1000, object boundaries begin to refine, although outliers such as black and blue pixels are still present. By Epoch 2000, the generator produces sharper structures, more coherent textures, and higher PSNR values, demonstrating successful learning progression.

Figure 5.8 provides the same visual monitoring approach for Iteration 2. In this case, the top row shows the ground truth 8 cm tile and the actual low-resolution 25 cm input. The bottom row again presents the outputs generated at various training checkpoints. At Epoch 500, the roof surfaces of buildings are rendered with pixelation and structural misalignment. Epoch 1000 shows minor improvements but with a noticeable drop in PSNR and further misaligned roof geometries. At Epoch 1500, PSNR increases, yet the outputs still exhibit pixelation on the building edges. Notably, the generator also appears to experiment with a different roof texture. Finally, by Epoch 2000, the output achieves the highest PSNR among all iterations, with better roof texture consistency, orthogonal structure alignment, and visibly reduced artifacts.

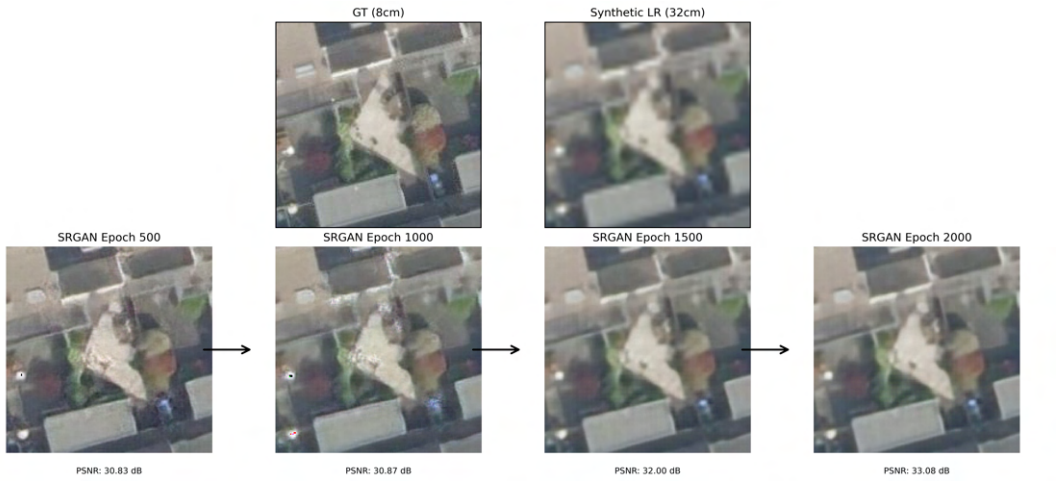

Figure 5.7.: Progressive outputs of SRGAN during Iteration 1 at Epochs 500, 1000, 1500, and 2000. Top row: ground truth and synthetic LR input. Bottom row: super-resolved outputs at different training stages.

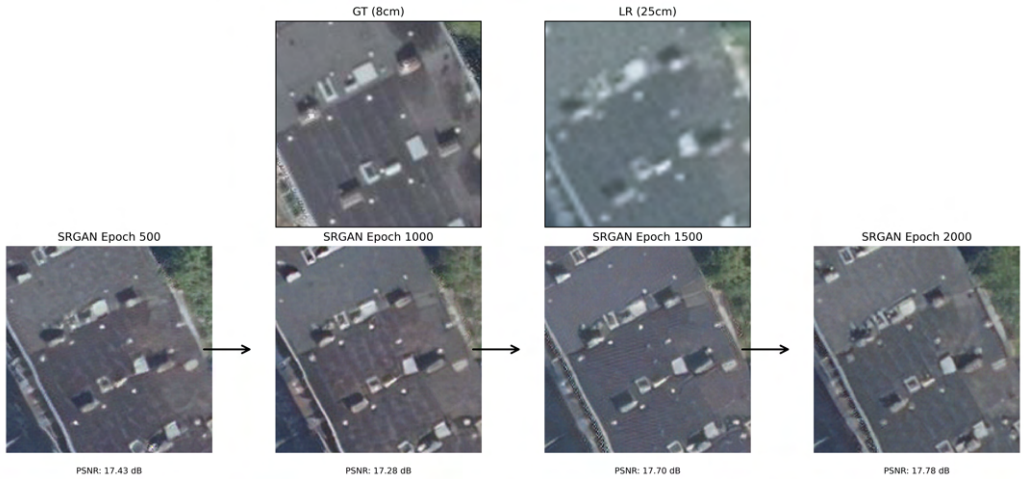

Figure 5.8.: Progressive outputs of SRGAN during Iteration 2 at Epochs 500, 1000, 1500, and 2000. Top row: ground truth and real LR input. Bottom row: outputs at various training stages.

This visual monitoring approach complements the recorded scalar metrics (L2 loss, PSNR, generator loss, and discriminator loss), which are included in this chapter to document convergence trends and learning stability. During both iterations, training checkpoints were saved for the generator, discriminator, and feature extractor (VGG). However, for evaluation purposes, only the generator was used to produce intermediate outputs.

In addition to visual inspection, training was monitored using scalar plots. For Iteration 1, which included a pre-training phase, Figure 5.9 shows the complete set of tracked metrics. Most notably, the L2 loss steadily decreased with each epoch, while the PSNR increased, indicating progressively more accurate reconstructions. Understanding the interplay between the generator and discriminator losses is particularly important. This relationship is often conceptualized as a minimax game, as outlined in Chapter 2, where the generator's objective is to successfully fool the discriminator, while the discriminator simultaneously works to identify the genuine ground truth outputs versus those created by the generator. A healthy adversarial balance is often indicated by opposing trends in their respective losses. Figure 5.9 reflects this interaction: the generator improves while not overpowering the discriminator, and both networks stabilize over time. This prevents the generator from collapsing into trivial solutions and ensures meaningful adversarial learning.

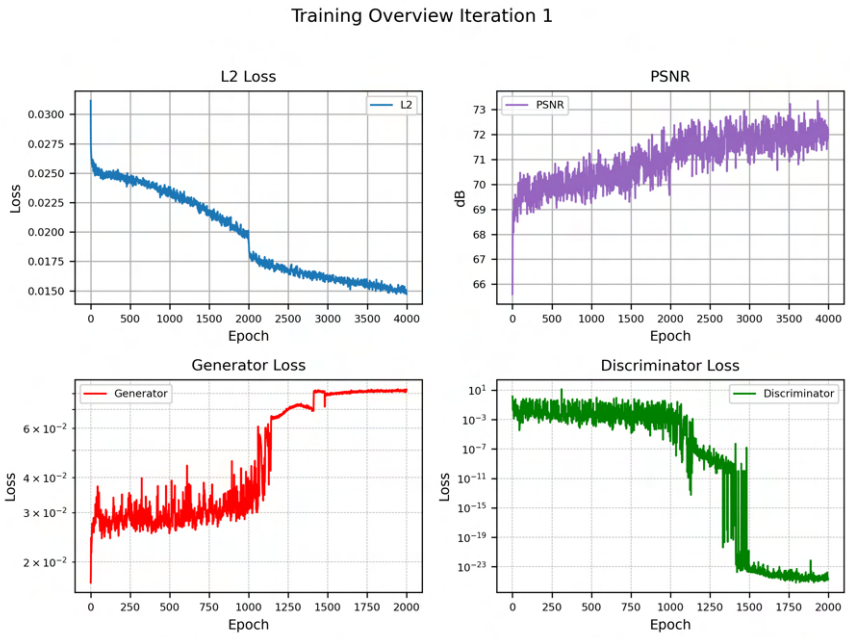

Figure 5.9.: Training losses for Iteration 1: L2 loss, PSNR, and adversarial losses. Note the smooth reduction in L2 loss and steady PSNR increase, with stable generator and discriminator dynamics.

Although the SRGAN training setup was designed to follow best practices, this project did not include a full hyperparameter tuning phase. Instead, a fixed set of configurations—such as patch size, batch size, and loss weights—was selected early on, guided by literature and feedback from Readar B.V., with the goal of maintaining a practical and interpretable pipeline. The primary strategy to improve generalization focused on increasing the number of training samples rather than tuning individual hyperparameters. This decision was necessary due to the time and computational resources available during the thesis period.

As visible in the training diagrams (Figure 5.9), the generator and discriminator losses initially evolve as expected, reflecting a balanced adversarial dynamic. However, in later stages, the discriminator loss sharply declines and stabilizes near zero, suggesting that it failed to distinguish between generated and real images. This behavior, known as discriminator collapse, can hinder further improvements in the generator's output quality. While the model still converged to reasonable results, this behavior likely limited the final performance gains. The observed issue highlights the known sensitivity of GANs to training instabilities and the importance of careful tuning—especially of adversarial loss weights, learning rates, and discriminator strength. Identifying and addressing this limitation was part of the learning process, and it underscores the trade-offs made when prioritizing architectural clarity and reproducibility over extensive tuning.

A similar pattern was observed in Iteration 2 (Figure 5.10), which served as a fine-tuning phase building on the pre-trained model from Iteration 1. Consequently, only adversarial losses are plotted. Once again, the generator and discriminator maintain a competitive equilibrium, confirming that neither network dominates. While a few early spikes are observed in the discriminator loss, this behavior is expected as the model adapts to a new data distribution. No collapse or instability is observed throughout the training run, supporting the

hypothesis that fine-tuning on real low-resolution inputs successfully refines the generator's capability without destabilizing adversarial dynamics.

Figure 5.10.: Adversarial loss trends during Iteration 2. The generator and discriminator maintain balanced competition, confirming healthy GAN training dynamics during fine-tuning.

5.5.5. Validation Strategy

Before validating Iteration 1 and Iteration 2, we also conducted a focused comparison between the baseline SRGAN and its edge-aware variant using the synthetic dataset. This allowed us to isolate the effect of architectural changes under controlled input conditions.

Each training iteration follows an 80–20 train/test split of the available tile datasets, ensuring a consistent framework for evaluation across synthetic and real input scenarios.

Iteration 1 is trained exclusively on synthetically degraded imagery derived from high-resolution (HR) orthophotos of **Delft** and **Rotterdam**. These HR images at 8 cm resolution are downsampled to simulate low-resolution inputs at 32 cm. The model is trained on 80% of the resulting tile pairs and validated on the remaining 20%.

An overview of the geographic locations used for training and testing in both iterations can be found in Appendix A.1.

Iteration 2, by contrast, incorporates real low-resolution imagery captured at 25 cm resolution from **Delft, Rotterdam, and Utrecht**. The model is fine-tuned using these real LR images paired with corresponding resampled HR tiles (6.25 cm). While 80% of these tiles are used for training, the validation process in Iteration 2 goes beyond a simple held-out 20% test set.

Specifically, in Iteration 2, we also investigate the model's ability to **generalize to spatially distinct areas**—regions that are geographically far from those seen during training. This helps evaluate whether the learned representations from the pre-trained and fine-tuned SRGAN can transfer across different urban structures, textures, and environmental conditions. Since the training data might be localized within certain neighborhoods of the three cities, testing on disjoint areas within the same cities provides a robust indicator of spatial generalization.

This two-tiered validation—(1) standard held-out split, and (2) generalization to distant tiles—ensures that the model is not simply overfitting to the training distribution, but is capable of meaningful super-resolution across the broader urban fabric of the Netherlands.

During validation, the following outputs are generated for each test tile:

- PSNR, SSIM, and LPIPS metric scores
- Visual comparisons against the bicubic baseline
- Segmentation outputs using the Segment Anything Model (SAM)
- Semantic outputs using Semantic Segmentaton Model provided by Readar B.V.

Together, these metrics assess both the perceptual and functional performance of the super-resolution pipeline.

5.6. Implementation Details

This project was implemented using Python, with the core deep learning framework developed in PyTorch. Development and debugging were conducted in the PyCharm IDE, running on a Windows system with access to the Windows Subsystem for Linux (WSL) to enable Linux-based execution. Matplotlib was used to plot the results.

5.6.1. Hardware Setup

Training and testing were performed on different hardware setups:

- **Training (SRGAN)**:
  - NVIDIA GeForce RTX 2070
- **Preprocessing and Evaluation**:
  - NVIDIA GeForce RTX 3060 (local machine)

5.6.2. Software and Tools

The following tools were used throughout the project:

- **QGIS**: Visualization and spatial analysis of aerial and land-use data.
- **GDAL (Geospatial Data Abstraction Library)**: Raster pre-processing, virtual raster creation, and mosaic generation using tools.
- **DBeaver + SQL**: Database access and spatial query execution.
- **Overleaf**: LaTeX-based thesis writing and formatting.
- **Miro** and **draw.io**: Architecture diagrams and process illustrations.

5.6.3. Runtime Performance

Runtime performance during preprocessing and training stages was monitored to evaluate the feasibility of large-scale experimentation. Table 5.3 reports average processing speeds for key stages, including tile generation, land-use-based categorization, evaluation metric computation (LPIPS, PSNR, SSIM), and SAM-based semantic segmentation. Additionally, the table summarizes the training configuration and total runtime for each SRGAN iteration on an RTX 2070 GPU.

Table 5.3.: Runtime performance and training configuration overview.

Preprocessing and Evaluation Steps		
Process	**25 cm (LR)**	**8 cm (HR)**
Tiling and Saving (Tiles/s)	2.5 – 3.5	1.5 – 2.5
Categorization (Tiles/s)	6 – 7	9 – 10
Metric Calculation (Tiles/s)	10 – 12	12 – 14
SAM Segmentation (Min/Tile)	1 – 2	2 – 3
SRGAN Training		
Iteration	**1**	**2**
Pre-train Epochs	4000	0
Fine-tune Epochs	2000	2000
Train Samples	2,300	23,800
Total Runtime	15h 36m 46s	119h 17m 57s

6. Benchmarks and Results

This chapter presents a detailed evaluation of the super-resolution models developed in this thesis. The analysis covers architectural variants, including the baseline SRGAN and the edge-aware SRGAN, and investigates model performance under different data conditions and application scenarios. Results are organized into architectural comparisons, quantitative benchmarks, qualitative visual assessments, and generalization studies. Particular attention is given to Iteration 1, which uses synthetically degraded data, and Iteration 2, which processes real low-resolution inputs with temporal misalignment. Model effectiveness is assessed using image quality metrics (PSNR, SSIM, LPIPS), visual inspection, and downstream segmentation performance using the Segment Anything Model (SAM). This multi-faceted evaluation examines the extent to which GAN-based super-resolution can enhance 25cm aerial imagery to 8cm, with a focus on its applicability for object detection tasks.

