UNIVERSIDADE
BEIRA INTERIOR

# Super-resolution Satellite Imagery for Crop Health Monitoring

**Ana Margarida Mendes Dias**

Dissertação para obtenção do Grau de Mestre em
**Engenharia Informática**
(2º ciclo de estudos)

Orientador: Prof. Doutor João Carlos Raposo Neves

**junho de 2024**

# Declaração de Integridade

Eu, Ana Margarida Mendes Dias, que abaixo assino, estudante com o número de inscrição M12563 do Mestrado em Engenharia Informática da Faculdade de Engenharia, declaro ter desenvolvido o presente trabalho e elaborado o presente texto em total consonância com o Código de Integridades da Universidade da Beira Interior.

Mais concretamente afirmo não ter incorrido em qualquer das variedades de Fraude Académica, e que aqui declaro conhecer, que em particular atendi à exigida referenciação de frases, extratos, imagens e outras formas de trabalho intelectual, e assumindo assim na íntegra as responsabilidades da autoria.

Universidade da Beira Interior, Covilhã 11/06/2024

Assinado por: **Ana Margarida Mendes Dias**
Num. de Identificação: 15768747
Data: 2024.06.11 22:38:54+01'00'

CHAVE MÓVEL

# Agradecimentos

Agradeço o todos os que me ajudaram, de alguma forma, a tornar possível esta dissertação de mestrado.

À Universidade da Beira Interior, enquanto instituição que me recebeu desde o início do 1º Ciclo de estudos e disponibilizou os recursos para chegar a esta fase da minha formação.

Ao Professor Doutor João Carlos Raposo Neves, meu orientador desta tese de mestrado, agradeço em particular, pela sua disponibilidade e apoio totais, pelos conhecimentos que me transmitiu, pela ajuda na resolução dos problemas que foram surgindo, pela sua orientação e incentivo ao longo do desenvolvimento deste projeto.

À equipa da empresa TeroMovigo por fornecer as imagens e os recursos, e pelo tempo dedicado às discussões indispensáveis para a realização deste trabalho de investigação.

Aos meus amigos, colegas de curso e colegas de laboratório que me acompanharam durante esta fase, pelo incentivo e o apoio que me proporcionaram.

Por fim, agradeço aos meus pais e irmão, pelo seu amor e apoio incondicional.

# Resumo

O crescimento exponencial da população global e as acentuadas preocupações com a sustentabilidade têm promovido o interesse na agricultura de precisão. Esta dissertação visa desenvolver uma metodologia de super-resolução que integra dados obtidos de dispositivos de deteção remota com dados multiespectrais para produzir imagens de alta resolução, e permitir uma monitorização eficiente da saúde do solo e das plantas em áreas agrícolas.
O uso da tecnologia de drones foi explorada nos últimos anos para monitorização de plantas e campos agrícolas, permitindo determinar diferentes métricas de saúde de plantas e solo. No entanto, o uso de drones requer intervenção humana o que aumenta o seu o custo de uso, sendo inacessível para pequenos agricultores. Além disso, a autonomia limitada destes dispositivos restringe a monitorização em áreas de grande tamanho. O uso de imagens de satélite contorna estes problemas, mas, por sua vez, a resolução das imagens, quando comparada com dados capturados por drones, representa um desafio. Este trabalho visa abordar o problema da baixa resolução das imagens de satélite usando técnicas inovadoras de processamento de imagem e aprendizagem automática, especificamente na área da super-resolução.
A metodologia envolve o uso de imagens de satélite em conjunto com dados adquiridos por drones, para treinar um modelo capaz de aproximar a alta resolução das imagens de drones. Em particular, este trabalho propõe uma nova estratégia baseada em *clustering* para melhorar o mecanismo de atenção dos *vision transformers*. Os resultados obtidos em *datasets* de imagens obtidas a partir de dispositivos de deteção remota sugerem que a estratégia proposta é capaz de alcançar um desempenho competitivo com abordagens do estado da arte. Mais importante ainda, os resultados obtidos sugerem que o uso da abordagem proposta pode ser utilizada para determinar o nível da saúde das plantas usando imagens de satélite, uma vez que o Normalized Difference Vegetation index (NDVI) estimado a partir dos dados obtidos a partir da super-resolução difere num máximo de 13% dos valores derivados de dados adquiridos por drones, onde pelo menos 50% dos valores apresentam uma diferença igual ou inferior a 3.6%.

# Palavras Chave

Super-resolução, Deteção Remota, Agricultura de Precisão, Imagens Multi-espectrais, Visão Computacional

# Resumo Alargado

Esta tese aborda o desenvolvimento de uma técnica avançada de super-resolução para converter imagens de satélite de baixa resolução em imagens de alta resolução comparáveis às capturadas por drones. Este estudo é motivado pela necessidade crítica de inovações tecnológicas que apoiem a gestão sustentável de recursos agrícolas, uma preocupação crescente devido ao aumento acelerado da população mundial e aos desafios da sustentabilidade ambiental.

O método proposto combina o uso de imagens multiespectrais, obtidas por satélite e drones, utilizando técnicas de aprendizagem automática e visão computacional. O ponto principal da metodologia envolve a melhoria do mecanismo de atenção dos *vision transformers*, com um mecanismo de *clustering* projetado para aumentar a capacidade do modelo em focar nas regiões críticas das imagens, melhorando assim a resolução e a utilidade dos dados para monitorização precisa.

Além da introdução de técnicas de super-resolução, esta tese aborda extensivamente o estado da arte, comparando várias abordagens e modelos existentes e destacando as suas vantagens e limitações quando aplicados ao contexto específico da agricultura de precisão. O modelo proposto é avaliado através de métricas padrão da área da super-resolução, assim como através da qualidade da estimativa da fitossanidade de plantas através do uso de um indíce vegetativo em *datasets* obtidos de áreas agrícolas. Os resultados demonstram que o método de super-resolução desenvolvido consegue alcançar um desempenho competitivo com abordagens do estado da arte nas várias métricas avaliadas.