6.1. Comparison of Baseline and Edge-Aware SRGAN

In the first experiment, the original SRGAN model is compared to the modified variant that incorporates an edge-mask refinement block. Visual analysis (Figure 6.1) reveals that the edge-aware SRGAN reconstructs sharp boundaries more effectively, particularly along building outlines and rooftop edges. To quantify this improvement, PSNR is computed between the edge maps extracted from each output and the corresponding high-resolution reference, highlighting the extent to which structural detail is preserved.

In the example shown, the edge-aware SRGAN not only improves the visual clarity of fine structures—such as solar panels and rooftop edges—but also achieves a higher PSNR compared to the baseline SRGAN. This confirms that sharper edge reconstruction contributes positively to the metric and reflects better alignment with the structural content of the ground truth.

As illustrated in Figure 6.1, the ground truth tile and its extracted edge map are presented on the left. The ground truth edge map successfully captures nearly all structural details from the original HR tile. In the middle, the SRGAN baseline output and its corresponding edge map are shown. While the baseline achieves a relatively high PSNR, it fails to recover many fine structures, particularly the smaller rooftop elements. On the right, the edge-aware SRGAN output demonstrates a marked improvement: almost all rooftop solar panels and structural details are reconstructed, producing an edge map that closely matches the ground truth. The results suggest that the inclusion of the edge-mask refinement block leads to both perceptual and quantitative gains in structural fidelity.

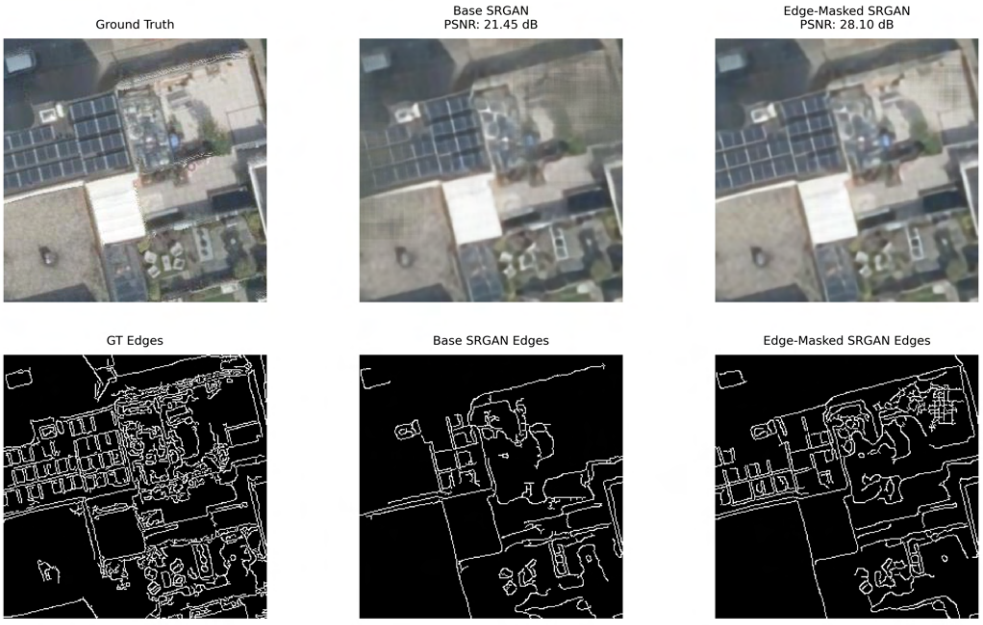

Figure 6.1.: Edge map comparison between Ground Truth, baseline SRGAN, and edge-aware SRGAN. The edge-aware model recovers rooftop structures more accurately, closely matching the ground truth.

Figure 6.2 presents a similar comparison using outputs from Iteration 2. Once again, the edge-aware SRGAN outperforms the baseline by capturing finer rooftop details, as evidenced by the enhanced edge maps. In this case as well, the edge-aware model achieves a higher PSNR, further confirming its ability to reconstruct high-frequency structures more accurately. The consistency of these improvements across both iterations supports the robustness of the edge-mask refinement strategy.

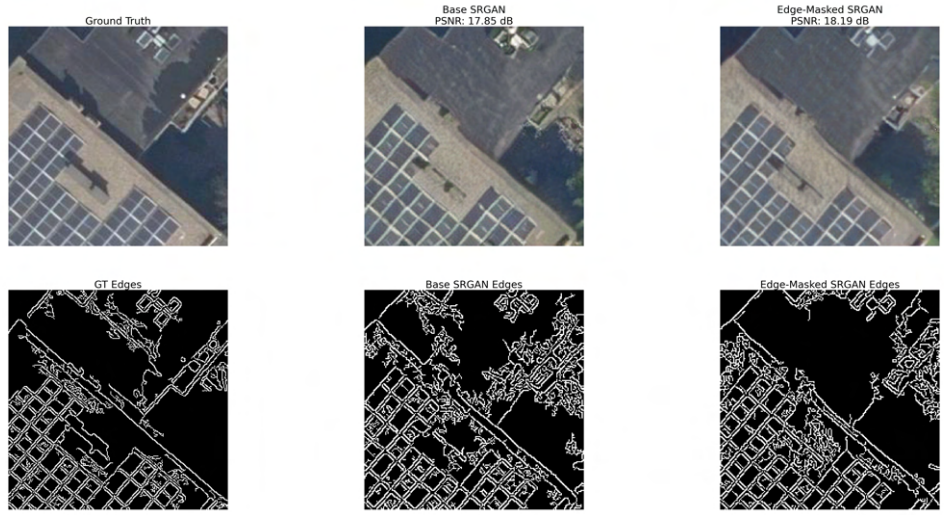

Figure 6.2.: Edge map comparison between Ground Truth, baseline SRGAN, and edge-aware SRGAN (Iteration 2). Enhanced structural fidelity is visible in the edge-aware output, along with higher PSNR.

To avoid repetition, the edge-aware SRGAN introduced in this work will hereafter be referred to simply as *SRGAN*, unless the baseline model is explicitly referenced.

6.2. Quantitative Evaluation

To assess the overall benefit of the edge-aware refinement block, a direct comparison was made between the baseline SRGAN and the edge-aware variant for both Iteration 1 and Iteration 2. Table 6.1 summarizes the average PSNR, SSIM, and LPIPS values, allowing for a side-by-side evaluation of their performance.

Table 6.1.: Average evaluation metrics comparing SRGAN with and without edge-mask refinement for Iteration 1 and Iteration 2. Best values are shown in bold.

Metric	**PSNR (dB)**		**SSIM**		**LPIPS**	
	Base	Edge-Aware	Base	Edge-Aware	Base	Edge-Aware
Iteration 1	28.2952	**29.7570**	0.8423	**0.8597**	0.3027	**0.2769**
Iteration 2	17.4202	**17.7979**	0.3916	**0.4202**	0.5636	**0.5572**

The edge-aware SRGAN outperforms the baseline across all three metrics in both iterations. In Iteration 1, the average PSNR improves by approximately 1.46 dB, SSIM increases by 0.0174, and LPIPS decreases by 0.0258. Similarly, in Iteration 2, the edge-aware model shows consistent gains with a 0.38 dB PSNR increase, a 0.0286 improvement in SSIM, and a slight reduction in LPIPS. These results confirm that the proposed architectural enhancement leads to both perceptual and quantitative improvements in super-resolution quality.

In Table 6.2, the performance of Iteration 1 is evaluated in terms of PSNR, SSIM, and LPIPS metrics. In this table, we compare the results of our super-resolution model against standard bicubic upsampling. Overall, bicubic upsampling performed better than the SRGAN model in terms of PSNR, with average differences across categories ranging from 1 to 3 dB. This indicates that bicubic interpolation provides slightly higher pixel-level fidelity compared to the SRGAN output. However, this outcome was expected, as bicubic interpolation operates by smoothly estimating pixel values within the existing image, whereas the SRGAN model reconstructs details from noisy low-resolution inputs, naturally leading to a slight reduction in PSNR. In terms of SSIM, the results are inverted: the SRGAN model outperforms bicubic interpolation by achieving higher SSIM values across all categories. This demonstrates that SRGAN produces images that are structurally more similar to the ground truth, capturing important textures and edges more effectively than bicubic interpolation. Regarding LPIPS, the SRGAN model outperforms bicubic interpolation in the High-Density Urban and Industrial & Infrastructure categories, which typically contain more complex textures and finer structures. In the remaining categories, bicubic slightly outperforms SRGAN, but the margins are small.

Table 6.2.: Evaluation Metrics Iteration 1. Higher PSNR and SSIM values indicate better reconstruction quality, while lower LPIPS values indicate higher perceptual similarity to the ground truth.

Category	**PSNR**		**SSIM**		**LPIPS**	
	SR	**Bicubic**	**SR**	**Bicubic**	**SR**	**Bicubic**
High-Density Urban	30.7008	**32.6048**	**0.8262**	0.7842	**0.2762**	0.2835
Low-Density Urban	30.7107	**33.3904**	**0.8768**	0.8572	0.3059	**0.3017**
Industrial & Infrastructure	31.8787	**34.1510**	**0.9098**	0.8760	**0.2638**	0.2690
Non-Urban (Green)	32.6888	**36.6673**	**0.9637**	0.9590	0.3009	**0.2767**

Table 6.3 presents the evaluation results for Iteration 2 and shows that SR consistently achieves higher PSNR values across all categories compared to Bicubic interpolation, indicating better pixel-wise reconstruction fidelity. However, SR underperforms in SSIM relative to Bicubic. This is likely due to SR recovering finer edges and textures that are penalized by SSIM, which is highly sensitive to small structural differences, especially when evaluated on the Y channel.

Regarding perceptual quality, SR achieves better (lower) LPIPS scores in the Low-Density Urban and Non-Urban (Green) categories. This suggests that in these environments, characterized by more natural and less repetitive textures, SR enhances perceptual similarity to the ground truth more effectively than Bicubic upsampling.

Table 6.3.: Evaluation Metrics Iteration 2. Higher PSNR and SSIM values indicate better reconstruction quality, while lower LPIPS values indicate higher perceptual similarity to the ground truth.

Category	PSNR		SSIM		LPIPS	
	SR	Bicubic	SR	Bicubic	SR	Bicubic
High-Density Urban	**19.3522**	19.2876	0.4748	**0.6220**	0.5879	**0.5585**
Low-Density Urban	**18.9724**	18.7459	0.4460	**0.5952**	**0.5570**	0.5812
Industrial & Infrastructure	**20.2038**	20.0323	0.5248	**0.6558**	0.5756	**0.5652**
Non-Urban (Green)	**22.0752**	21.5057	0.5614	**0.6528**	**0.5761**	0.5833

While quantitative metrics such as PSNR and SSIM provide a standardized means of evaluating reconstruction performance, they often underestimate the perceptual quality of GAN-generated outputs. This is particularly evident in urban scenes where SRGAN successfully reconstructs fine architectural details that are visually clear but differ slightly in pixel alignment or intensity from the ground truth. Metrics like PSNR are highly sensitive to such differences, penalizing outputs that deviate from exact pixel values even when those deviations improve visual clarity. Similarly, SSIM emphasizes local structural similarity but can misrepresent quality in cases where GANs reconstruct sharp edges with minor shifts. As a result, despite SRGAN outperforming Bicubic interpolation in visual coherence and segmentation utility—as seen in later evaluations—its metric scores may not fully reflect the perceptual improvements achieved. This gap reinforces the importance of incorporating both quantitative and qualitative analyses when evaluating super-resolution models, especially in the context of GANs.

6.3. Qualitative Evaluation

This section complements the quantitative metrics with a visual inspection of the outputs generated by SRGAN and Bicubic interpolation. Visual analysis provides additional insights into perceptual quality and structural clarity that may not be fully captured by numeric metrics alone. Additional visual comparisons for each example shown in this section are included in Appendix A.2 to provide a more detailed overview of the SRGAN and Bicubic outputs across various tiles and categories.

6.3.1. Visual Comparisons

Iteration 1

This section presents visual comparisons, alongside the quantitative results. Since numerical metrics alone cannot fully capture the perceptual quality of the super-resolved images, a visual inspection was also conducted to assess the outputs more thoroughly. This section includes plots showing results from both Iteration 1 and Iteration 2.

For Iteration 1, Figure 6.3 presents the results grouped by category, allowing the performance of the algorithm to be evaluated across different types of tiles. In each comparison, the images are organized from left to right as follows: the ground truth full tile at 8 cm resolution, a zoomed-in crop of the ground truth, the synthetic low-resolution input at 32 cm, the bicubic upsampled image at 8 cm, and the SRGAN super-resolved output at 8 cm. Below the bicubic and SRGAN outputs, the corresponding PSNR, SSIM, and LPIPS scores are

shown to facilitate side-by-side comparison.

Figure 6.3.: Visual comparison of category-wise results for Iteration 1, alongside PSNR, SSIM, and LPIPS scores.

Despite the quantitative evaluation often indicating that the SRGAN model generally underperforms Bicubic interpolation in terms of raw PSNR metrics, the visual evidence, particularly in the tiles plotted in Figure 6.3, reveals a different story regarding perceptual quality. It becomes evident that the SRGAN outputs are significantly more visually pleasing compared to those produced by Bicubic upsampling, which tend to appear noticeably blurry. In contrast, the SRGAN results exhibit considerably greater clarity and detail. For example, in the High-Density Urban tile presented, features such as solar panels on rooftops are clearly distinguishable, and building outlines are reconstructed with high sharpness, accurately reflecting their structural contours. This demonstrates SRGAN's capability to successfully enhance fine architectural details, which are typically lost or heavily smoothed by Bicubic upsampling.

A similar pattern is observed within the Industrial & Infrastructure tile shown in the figure, where the SRGAN output closely resembles the ground truth image visually. This superior visual consistency is also supported by the higher SSIM scores obtained by SRGAN

in this category, indicating a better preservation of structural similarity. Furthermore, for both the High-Density Urban and Industrial & Infrastructure categories, SRGAN achieved higher PSNR scores than Bicubic upsampling, underscoring its superior reconstruction fidelity in these complex environments. The lower LPIPS values recorded for SRGAN in these categories further reinforce its advantage in terms of perceptual similarity to the ground truth.

Moving to the Low-Density Urban tile, the SRGAN model successfully reconstructs a variety of roof textures, such as those of metal and gravel surfaces, while effectively preserving crucial structural features. Finally, when examining the Non-Urban Green tile, although the visual distinctions might be less immediately striking to the human eye compared to urban areas, SRGAN still demonstrates an improved reconstruction of natural textures like tree canopies and vegetation patterns. These improvements in the Non-Urban Green category are likewise supported by the metric results, which show higher perceptual and structural similarity for SRGAN compared to Bicubic interpolation.

Iteration 2

In the quantitative evaluation of Iteration 2, it was observed that SRGAN generally outperformed bicubic interpolation in terms of PSNR, but underperformed in SSIM, with LPIPS scores showing a balanced outcome without a consistent advantage across all categories. Iteration 2 is the primary focus of this work, as it demonstrates how a model trained with LR-HR tile pairs captured during different seasons can reconstruct a high-resolution image from a low-resolution input whose ground truth does not truly exist. The ground truth tile and the LR input differ due to seasonal changes, variations in shadowing, and other environmental factors.

Figure 6.4 presents category-wise visual examples. From left to right, each comparison includes the ground truth tile at 8 cm resolution (not used during training), a zoomed-in window for better detail visibility, the 25 cm low-resolution input, the bicubic upsampled version at 8 cm, and finally the SRGAN super-resolved output at 8 cm.

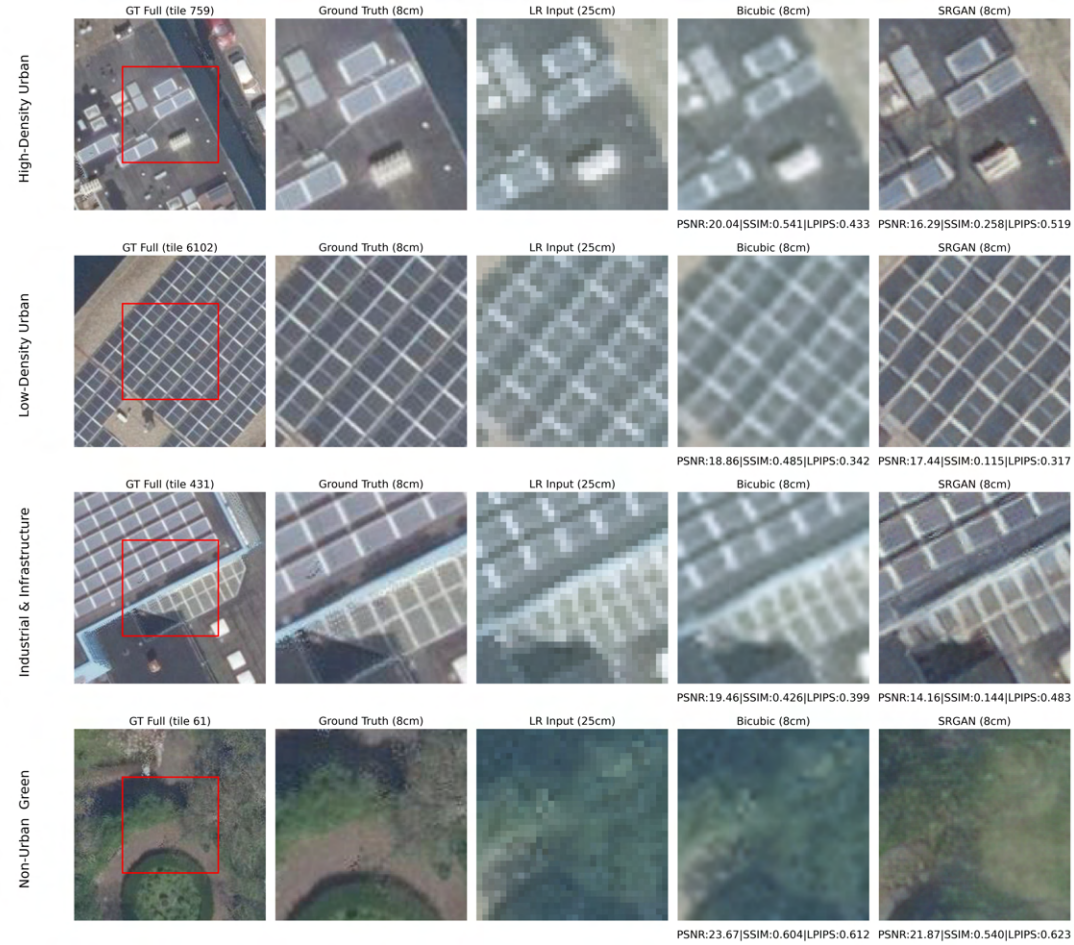

Figure 6.4.: Visual comparison of category-wise results for Iteration 2, alongside PSNR, SSIM, and LPIPS scores.

For the high-density urban tile, SRGAN successfully recovers details absent from the low-resolution input and reconstructs fine textures that resemble the original high-resolution content. Although the PSNR score is slightly lower, this is largely due to an artifact in another region of the tile that disproportionately affects the metric. SSIM is also lower for SRGAN, likely because of subtle spatial shifts and structural inconsistencies that SSIM penalizes despite the visual result being coherent. LPIPS scores are comparable between SRGAN and Bicubic interpolation, suggesting similar perceptual quality. Visually, however, Bicubic output appears blurred and merges adjacent rooftop features, whereas SRGAN preserves sharper building outlines and reconstructs distinct structures with greater clarity.

In the low-density urban tile, the input depicts a building with a complex roof structure and numerous solar panels. SRGAN manages to reconstruct these shapes with high fidelity, accurately preserving geometries that are difficult to discern in the LR input. The solar panels are clearly defined in the SRGAN output, with both edges and internal divisions modeled with sharpness on each side of the geometry. PSNR decreases slightly compared to Bicubic interpolation, and SSIM remains lower due to small inconsistencies in structure continuity. Nonetheless, LPIPS scores are comparable, with SRGAN showing better percep-

tual similarity to the ground truth in qualitative terms. Additionally, the SRGAN output exhibits a cooler, winter-like tone more aligned with the ground truth, while the LR input retains a green, summer-like appearance—highlighting SRGAN's ability to adapt across seasonal variations.

This difference arises because SRGAN learns to reconstruct images based on patterns observed during training. Its outputs are generated to match an ideal high-resolution distribution, which includes seasonal tones, object textures, and lighting characteristics. Bicubic interpolation, on the other hand, simply resamples the input pixel values without reference to a learned goal. As a result, it preserves the original color distribution and cannot shift toward ground truth conditions, even if they differ.

For the industrial and infrastructure tile, which contains a variety of textures such as solar panels, gravel, rooftop vegetation, and structural elements, SRGAN again shows a marked improvement. Bicubic results tend to blur these features and preserve the color characteristics of the LR input, while SRGAN more closely approximates the HR tile in both color and geometry. Roof structures are delineated with fine lines, and SRGAN accurately reconstructs sharp transitions between surfaces. Tree colors remain consistent across datasets, avoiding color hallucination. Although Bicubic yields slightly higher PSNR and SSIM values, the SRGAN output offers improved perceptual clarity and spatial coherence, which are more relevant for interpretation. Notably, the SRGAN reconstruction preserves the position and orientation of shadows from the LR input, producing shadows in the HR output that closely resemble those in the reference tile—demonstrating strong spatial adaptation in complex lighting conditions.

Finally, the urban green tile presents an interesting case. This tile contains both vegetation and a road. Seasonal differences are clearly visible: the ground truth trees exhibit a more reddish tone, while the LR input trees are greener. This confuses the SRGAN model to some extent, particularly for pixel-wise color reconstruction, leading to lower performance in all metrics. Nonetheless, the road passing through the vegetation is sharply reconstructed, indicating SRGAN's ability to preserve structural elements even when color domain shifts occur.

Overall, these examples demonstrate that SRGAN, trained with LR-HR tile pairs captured during different seasons, can successfully reconstruct unseen 8 cm HR tiles based solely on LR inputs at 25 cm. The model not only learns to reconstruct the spatial content (such as building contours and fine details) but also captures important relationships between pixel colors influenced by seasonal variations. This ability to generalize beyond direct supervision highlights the strength of the approach in handling real-world aerial imagery where exact temporal alignment is not possible.

6.3.2. Adaptability to New Geographic Areas

In this section, the adaptability of the Iteration 2 model to unseen geographic areas is evaluated. Only the Iteration 2 model is considered for this analysis, as the primary research focus is on assessing the model's performance when processing real low-resolution (LR) inputs, rather than synthetically degraded ones. For the purpose of this evaluation, two new areas were deliberately selected: The Hague and Zwolle. These locations were chosen to test the model's inherent ability to generalize to different cities without requiring site-specific retraining. Evaluating this generalization capability is crucial for understanding the model's viability in real-world applications, particularly in scenarios where obtaining large-scale, perfectly matched local training data might be constrained by computational resources, storage limitations, or data acquisition challenges. By testing in both a geographically closer area to the training data (The Hague) and a more distant one (Zwolle), the

evaluation assesses how variations in the environment, seasonality of imagery capture, and specific imaging conditions in different regions impact the model's super-resolution performance. The evaluation metrics for adaptability testing in The Hague and Zwolle are presented in Table 6.4.

Table 6.4.: Adaptability Evaluation Metrics for Zwolle and The Hague. Higher PSNR and SSIM values indicate better reconstruction quality, while lower LPIPS values indicate higher perceptual similarity to the ground truth.

Category	**SRGAN**	**Bicubic**	**SRGAN**	**Bicubic**	**SRGAN**	**Bicubic**
	PSNR		**SSIM**		**LPIPS**	
Zwolle						
High-Density Urban	15.6968	**16.7157**	0.3818	**0.5154**	0.6639	**0.5789**
Low-Density Urban	16.7844	**17.6015**	0.3849	**0.5187**	0.6670	**0.5959**
Industrial & Infrastructure	15.8452	**16.5261**	0.4083	**0.5164**	0.6609	**0.6136**
Non-Urban Green	18.5967	**18.7134**	0.4024	**0.5100**	0.6603	**0.6423**
The Hague						
High-Density Urban	18.3550	**19.7283**	0.4740	**0.6274**	0.6064	**0.5525**
Low-Density Urban	18.8079	**19.9894**	0.5185	**0.6505**	0.5991	**0.5598**
Industrial & Infrastructure	19.3459	**21.0076**	0.5630	**0.6908**	0.5909	**0.5549**
Non-Urban Green	19.5732	**21.6849**	0.6369	**0.7573**	0.5573	**0.4760**

The Hague, being geographically closer to the training areas used in Iteration 2 (Rotterdam and Delft), serves as a suitable test case for evaluating performance in a semi-familiar environment. Although the quantitative metrics presented in Table 6.4 indicate that Bicubic interpolation generally outperforms the SRGAN model across PSNR, SSIM, and LPIPS for The Hague, it is important to interpret these results with caution. PSNR, which heavily penalizes luminance differences (calculated on the Y channel), is particularly sensitive to the slight shifts in brightness and contrast that SRGAN outputs sometimes exhibit when reconstructing images based on learned patterns from training data, especially when dealing with different seasonal and lighting conditions than the ground truth. Qualitatively, as shown in the visual examples in Figure 6.5, the SRGAN outputs for The Hague demonstrate better building contours and rooftop detail compared to Bicubic interpolation, despite the metric scores suggesting otherwise. The lower PSNR can be understood as a consequence of SRGAN actively reconstructing pixel values based on complex learned relationships, a process distinct from Bicubic's straightforward spatial interpolation. Specifically, in High-Density Urban areas within The Hague, SRGAN successfully recovers clearer roof structures, though it may struggle slightly with very fine rooftop artifacts. In Low-Density Urban areas, SRGAN outputs sharper building outlines, occasionally showing some mixed pixel colors in areas where the model may lack strong prior knowledge from the training set for those specific textures or features. For Industrial areas, SRGAN effectively reconstructs layered rooftops and handles shadows better than Bicubic. However, in Non-Urban Green areas, seasonal differences, such as variations in leaf colors and shadow patterns between the LR input and the HR ground truth, cause the SRGAN outputs to be less consistent with the ground truth, which contributes to lower metric scores in this category.

Figure 6.5.: Example of model performance on tiles from The Hague.

Zwolle presents a distinctly different scenario for evaluating generalization, as it is geographically further removed from the training region and its imagery was captured at slightly different altitudes and under varying lighting conditions. As is evident from the visual examples shown in Figure 6.6, the SRGAN's performance declines more noticeably in Zwolle compared to The Hague, a trend also reflected in the lower metric scores in Table 6.4. Observations from the Zwolle tiles reveal that in High-Density Urban areas, SRGAN struggles more to preserve clean building outlines, although it may still offer better sharpness in certain parts compared to Bicubic. In Low-Density Urban tiles from Zwolle, the model exhibits notable issues, such as the misinterpretation of tiled roofs as grassy surfaces, resulting in the appearance of greenish tones and blurring. For Industrial tiles, SRGAN manages to reconstruct large rooftop structures and glass facades with reasonable clarity, even when working with noisy LR inputs. In Non-Urban Green tiles, vegetation generally appears relatively smooth, although the model sometimes hallucinates textures that are not present in the ground truth.

Figure 6.6.: Example of model performance on tiles from Zwolle.

Throughout these adaptability experiments in both The Hague and Zwolle, a clear pattern emerges regarding the behavior of the SRGAN model compared to Bicubic interpolation. SRGAN learns to reconstruct textures and spatial relationships based on the prior examples it was trained on, a process fundamentally different from Bicubic's method of simply upscaling pixel values through smooth estimation. This more complex, learning-based behavior allows SRGAN to produce outputs that are often visually clearer and contain more detail than Bicubic. However, it also makes SRGAN susceptible to penalties from pixel-wise metrics like PSNR and structural metrics like SSIM when the reconstructed structures or textures, while visually plausible, exhibit slight deviations in pixel intensity, color, or precise shape from the ground truth data for unseen areas. SRGAN generates high-resolution outputs by inferring plausible details based on learned patterns, which may not match the ground truth exactly at the pixel level. As a result, even when the output appears sharper and perceptually closer to a realistic high-resolution image, traditional metrics may still assign lower scores.

Examples of failure cases observed in The Hague further illustrate some of the challenges the model faces when generalizing. Failure examples from The Hague are shown in Figures 6.7 and 6.8. For instance, in Figure 6.7, a rooftop is reconstructed with an unintended

greenish tone, which is likely a result of the model confusing roof material textures with surrounding vegetation patterns present in the training data or similar textures in the input. Figure 6.8 shows an example where solar panels lose their distinct blue tone and merge indistinctly into the building's surface, highlighting how the model can sometimes struggle to preserve fine details or specific color characteristics when it encounters features for which it lacks sufficiently diverse or representative examples in the training data.

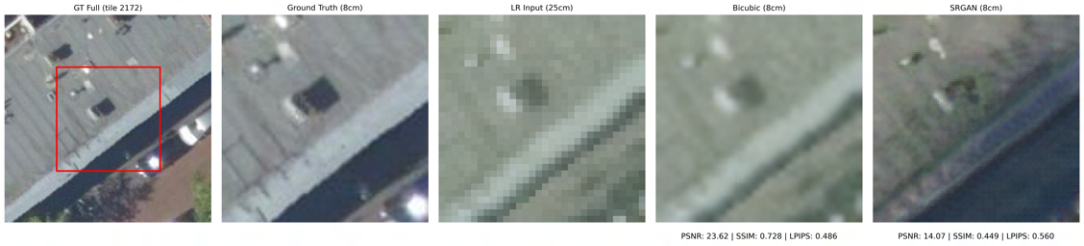

Figure 6.7.: Failure case: Roof reconstructed with greenish tones in The Hague.

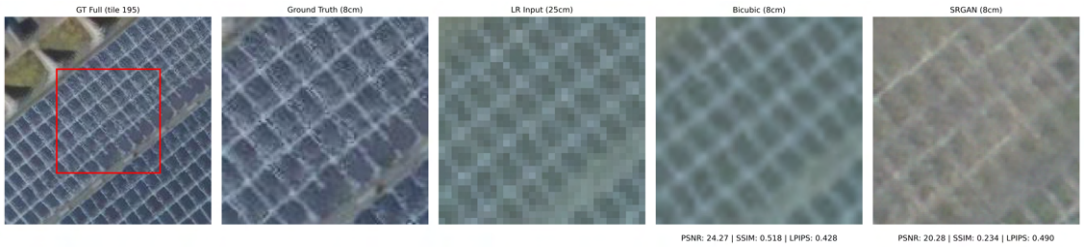

Figure 6.8.: Failure case: Solar panels misinterpreted as merged roof textures in The Hague.

Additional category-wise tile comparisons for The Hague and Zwolle are provided in Appendix A.2.

6.3.3. Failure Cases

This section presents examples of failure cases to highlight that the super-resolved tiles were not always satisfactory. Problems often occurred when the algorithm failed to accurately reconstruct the super-resolved output, leading to artifacts such as aliasing, ghosting, blurring, and glitches, which degraded the overall quality of the reconstructed images. These failure cases provide valuable insights into the limitations of the model, particularly when facing challenges related to high-frequency textures, seasonal variations, and structural changes between the low-resolution and high-resolution image pairs. Separate discussions are provided for Iteration 1 and Iteration 2 to illustrate the specific types of errors encountered in each phase.

Iteration 1

Figure 6.9 shows a tile containing a roof with solar panels. Although the low-resolution input tile at 25 cm is relatively clear and the bicubic interpolation manages to upscale it reasonably well, the SRGAN implementation struggles to reconstruct a usable output. A

large patch-like artifact is observed, creating a ghosting effect that severely impacts the visual quality. The texture of the artifact resembles grass, commonly seen in non-urban green tiles. This confusion likely arises when high-frequency patterns interact poorly with the upsampling process. Increasing the diversity and quantity of training samples could mitigate such issues by enabling the model to distinguish more clearly between different textures and structures. This inability to reconstruct a realistic output is also reflected in the evaluation metrics for this tile, where PSNR, SSIM, and LPIPS all show very low performance compared to other examples.

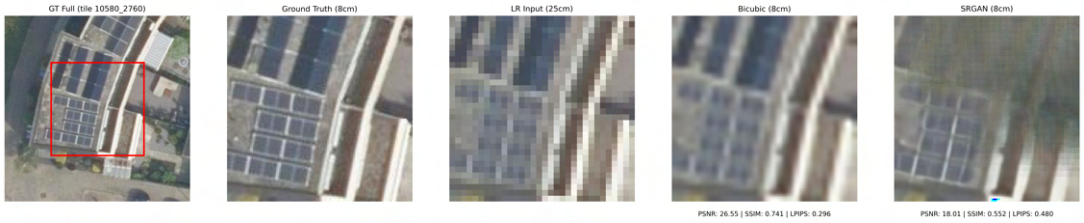

Figure 6.9.: Failure case for Iteration 1, showing a ghosting artifact on a solar panel roof.

Iteration 2

Due to the nature of Iteration 2, where the model was required to reconstruct outputs based on limited knowledge, failure cases were more frequent and more interesting to interpret.

Figure 6.10 illustrates an issue previously discussed: when the model had to deal with LR and HR tiles that differed significantly in color distribution due to seasonal changes, it often struggled to maintain consistency. As a result, the reconstructed output exhibits a merging of different vegetation colors, causing a loss of structural clarity.

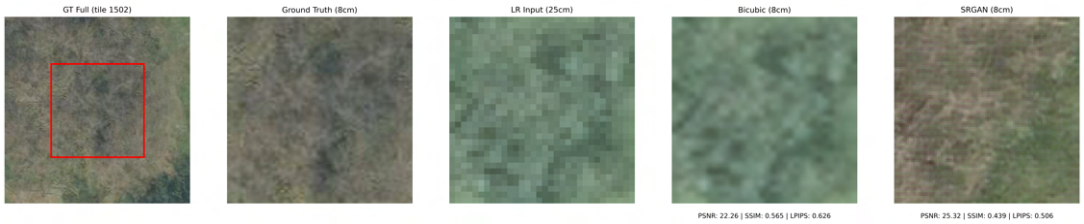

Figure 6.10.: Failure case showing vegetation color blending due to seasonal variation.

Seasonal change not only affected vegetation but also altered the appearance of man-made structures. Because the LR and HR images were captured at different times, some building structures had changed between acquisitions. This created LR-HR tile pairs that, despite spatial alignment, no longer accurately represented the same ground truth. As a result, the SR model was tasked with reconstructing structures that no longer existed or had significantly changed. In Figure 6.11, the SR output attempts to mimic the building structure present in the LR input but generates vague textures resembling vegetation instead. Such issues are inherent to temporal misalignment and could only be minimized by using images captured closer in time.

Figure 6.11.: Failure case caused by temporal changes in building structures between LR and HR images.

When using generative algorithms such as GANs, artifacts related to unseen patterns are common. In Figure 6.12, a swimming pool is present in the input, but because the model had limited exposure to such features during training, it fails to reconstruct the pool correctly. Instead, it outputs a black area with random noise. This type of failure could be addressed by expanding the dataset to include more examples of such features.

Figure 6.12.: Failure case where the model fails to reconstruct a swimming pool due to lack of similar training samples.

Finally, another common issue observed in Iteration 2 was the confusion between highly textured roofs and grass surfaces. As shown in Figure 6.13, the model occasionally reconstructed tiled roofs with textures resembling grass. This confusion is likely caused by the high-frequency patterns shared between tiled roofs and vegetation surfaces. While increasing the number of training samples would help mitigate this problem, further improvements could also be achieved by optimizing the generator architecture to better handle high-frequency details.

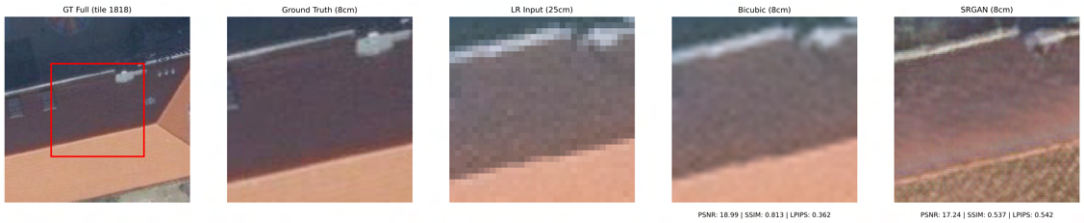

Figure 6.13.: Failure case where rich roof textures are confused with vegetation surfaces.

6.4. Impact on Downstream Tasks

After presenting the quantitative and qualitative results, this chapter focuses on the primary research goal: to assess whether the super-resolved outputs can be effectively utilized in downstream tasks. In particular, the Segment Anything Model (SAM) was applied to both Bicubic and SRGAN outputs, and the resulting segmentation masks were compared side by side.

It is important to note that SAM generates masks by identifying regions it deems segmentable based on its architecture and internal representations. It is not a dedicated detection algorithm trained specifically for objects such as solar panels or buildings. Therefore, it should not be viewed as a specialized detector, but rather as a proxy to evaluate how much useful structural and textural information the super-resolved images preserve. If SAM can successfully segment fine structures from the outputs, it suggests that SRGAN is not only enhancing perceptual quality but also creating outputs suitable for integration into broader geospatial analysis pipelines.

In addition to SAM, the proprietary segmentation model developed by Readar B.V. was also applied to the outputs. This model is trained to detect buildings, solar panels, and dormers, and incorporates a digital surface model (DSM) to improve building footprint delineation. To enable this, a full 1 km $\times$ 1 km test area in Rotterdam was prepared by first generating a virtual raster from the 256 px inference tiles using `gdalbuildvrt`, followed by stitching into a single georeferenced image using `gdal_translate` [GDAL, 2024]. This composite tile was created from tiles that were part of the SRGAN test set and includes a representative urban sample.

6.4.1. Segment Anything Evaluation

The Segment Anything Model (SAM) was applied to the outputs of both Iteration 1 and Iteration 2. The goal was to assess whether the super-resolved outputs not only improve perceptual quality but also enhance the ability to perform downstream tasks, such as object segmentation. By analyzing how SAM responds to the different outputs, it becomes possible to evaluate if the reconstructed images retain enough structural and textural information to support further automated processing beyond simple visual inspection.

To complement the visual analysis, Intersection-over-Union (IoU) was also calculated between the binary masks generated by SAM on the SRGAN and Bicubic outputs versus the ground truth HR tile. This additional metric was computed specifically for the example tiles shown in the figures that follow. However, due to the computational cost of SAM, it was not feasible to extend the IoU calculation to larger spatial areas.

While the number of masks detected by SAM is included in the visualizations for reference, it is **not used as a formal evaluation metric**. A higher mask count does not inherently reflect better performance. Instead, it is shown as a qualitative cue that complements the visual comparison—highlighting whether finer object boundaries and structures are preserved or lost.

Iteration 1

In Figure 6.14, the segmentation masks generated by SAM are shown for different inputs. From left to right, the ground truth 8 cm tile, the synthetic low-resolution input at 32 cm, the bicubic upsampling output at 8 cm, and the SRGAN super-resolved output at 8 cm are presented. The HR ground truth tile produced 66 masks, successfully capturing almost all fine details and distinct objects within the scene. In contrast, the LR 32 cm input resulted in

only 29 masks, missing several structures and often merging distinct elements into a single object.

The bicubic upsampling produced 30 masks, showing slight improvement over the LR input but still suffering from over-smoothing and object merging. The SRGAN output generated 54 masks, significantly outperforming the bicubic output. SRGAN not only recovered a greater number of individual structures but also successfully separated objects that bicubic interpolation failed to distinguish — for example, individual solar panels on rooftops were segmented separately rather than being merged into a single object.

These results suggest that the increased clarity and fine detail reconstruction provided by SRGAN directly benefits models like Segment Anything, enabling them to detect and segment scene components more effectively.

Figure 6.14.: Segmentation masks generated by SAM for Iteration 1 outputs: HR tile (8 cm), synthetic LR tile (32 cm), Bicubic upsampling, and SRGAN output.

These visual findings are supported by IoU calculations: the intersection between the SRGAN and HR masks reached 0.51, while the intersection between Bicubic and HR was only 0.30. This quantifies the visible improvement in object-level alignment when using SRGAN outputs.

Another example is shown in Figure 6.15, where SAM performs almost equally well on the ground truth HR tile and the SRGAN output, generating 71 and 54 masks, respectively. However, the evaluation is not solely based on the number of masks, as that would not be an objective criterion. Instead, it is also important to assess the quality of the segmentation, particularly the ability to individually detect all the small solar panels present on the rooftop. The SRGAN output successfully preserves enough detail to allow SAM to segment these fine structures separately, which is a strong indication of improved perceptual quality compared to the bicubic upsampled result.

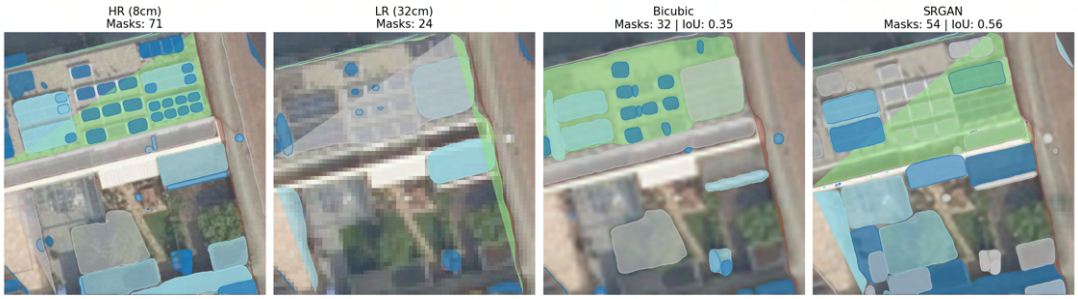

Figure 6.15.: Segmentation masks generated by SAM for Iteration 1 outputs: HR tile (8 cm), synthetic LR tile (32 cm), Bicubic upsampling, and SRGAN output.

In this example as well, SRGAN achieved a higher IoU of 0.56 compared to 0.35 for Bicubic. This reinforces the conclusion that SRGAN preserves object boundaries more consistently in a way that matches the HR reference.

Iteration 2

A tile containing multiple solar panels was selected to further evaluate the ability of SAM to perform on the super-resolved outputs, compared to the bicubic upsampled results. As shown in Figure 6.16, SAM was able to detect almost every solar panel in the SRGAN output, generating a total of 79 masks, compared to 125 masks on the ground truth tile. In contrast, the bicubic upsampled tile yielded only 21 masks. This highlights that, although bicubic upsampling can achieve reasonable results in terms of some quantitative metrics, its blurry output limits its effectiveness when applied to downstream tasks such as segmentation. In contrast, the sharper and more detailed SRGAN outputs significantly enhance the performance of segmentation models like SAM.

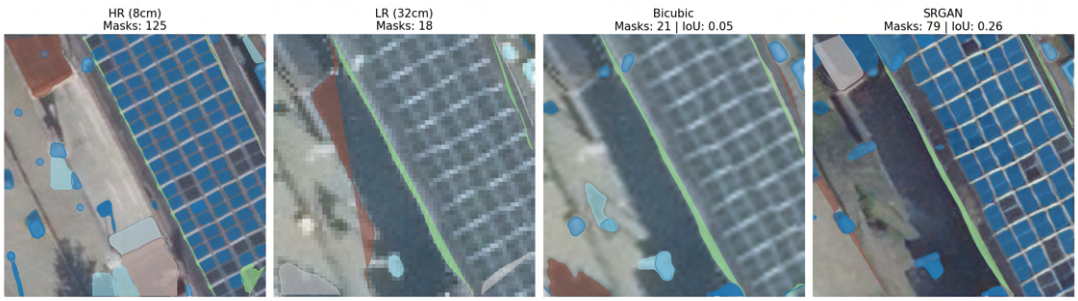

Figure 6.16.: Segmentation masks generated by SAM for Iteration 2 outputs: HR tile (8 cm), real LR tile (25 cm), Bicubic upsampling, and SRGAN output.

The IoU for this example further illustrates the difference: SRGAN achieved an IoU of 0.26 with the HR reference masks, compared to just 0.05 for Bicubic. Despite the larger visual gap due to seasonal misalignment, SRGAN provides a better structural fit to the ground truth.

Another example of a tile containing solar panels can be seen in Figure 6.17. Once again, it is evident that SRGAN is capable of producing outputs that are sufficiently clear and ready

for use as inputs to object detection pipelines. Rich details, clear building contours, and structural elements are precisely reconstructed, resulting in 89 masks generated from the SRGAN output, compared to only 59 masks generated from the bicubic upsampled tile.

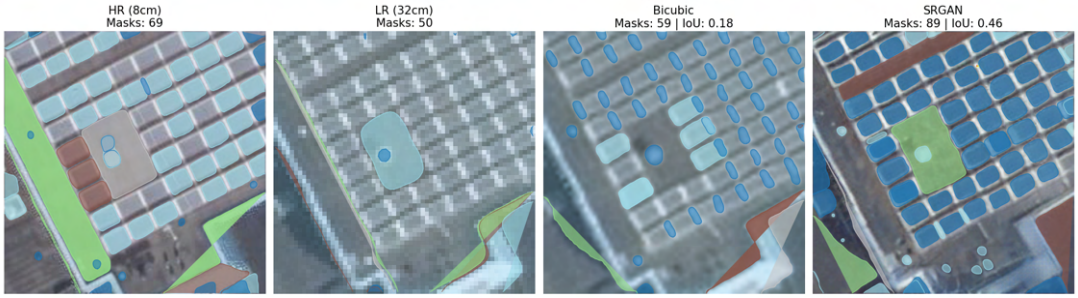

Figure 6.17.: Segmentation masks generated by SAM for a tile with solar panels: HR tile (8 cm), LR input (25 cm), Bicubic upsampling, and SRGAN output.

Similarly, the SRGAN masks reached an IoU of 0.46, while Bicubic only reached 0.18. This confirms that SRGAN reconstructions enable segmentation masks that more closely match the HR structure, even in complex urban scenes.

Adaptability to New Geographic Areas: SAM Evaluation

In addition to the quantitative evaluation, the adaptability of the SRGAN model to new areas was also assessed using the Segment Anything Model (SAM). Example tiles from The Hague and Zwolle were processed to evaluate how well the SR outputs could serve as inputs for segmentation tasks.

In Figure 6.18, SAM was applied to a tile from The Hague. From left to right, the ground truth 8 cm tile, the low-resolution 32 cm tile, the bicubic upsampling result, and the SRGAN output are shown. SAM detected 110 masks in the HR tile, but only 10 masks in the LR input, demonstrating how information loss at lower resolution impairs segmentation ability. For the SRGAN output, 66 masks were successfully generated, capturing most of the fine structures — including a series of solar panels on the rooftop — that the bicubic output failed to segment (only 10 masks). This illustrates that SRGAN reconstructions preserve sufficient detail and object boundaries to be beneficial for segmentation models like SAM.

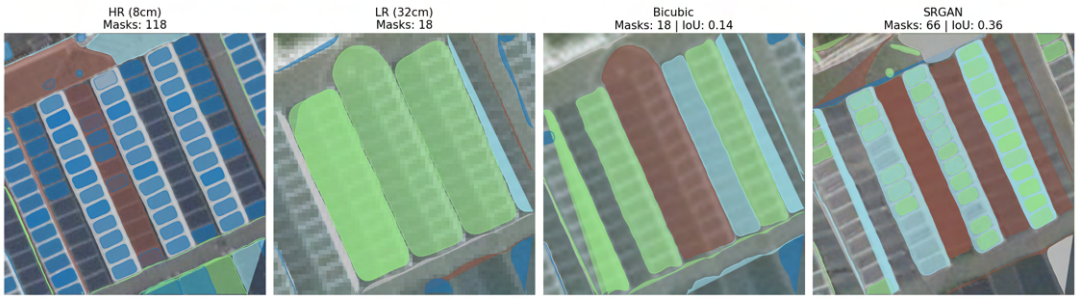

Figure 6.18.: Segmentation masks generated by SAM for The Hague tile: HR 8 cm tile, LR 32 cm tile, Bicubic upsampling, and SRGAN output.

Figure 6.19 presents example tiles from Zwolle for evaluation. In the HR 8 cm tile, SAM detected 70 masks, while the LR 32 cm input resulted in almost no meaningful segmentation. In contrast, the SRGAN output enabled SAM to generate 72 masks, successfully segmenting almost all objects present on the rooftops. The bicubic output performed poorly, producing only 32 noisy masks that did not correspond clearly to meaningful structures. Again, this highlights the SRGAN model's advantage in reconstructing images with enough perceptual quality and structural clarity to support downstream tasks even in unseen geographic areas.

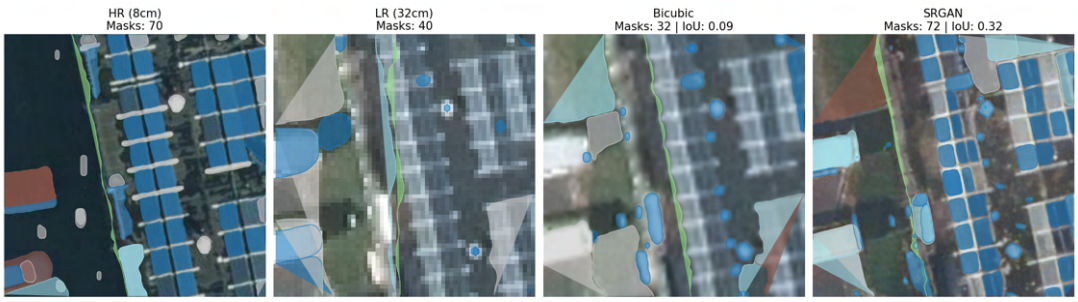

Figure 6.19.: Segmentation masks generated by SAM for Zwolle tile: HR 8 cm tile, LR 32 cm tile, Bicubic upsampling, and SRGAN output.

6.4.2. Semantic Segmentation (Readar B.V.)

The company's segmentation model was executed on four versions of the same spatial extent: (1) the SRGAN-enhanced 8 cm output, (2) the original low-resolution 25 cm input, (3) the high-resolution 8 cm reference image, and (4) the Bicubic-upsampled 8 cm image derived from the 25 cm input. This setup enabled comparative evaluation of how each representation supports segmentation in operational pipelines, particularly in detecting rooftop features and delineating class boundaries.

The segmentation model applied to the SRGAN and Bicubic outputs was trained exclusively on native 8 cm imagery. This means the model was not exposed to super-resolved or interpolated inputs during training. As a result, its performance on SRGAN and Bicubic tiles provides insight into how well these generated outputs align with the distribution of real high-resolution data.

In terms of height input, the model used DSM data matching the resolution of the RGB image: 8 cm DSMs were used with both the SRGAN and HR inputs, while 25 cm DSMs were used with the LR and Bicubic inputs. This ensured a fair comparison within each scale and emphasized the effects of spatial resolution on both texture and height-based segmentation.

The HR and LR segmentations were used as reference points to approximate the upper and lower bounds of expected performance, allowing the SRGAN and Bicubic results to be interpreted relative to these benchmarks. Although HR and LR segmentations were used internally as reference bounds, only the results from the SRGAN and Bicubic inputs are presented in the analysis below, as they reflect the outputs of interest for super-resolution evaluation.

The following table summarizes the detection performance across two structurally relevant object classes: PV panels and dormers. Precision, recall, and F1-score are reported for each class. In addition, Intersection over Union (IoU) is provided for the building footprint class.

Table 6.5.: Object Detection Performance per Class using SRGAN and Bicubic segmentation. Precision, recall, and F1 scores are calculated per object type. Building segmentation performance is reported separately using Intersection-over-Union (IoU).

Class	**Precision (%)**		**Recall (%)**		**F1 Score (%)**	
	SRGAN	Bicubic	SRGAN	Bicubic	SRGAN	Bicubic
PV Panel	**87.5**	20.0	**33.1**	0.8	**48.3**	1.5
Dormer	**97.2**	94.2	**76.9**	53.5	**85.9**	68.3
	Intersection over Union (IoU)					
	SRGAN			Bicubic		
Buildings	**0.9474**			0.9473		

The results show a clear advantage for SRGAN over bicubic interpolation. For PV panels, SRGAN achieves a precision of 87.5% and recall of 33.1%, compared to only 20.0% precision and 0.8% recall for bicubic. This indicates that SRGAN not only avoids false detections but also recovers more of the real panels present in the image. Dormer segmentation shows a similar trend: SRGAN reaches a recall of 76.9% and an F1-score of 85.9%, while bicubic lags behind at 53.5% recall and 68.3% F1. Although both methods achieve nearly identical IoU scores for buildings, this is largely because both pipelines use the same DSM input for building geometry—rather than due to image resolution. In contrast, the object-level metrics for PVs and dormers highlight SRGAN's superior ability to recover small-scale urban features essential for inventory or structural analysis.

To validate these metrics more concretely, additional evaluation was performed using manually placed ground truth points for PV panels and dormers. These points were used to identify true positives, false positives, and false negatives through spatial intersection. Each detected object was classified accordingly, and the resulting counts are visualized in the figure below.

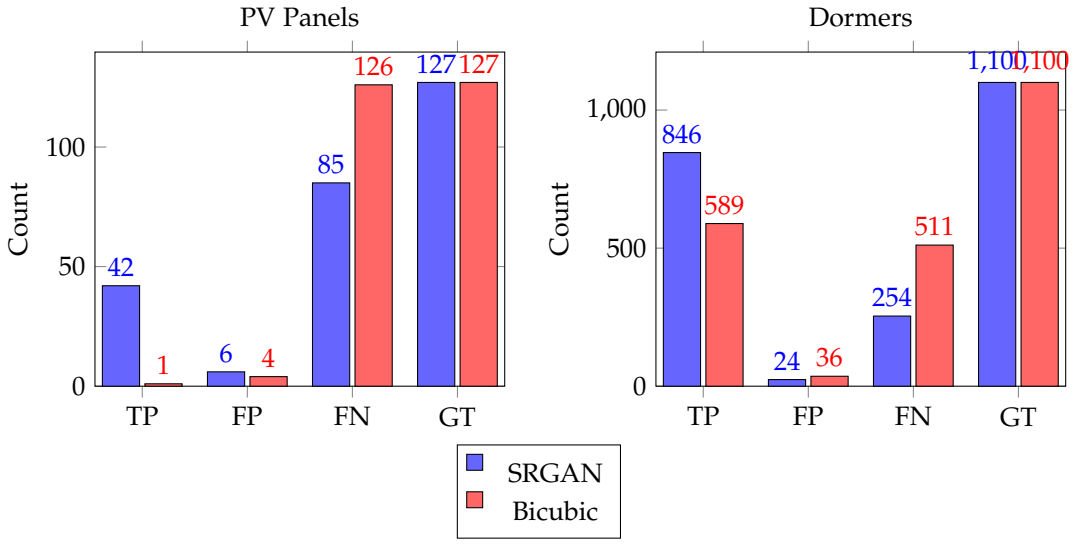

Figure 6.20.: Comparison of TP, FP, FN, and Ground Truth counts for SRGAN and Bicubic segmentations, for PV panels and dormers. Bar labels indicate absolute counts.

The manual validation confirms the trends observed in the model-based metrics. SRGAN consistently detects a higher number of true objects while limiting false positives. Particularly for PV panels, SRGAN recovers 42 of the 127 manually identified panels, while Bicubic only detects one. Dormer results are more balanced but still favor SRGAN in both precision and recall. These findings underscore SRGAN's advantage not only in metric scores but also in real-world usability, especially for downstream geospatial tasks where object-level detail matters.

Besides reporting standard metrics, it is also important to visually demonstrate how semantic segmentation performs on super-resolved versus bicubic-upsampled tiles. Figures 6.21 and 6.22 present a portion of the 1 km $\times$ 1 km test area in Rotterdam, with segmentation results overlaid on the imagery. Each map includes four zoomed regions at a larger scale, highlighting how the model performed on detecting individual object classes—PV panels, dormers, and buildings.

In Figure 6.21, the underlying imagery is the result of bicubic upsampling from the original 25 cm LR tiles. As shown in the bottom-left zoomed region, the segmentation model fails to capture the full extent of a solar panel installation, and in the top-left region, only a small portion of the panel is detected. For buildings, the bottom-right zoomed region illustrates a reasonably accurate footprint extraction, even for a complex roof structure. Dormer detection (top-right region) is partially successful but misses three clearly visible dormers on the same rooftop. Overall, the segmentation on bicubic-upsampled input severely underperforms for solar panels and dormers—an observation consistent with the low recall and F1 scores reported in the evaluation metrics. This suggests that bicubic output is not sufficiently detailed for reliable use in downstream tasks.

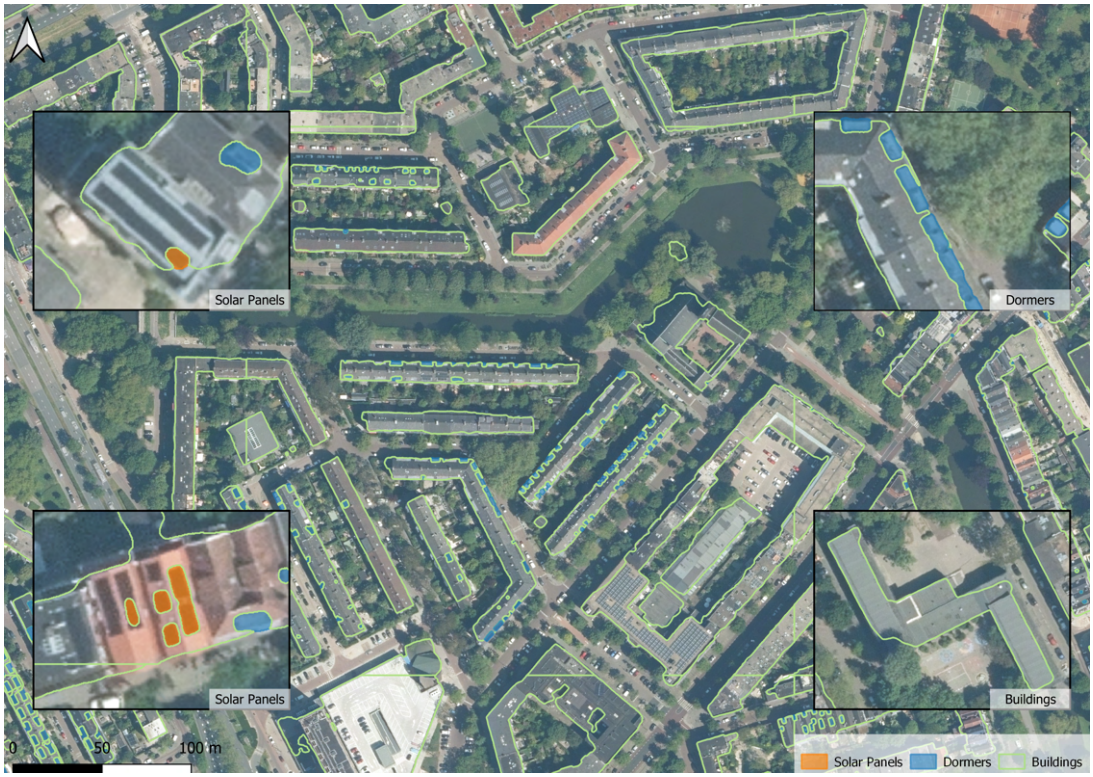

Figure 6.21.: Semantic segmentation results using bicubic-upsampled input. Zoomed regions highlight class-specific areas.

Figure 6.22 shows a significant improvement when the segmentation model is applied to the SRGAN-enhanced output. In the left zoomed regions, nearly all solar panels are correctly detected—even in complex, multi-row arrangements. The dormer region (top right) shows complete detection of all dormers on the roof. Building footprints remain as consistent and precise as those seen in the bicubic case. This result demonstrates that the SRGAN output not only enables the model to detect a much higher number of PV panels and dormers, but also reconstructs them with sufficient clarity and completeness to match their true spatial extent. In other words, super-resolution facilitates both recognition and full object recovery, improving segmentation usability for downstream geospatial analysis.

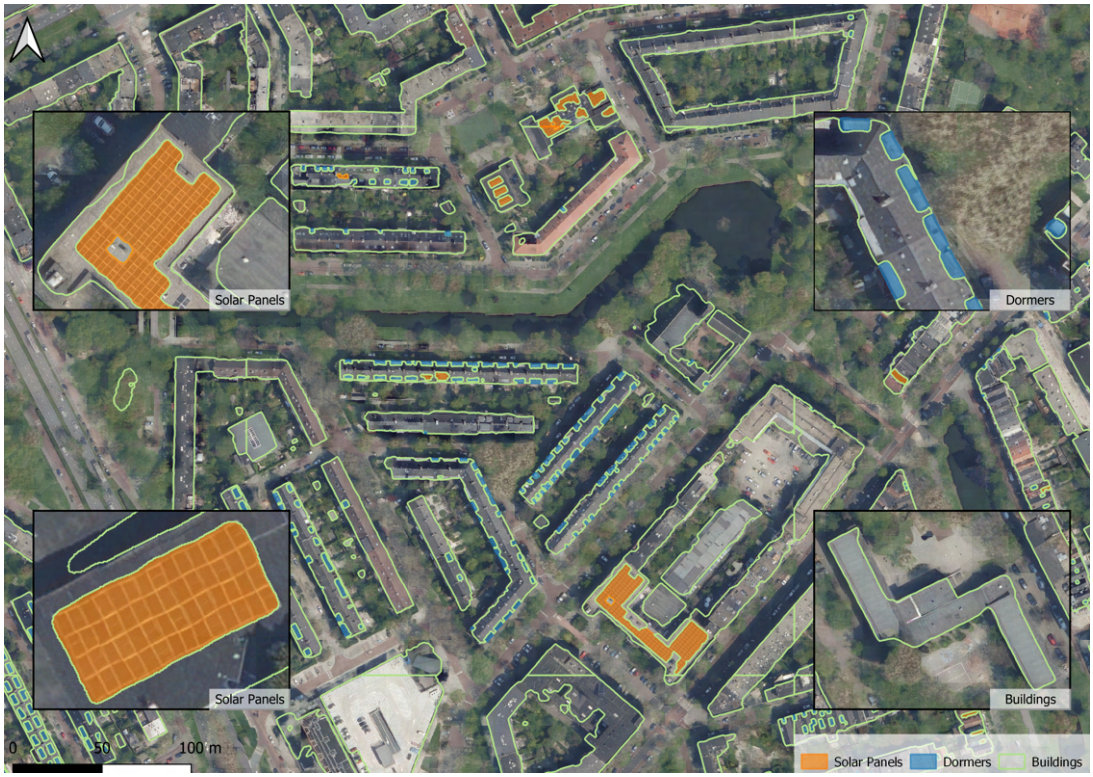

Figure 6.22.: Semantic segmentation results using SRGAN-enhanced input. Zoomed regions highlight class-specific areas.

The complete segmentation outputs over the full 1 km × 1 km test tile in Rotterdam can be seen in Appendix A.3, which includes the HR, LR, SRGAN, and Bicubic variants for full-scene comparison.

6.5. Discussion

This section reflects on the observed outcomes, interpreting the results from multiple perspectives—architectural design, land cover variation, geographic generalization, and downstream segmentation performance. The aim is to contextualize the quantitative metrics and visual assessments by linking them back to the model design choices, data characteristics, and intended application scenarios.

6.5.1. Impact of Edge-Aware Technique

The introduction of the edge-mask refinement block in the SRGAN architecture was aimed at enhancing the reconstruction of fine structural details, particularly building outlines and rooftop features. Visual inspections across multiple categories, as illustrated in Figure 6.1 and Figure 6.2 confirmed that the edge-aware SRGAN consistently produced sharper and more orthogonally aligned structures compared to the baseline SRGAN. Overall PSNR values across the dataset were higher for the edge-aware model compared to the baseline. The selected example was deliberately chosen to emphasize the improved reconstruction of fine

edge structures, where the edge-aware SRGAN demonstrated clearer roof contours and finer structural delineations. These results indicate that the proposed architectural modification was effective in enhancing geometric clarity without compromising overall image quality, particularly in urban environments characterized by strong man-made patterns.

6.5.2. Performance Across Different Land Cover Categories

The super-resolution performance varied depending on the type of land cover. In both Iteration 1 and Iteration 2, the SRGAN model showed superior results in urban categories specifically High-Density Urban and Industrial & Infrastructure areas—where geometric regularity and texture patterns supported effective reconstruction. For these classes, SRGAN often achieved competitive or superior LPIPS scores and visual quality compared to Bicubic interpolation, even when PSNR scores were slightly lower.

It is important to note that some tiles categorized as urban, particularly in the High-Density Urban class, contained not only buildings but also peripheral areas of vegetation, such as trees surrounding the structures. In these cases, even when the SRGAN model accurately reconstructed the building geometries with high visual fidelity, the presence of trees introduced greater ambiguity in the low-resolution inputs. As a result, small inconsistencies or hallucinations in these ambiguous areas could negatively affect the overall PSNR for the tile, despite the successful reconstruction of the primary man-made structures.

In contrast to tiles characterized by more urban settings, Non-Urban Green areas, consisting primarily of vegetation and irregular surfaces, posed a greater challenge. The absence of repetitive geometric structures and the higher seasonal variability of natural elements made it significantly more difficult for the model to accurately reconstruct these regions. Consequently, both PSNR and perceptual quality gains were less pronounced for non-urban tiles compared to their urban counterparts. These observations indicate that the model performs best when distinct, regular structural cues are present in the low-resolution input.

6.5.3. Generalization Across Cities

Generalization experiments conducted on The Hague and Zwolle confirmed that the SRGAN model was able to transfer knowledge beyond the training areas, though with varying success. The results suggested that geographic proximity alone was not the sole determining factor for performance. Instead, generalization was more strongly affected by the similarity in land cover types and spatial patterns—such as the arrangement of buildings, roads, and vegetation—between the test and training tiles.

For The Hague, where tile characteristics more closely matched the training data, SRGAN outputs maintained good visual quality, albeit with some drop in quantitative metrics such as PSNR and SSIM due to seasonal and acquisition differences. In Zwolle, where imagery conditions diverged more significantly, the model encountered greater difficulty, leading to lower reconstruction fidelity, particularly in low-density residential and green areas. Nevertheless, SRGAN consistently produced sharper and more visually coherent outputs than Bicubic interpolation, reinforcing its potential for generalization when the input domain remains reasonably close to the training distribution.

6.5.4. Downstream Task Performance

Evaluation of the super-resolved outputs using the Segment Anything Model (SAM) demonstrated that SRGAN outputs were significantly better suited for segmentation tasks than Bicubic-upsampled counterparts. Across both Iteration 1 and Iteration 2, SRGAN-enhanced

images enabled SAM to generate a higher number of meaningful masks, recovering fine structures such as individual solar panels, building boundaries, and road networks that were otherwise lost in the Bicubic results. These findings confirm that super-resolution using SRGAN not only improves perceptual quality but also enhances the functional usability of the outputs for subsequent geospatial analysis tasks.

A reason why SRGAN outperforms Bicubic in detecting objects like PV panels lies in its ability to perform domain transfer. While Bicubic interpolation passively upsamples image resolution through pixel resampling, SRGAN learns how to reconstruct high-resolution structure and texture from degraded inputs. This learned transformation enables SRGAN to generate plausible high-frequency details that are entirely missing from the LR input. As a result, SRGAN is not only able to increase perceptual sharpness but also recover objects that are no longer visible in the original input.

This is especially relevant for PV panel detection, where fine structure and subtle edges are essential for successful segmentation. Because PV panels are thin, low-profile structures, they are almost invisible in DSM data and rely heavily on RGB cues. SRGAN enhances these cues by reconstructing edges and textures, while Bicubic fails due to its lack of learned structural priors.

In contrast, the near-identical IoU scores between SRGAN and Bicubic for building segmentation can be explained by the presence of DSM in the Readar segmentation model. Buildings are large, elevated structures with strong DSM signals, so segmentation performance does not depend on image clarity alone. Since DSM is used as an input to the model, both SRGAN and Bicubic produce similarly accurate results for buildings, regardless of visual quality. This effect highlights that for DSM-supported classes, image super-resolution may have less impact.

Importantly, SRGAN's ability to perform domain transfer means that retraining the downstream segmentation model is not required. The segmentation model trained on HR imagery can be directly applied to SRGAN outputs, despite the LR inputs being from a different season or acquisition context. This capability has significant implications for operational scalability: it allows super-resolved imagery to be used as a drop-in replacement for HR data in existing pipelines without additional model tuning.

6.6. Limitations and Artifact Analysis

While the proposed SRGAN-based approach demonstrated strong results in perceptual quality and downstream utility, a number of artifacts and methodological limitations were observed throughout the experiments. This section provides a structured analysis of the most common visual artifacts, contextual challenges related to data and domain differences, and broader limitations inherent to the chosen training strategy. Understanding these issues is essential for interpreting the model's behavior and identifying opportunities for future refinement.

6.6.1. Artifacts in Iteration 1: Ghosting and Grass-like Patterns

Iteration 1 was trained entirely on synthetic low-resolution (LR) images created through bicubic downsampling. While this provided perfectly aligned LR-HR pairs for stable pretraining, it also introduced limitations. Bicubic downsampling represents a simplified degradation process, lacking the complexity of real-world distortions such as atmospheric blur, sensor noise, and compression artifacts. As a result, the model primarily learned to reverse smooth interpolation effects, rather than generalizing to more complex degradations.

This constraint led to ghosting effects and grass-like patterns in Iteration 1 outputs. The SRGAN occasionally hallucinated artificial textures in areas with ambiguous information, especially in tiles containing high-frequency surfaces such as rooftops or vegetation. Such artifacts are characteristic of GAN-based models trained to prioritize perceptual realism, even at the risk of inventing plausible but incorrect details. The reliance on synthetic bicubic degradation thus limited the diversity of textures the model could reconstruct, highlighting the need for subsequent fine-tuning on real-world imagery.

6.6.2. Artifacts in Iteration 2: Greenish Roofs and Texture Errors

In Iteration 2, the model was fine-tuned using real LR images captured at 25 cm resolution. Although fine-tuning successfully adapted the generator to realistic degradation patterns, several artifacts persisted, notably a greenish tint on rooftop surfaces and occasional hallucination of vegetation-like textures.

A contributing factor was the transfer of bias from the pre-trained model of Iteration 1. Since the model had learned specific color distributions and textural priors from synthetic data, these biases influenced its behavior even after fine-tuning. When encountering less familiar textures in new geographic areas, particularly in The Hague and Zwolle, the generator sometimes incorrectly associated certain structures with vegetation-like colors.

Additionally, the differences in seasonal capture between the real LR inputs and HR reference images further exacerbated these problems. Variations in foliage, shadows, and lighting conditions introduced discrepancies that the model attempted to reconcile, occasionally leading to unnatural colorization or texture blending. These effects were more pronounced in non-urban green areas, where seasonal variability was higher and structural cues were weaker.

6.6.3. General Limitations of the SRGAN Approach for Aerial Imagery

Several intrinsic challenges limited the full potential of SRGAN for aerial image super-resolution:

- **Synthetic vs. Real Degradation Mismatch:** Training on bicubic-downscaled images could not fully simulate the diverse degradations present in real aerial imagery, limiting robustness.
- **Domain Shift:** Seasonal and environmental differences between LR and HR pairs, especially when temporally misaligned, introduced inconsistencies that adversely affected pixel-wise metrics like PSNR and SSIM.
- **Bias from Pre-trained Weights:** While transfer learning accelerated convergence, it also transferred learned biases, particularly in color and texture associations.
- **Challenges with Natural Surfaces:** Non-urban green areas, characterized by irregular textures and seasonal variability, remained more difficult for the model to reconstruct compared to urban settings with strong geometric regularities.
- **Tile Categorization Noise:** The tile categorization used during evaluation may introduce label noise. Some tiles assigned to urban classes contained green or open areas, which affected the reliability of category-level metrics.

- **Temporal Misalignment in Ground Truth:** Temporal differences between the real HR and LR tiles caused structural inconsistencies that complicated the evaluation of true performance, particularly for solar panels and vegetation.
- **Limited Experimental Scope:** Due to hardware and time constraints, training and evaluation were performed on full batches of large tile sets. This limited the exploration of additional architectural variants and fine-tuning configurations within the available timeframe.

Despite these limitations, the overall improvements observed in perceptual quality, structural coherence, and downstream segmentation performance validate the effectiveness of the two-stage training strategy employed.

6.6.4. Necessity of Pre-Training on Synthetic Data

Pre-training on synthetic LR/HR pairs was a critical step for achieving stable and efficient model convergence. Real LR images lacked perfect pixel alignment with the HR references due to environmental changes and acquisition differences, making direct supervised learning impractical. Iteration 1 allowed the model to learn fundamental upsampling behavior under controlled conditions, ensuring that the generator acquired essential super-resolution capabilities. Fine-tuning in Iteration 2 subsequently adapted these capabilities to realistic degradation patterns. Without synthetic pre-training, training stability would have been compromised, likely leading to poor generalization and slower convergence when learning from real-world LR inputs.

7. Conclusions

This thesis demonstrated the effectiveness of GAN-based super-resolution for enhancing aerial imagery from 25 cm to 8 cm resolution, enabling improved performance in downstream object detection tasks. The proposed two-phase training strategy—starting with synthetic HR-LR pairs and followed by fine-tuning on real, temporally misaligned imagery—proved essential for adaptingthe model to the degradation patterns observed in real aerial imagery.

The modified SRGAN architecture, which integrated edge-aware refinement and mask-guided filtering, significantly improved the reconstruction of fine-grained urban features such as rooftops, dormers, and solar panels. These architectural enhancements led to higher precision, recall, and F1-scores compared to traditional interpolation methods, particularly for small objects. Overall, the results confirm that domain-adapted, GAN-based super-resolution offers a promising pathway for operational use of low-resolution aerial datasets in urban analysis.

7.1. Research Questions

7.1.1. Main Research Question

To what extent can GAN-based super-resolution enhance 25 cm aerial imagery to 8 cm, ensuring its applicability for object detection tasks?

The study showed that GAN-based super-resolution was highly effective in enhancing aerial imagery from 25 cm to 8 cm resolution. The proposed method produces outputs with improved clarity, visual quality, and structural fidelity—successfully reconstructing fine features such as rooftop edges, solar panels, and dormers. These high-frequency details are particularly relevant for segmentation and object detection tasks.

Compared to traditional interpolation methods like bicubic upsampling,the GAN-based model consistently outperformed interpolation methods in both visual fidelity and segmentation performance. Moreover, by applying super-resolution to low-resolution imagery while using an existing segmentation model trained on 25 cm data, the study demonstrated that it is not necessary to retrain models on high-resolution inputs. The SRGAN output alone was sufficient to boost detection performance, confirming its applicability not only as a preprocessing step but also as a way to extend the use of existing models to higher-resolution domains. This makes the approach particularly suitable for operational pipelines that require consistent feature extraction from limited-resolution imagery.

7.1.2. Sub-Questions

Sub-question 1: How accurately can a GAN reconstruct 8 cm HR images from 25 cm LR aerial inputs, especially at edges and rooftop details?

The modified SRGAN accurately reconstructed 8 cm HR images from 25 cm LR inputs. This was confirmed by improvements across all key quantitative metrics (PSNR, SSIM, LPIPS), as well as visual inspection. Rooftop contours, panel boundaries, and building edges were preserved with high fidelity, indicating that the model captured the spatial relationships necessary to recover structural detail. Unlike interpolation methods, SRGAN generated outputs that retained geometric integrity and avoided blurring or distortion, showing a deep understanding of how urban structures should be represented at higher resolution.

Sub-question 2: How does temporal misalignment (e.g., winter HR vs. summer LR) affect GAN performance, and can domain adaptation mitigate these effects?

Temporal misalignment introduced seasonal differences in illumination, vegetation, and shading between the LR and HR tiles. The two-phase training strategy—consisting of pretraining on synthetic data and fine-tuning on real, temporally misaligned imagery—helped mitigate some of these effects. However, the model's performance remained highly dependent on the diversity of the training data. When encountering structures or configurations that were absent during training, the model occasionally produced artifacts or inaccurate color reconstructions. Despite these limitations, it was able to learn robust mappings within the RGB domain, effectively performing domain transfer between seasonal variations and appearance changes.

Sub-question 3: What are the limitations of GANs in preserving geometric fidelity (e.g., artifacts, hallucinations) for geospatial use cases?

Despite improvements in geometric preservation, limitations remain. In particular, areas with irregular textures—such as green or non-urban zones—were more susceptible to hallucinated details or ghosting artifacts. These issues were more pronounced in regions where the model lacked representative examples during training. Nevertheless, the inclusion of edge-aware modules significantly reduced such effects compared to baseline models. Experiments also showed that extending training (e.g., increasing epochs) helped suppress these artifacts further, indicating that some of the limitations are not fundamental but can be addressed through longer or more diverse training.

Sub-question 4: What metrics best assess SR image quality for downstream object detection tasks?

The study employed a combination of pixel-based (PSNR), structural (SSIM), and perceptual (LPIPS) metrics, along with functional benchmarks such as segmentation accuracy (precision, recall, and F1-score). While LPIPS and SSIM captured visual and structural improvements, segmentation metrics provided the most direct insight into the model's utility for downstream tasks. Importantly, performance was also evaluated per object class (e.g., PV panels, dormers), recognizing that not all metrics reflect the same aspects of quality. When comparing SRGAN with bicubic upsampling, it is also necessary to account for the fundamental difference in approach—generative vs. interpolation-based—which influences metric interpretation. Ultimately, no single metric is sufficient on its own; evaluation must be task-aware and weighted accordingly.

7.2. Future Work

This section outlines several directions for extending the research. These recommendations address limitations encountered during the project and propose architectural, data-driven,

and methodological improvements to enhance the super-resolution pipeline and its integration into geospatial workflows. The proposed directions are organized thematically.

7.2.1. Architectural and Loss Function Enhancements

- **Edge-aware loss functions:** While the EdgeMaskBlock improved the reconstruction of urban geometry, future work could incorporate explicit edge-aware loss terms into training. This would further encourage the network to prioritize edge fidelity, particularly along building contours and rooflines.
- **Mask-guided input refinement:** Preliminary experiments explored integrating binary building masks from the PDOK repository as a fourth input channel during training. The intuition behind this approach was to guide the model's attention toward semantically important regions—particularly building boundaries—by explicitly marking their locations in both LR and HR tiles. The mask was appended as a separate channel to the input image, resulting in a 4-channel input tensor.

  As shown in Figure 7.1, the binary masks were generated by rasterizing building polygons and aligning them with the tile grids. These masks were applied consistently across both high-resolution (HR) and low-resolution (LR) images to maintain spatial correspondence.

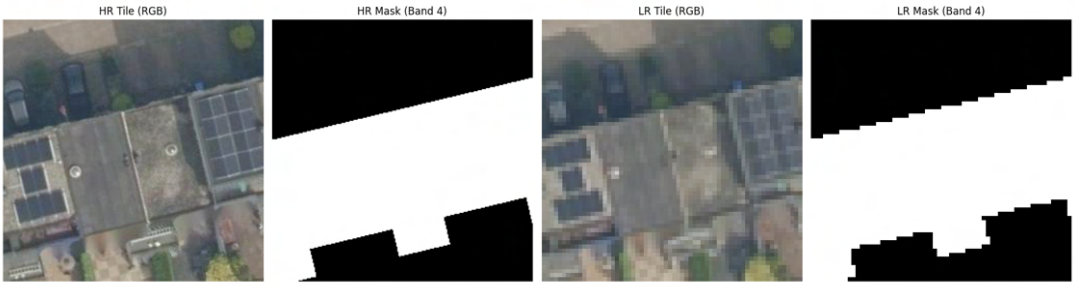

Figure 7.1.: Visualization of building mask integration into the SR input pipeline.

To test the effect of mask-guided input, the model was trained on the same synthetic LR–HR pairs used in Iteration 1, with all parameters kept constant except for the additional input channel. Figure 7.2 shows a side-by-side visual comparison of SRGAN outputs with and without mask input.

Figure 7.2.: Qualitative comparison between SRGAN with and without mask input.

Although the inclusion of the building mask was expected to enhance structural sharpness, the result was inconclusive. The perceptual difference was minimal, and PSNR slightly decreased. This suggests that simply concatenating a binary mask as a fourth channel was not sufficient to influence the generator's attention or internal representations meaningfully.

A likely reason for the limited effect is that the model did not have a mechanism to interpret or act upon the mask information effectively. Without explicit conditioning layers or a loss function that encourages edge preservation based on the mask, the extra channel may have been treated as a redundant input. Moreover, the building masks were binary and did not encode edge strength or uncertainty, which could have limited their utility for guiding finer-grained attention during feature learning.

Future iterations could explore more advanced conditioning strategies—such as spatial attention mechanisms, edge-aware feature fusion, or the use of soft masks derived from edge detectors. Additionally, including a mask-guided loss component during training could provide a more direct learning signal to enforce alignment between predicted and true structures in these regions.

7.2.2. Training Strategy and Dataset Improvements

- **Multi-city generalization:** Current models were trained on data from a single city, resulting in poor generalization to unseen areas like Zwolle. Training on a broader, geographically diverse dataset would improve robustness to variations in building types, land use, and environmental conditions.
- **Temporal-aware training:** The model does not account for differences in seasonality or acquisition time between HR and LR images. Incorporating temporal attention or multi-temporal data could improve the handling of foliage, shadows, and structural changes over time.
- **SR-assisted object detection:** Future work could optimize the model not only for perceptual quality but also for object detection performance. This could be achieved by linking the SR module directly to detection heads in a shared or end-to-end training pipeline, improving the relevance of outputs for tasks such as rooftop inventory or solar panel mapping.

7.2.3. Evaluation Pipeline and Methodological Refinements

- **Automated preprocessing pipeline:** The current workflow requires manual selection of cities, tile generation, and land-use filtering. Developing a fully automated preprocessing pipeline would increase reproducibility and scalability, enabling deployment across larger regions or by non-technical users.
- **Masking vegetation in metric evaluation:** Some quality metrics were biased by the presence of vegetation or green areas within the tiles. A possible improvement is to mask out green regions when computing metrics like PSNR and SSIM. However, this is non-trivial, as metrics rely on consistent pixel dimensions and introducing masked (e.g., black) pixels may skew results unless carefully handled.
- **Improved tile stitching methods:** Final segmentation outputs were generated after stitching super-resolved tiles into 1 km $\times$ 1 km mosaics using GDAL. While functional,

this stitching was simplified and did not account for potential alignment artifacts at tile borders. In operational pipelines, more accurate blending or edge-aware stitching methods could improve consistency—especially in areas near tile edges where segmentation performance sometimes degraded.

Although time and computational constraints limited the exploration of deeper architectural variants and broader hyperparameter tuning, the results demonstrate that perceptual super-resolution can meaningfully enhance aerial imagery. Future work should prioritize architectural refinement, generalization across time and geography, and tighter integration with task-specific applications.

A. Appendix

This appendix provides supporting visualizations referenced throughout the main chapters. It includes spatial maps of training and testing areas, full-sized super-resolved outputs, category-specific comparisons, and complete semantic segmentation results. These figures serve to complement the core results discussed in Chapter 6 and offer a more detailed view of the qualitative performance of each method.

A.1. Training and Testing Area Maps

Figure A.1.: Training and testing areas for Iteration 1. Green indicates training tiles and red indicates testing tiles. These locations correspond to the evaluation setup described in Chapter 6.

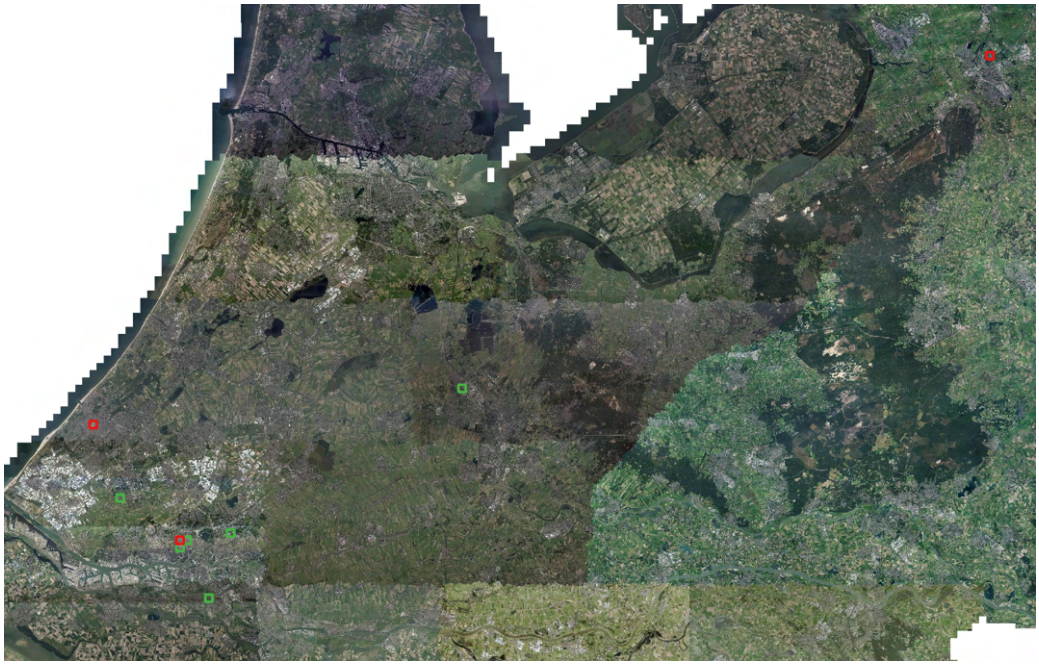

Figure A.2.: Training and testing areas for Iteration 2. Green shows training tiles and red indicates testing tiles. Generalization cities (e.g., Zwolle, The Hague) are labeled.

Figure A.3.: Location of the 1 km × 1 km test tile (Iteration 2) within the municipal bounds of Rotterdam. This is the same area analyzed in Chapter 6.

A.2. Full Super-Resolved Output and Category Comparisons

Figure A.4.: Full SRGAN output (Iteration 2) over the 1 km × 1 km test tile in Rotterdam. This complements the cropped views shown in Chapter 6.

Figure A.5.: Visual comparison of category-wise results for Iteration 1, alongside PSNR, SSIM, and LPIPS scores.

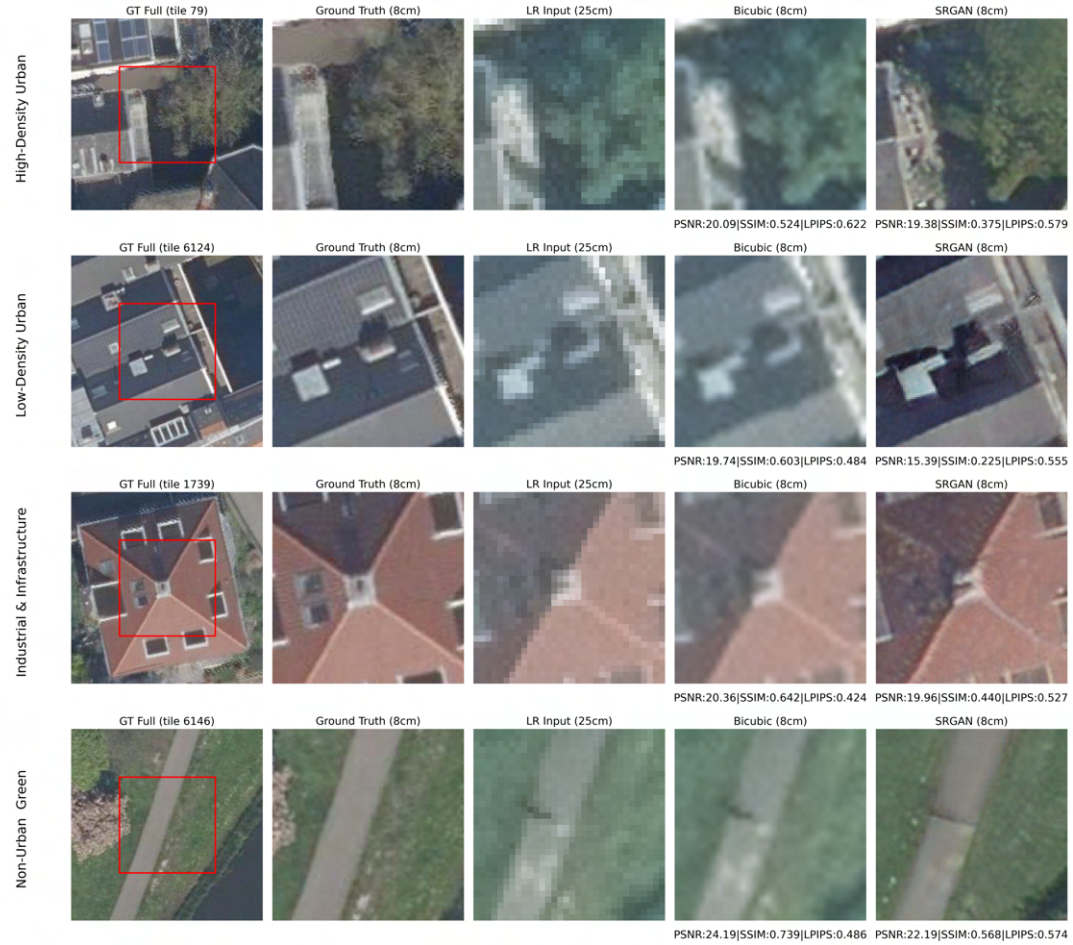

Figure A.6.: Visual comparison of category-wise results for Iteration 2, alongside PSNR, SSIM, and LPIPS scores.

Figure A.7.: Visual comparison of category-wise generalization results to The Hague (Iteration 2), alongside PSNR, SSIM, and LPIPS scores.

Figure A.8.: Visual comparison of category-wise generalization results to Zwolle (Iteration 2), alongside PSNR, SSIM, and LPIPS scores.

A.3. Complete Semantic Segmentation Outputs

Figure A.9.: Semantic segmentation result using the high-resolution (8 cm) input over the Rotterdam test tile. Full 1 km × 1 km output as referenced in Chapter 6.

Figure A.10.: Semantic segmentation result using the low-resolution (25 cm) input over the Rotterdam test tile. This represents the baseline scenario in Chapter 6.

Figure A.11.: Semantic segmentation result using the SRGAN-enhanced 8 cm imagery over the Rotterdam test tile. This is the main super-resolved input evaluated in Chapter 6.

Figure A.12.: Semantic segmentation result using the bicubic-upsampled input over the Rotterdam test tile. Compared alongside SRGAN in Chapter 6.

Bibliography

Agustsson, E. and Timofte, R. (2017). Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1122–1131.

Amhar, F., Jansa, J., and Ries, C. (1998). The generation of true orthophotos using a 3d building model in conjunction with a conventional dtm. *International Archives of Photogrammetry and Remote Sensing*, 32.

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.

Babu, K. K. and Dubey, S. R. (2021). Cdgan: Cyclic discriminative generative adversarial networks for image-to-image transformation.

Beeldmateriaal (2023). Beeldmateriaal: Aerial and satellite imagery services.

Bevilacqua, M., Roumy, A., Guillemot, C., and Alberi-Morel, M.-L. (2012). Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference*, pages 1–10.

Campbell, J. B. and Wynne, R. H. (2011). *Introduction to Remote Sensing, Fifth Edition*. Guilford Press.

Canny, J. F. (1983). A variational approach to edge detection. In *AAAI*, volume 1983, pages 54–58.

Conde, M. V., Choi, U.-J., Burchi, M., and Timofte, R. (2022). Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration.

Copernicus Land Monitoring Service (2024). Urban Atlas: European Land Use and Land Cover Data. Accessed: March 17, 2025.

Dai, D. and Yang, W. (2010). Satellite image classification via two-layer sparse coding with biased image representation. *IEEE Geoscience and Remote Sensing Letters*, 8(1):173–176.

Deshpande, A. and Patavardhan, P. P. (2019). Survey of super resolution techniques. *ICTACT Journal on Image & Video Processing*, 9(3).

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.

GDAL (2024). *GDAL/OGR Geospatial Data Abstraction Library: Version 3.x*. Open Source Geospatial Foundation. Available at `https://gdal.org/en/stable/`.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Google Developers (2023). Classification: Precision and recall. Accessed: 2024-05-09.

Gross, S. and Wilber, M. (2016). Training and investigating residual nets. http://torch.ch/blog/2016/02/04/resnets.html. Accessed: 2025-04-11.

Guo, M., Zhang, Z., Liu, H., and Huang, Y. (2022). NDSRGAN: A Novel Dense Generative Adversarial Network for Real Aerial Imagery Super-Resolution Reconstruction. *Remote Sensing*, 14:1574.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034.

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.

Huang, J.-B., Singh, A., and Ahuja, N. (2015). Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 448–456.

Jain, A., Sarkar, A., and Agrawal, A. P. (2023). Improved generative adversarial network for generating high-resolution images from low-resolution images. In *2023 13th International Conference on Cloud Computing, Data Science and Engineering (Confluence)*, pages 360–364.

Jebur, A., Abed, F. M., and Mohammed, M. (2017). *3D City Modelling by Photogrammetric Techniques*. PhD thesis.

Jiang, K., Wang, Z., Yi, P., Wang, G., Lu, T., and Jiang, J. (2019). Edge-Enhanced GAN for Remote Sensing Image Superresolution. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5799–5812. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

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.

Keys, R. (1981). Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(6):1153–1160.

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 *IEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 1063-6919.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. (2023). Segment anything.

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].

Lee, O.-Y., Shin, Y.-H., and Kim, J.-O. (2019). Multi-perspective discriminators based generative adversarial network for image super resolution. *IEEE Access*, PP:1–1.

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.

Li, X. and Orchard, M. (2001). New edge-directed interpolation. *IEEE Transactions on Image Processing*, 10(10):1521–1527.

Lu, Z., Li, J., Liu, H., Huang, C., Zhang, L., and Zeng, T. (2022). Transformer for single image super-resolution.

Mao, Q., Wang, S., Wang, S., Zhang, X., and Ma, S. (2018). Enhanced image decoding via edge-preserving generative adversarial networks. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the Eighth IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 416–423.

Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012a). No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21:4695–4708.

Mittal, A., Soundararajan, R., and Bovik, A. C. (2012b). Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20:209–212.

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.

Radford, A., Metz, L., and Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*.

Readar B.V. (2024). Readar b.v. - high-quality geospatial data solutions. https://readar.com/en/. Accessed: 2025-04-11.

Ren, Z., He, L., and Lu, J. (2024). Context aware edge-enhanced gan for remote sensing image super-resolution. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:1363–1376.

Rhee, S. and Kang, M. G. (1999). Discrete cosine transform based regularized high-resolution image reconstruction algorithm. *Optical Engineering*, 38(8):1348–1356.

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*.

Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883.

Singh, G. and Mittal, A. (2014). Various Image Enhancement Techniques - A Critical Review. *International Journal of Innovation and Scientific Research*, 10(2):267–274.

Sun, J., Zhu, J., and Tappen, M. F. (2010). Context-constrained hallucination for image super-resolution. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 231–238.

Tsai, R. Y. and Huang, T. S. (1984). Multipleframe image restoration and registration. In *Advances in Computer Vision and Image Processing*, pages 317–339. JAI Press Inc., Greenwich.

Vishnukumar, S., Nair, M. S., and Wilscy, 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, L., Xiang, S., Meng, G., Wu, H., and Pan, C. (2013). Edge-directed single-image super-resolution via adaptive gradient magnitude self-interpolation. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(8):1289–1299.

Wang, T., Sun, W., Qi, H., and Ren, P. (2018a). 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, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C. C., Qiao, Y., and Tang, X. (2018b). Esrgan: Enhanced super-resolution generative adversarial networks.

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.

Xia, G.-S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., and Lu, X. (2017). AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981.

Xu, H., Zhai, G., and Yang, X. (2013). Single image super-resolution with detail enhancement based on local fractal analysis of gradient. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(10):1740–1754.

Xu, Y., Luo, W., Hu, A., Xie, Z., Xie, X., and Tao, L. (2022). TE-SAGAN: An improved generative adversarial network for remote sensing super-resolution images. *Remote Sensing*.

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.

Yang, F., Yang, H., Fu, J., Lu, H., and Guo, B. (2020). Learning texture transformer network for image super-resolution.

Yang, W., Feng, J., Yang, J., Zhao, F., Liu, J., Guo, Z., and Yan, S. (2017). Deep edge guided recurrent residual learning for image super-resolution. *IEEE Transactions on Image Processing*, 26(12):5895–5907.

Yang, Y. and Newsam, S. (2010). Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 270–279, San Jose, CA, USA.

Yu, M., Shi, J., Xue, C., Hao, X., and Yan, G. (2024). A review of single image super-resolution reconstruction based on deep learning. *Multimedia Tools and Applications*, 83(18):55921–55962.

Yu, Y., Li, X., and Liu, F. (2020). E-dbpn: Enhanced deep back-projection networks for remote sensing scene image superresolution. *IEEE Transactions on Geoscience and Remote Sensing*, 58(8):5503–5515.

Zeyde, R., Elad, M., and Protter, M. (2012). On single image scale-up using sparse-representations. In *Curves and Surfaces*, pages 711–730. Springer.

Zhang, J., Lei, J., Xie, W., Fang, Z., Li, Y., and Du, Q. (2023). Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15.

Zhang, K., Gao, X., Tao, D., and Li, X. (2012). Single image super-resolution with non-local means and steering kernel regression. *IEEE Transactions on Image Processing*, 21(11):4544–4556.

Zhang, K., Liang, J., Gool, L. V., and Timofte, R. (2021). Designing a practical degradation model for deep blind image super-resolution.

Zhang, L. and Wu, X. (2006). An edge-guided image interpolation algorithm via directional filtering and data fusion. *IEEE Transactions on Image Processing*, 15(8):2226–2238.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018a). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595.

Zhang, Y., Fan, Q., Bao, F., Liu, Y., and Zhang, C. (2018b). Single-image super-resolution based on rational fractal interpolation. *IEEE Transactions on Image Processing*, 27(8):3782–3797.

Zou, Q., Ni, L., Zhang, T., and Wang, Q. (2015). Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, 12(11):2321–2325.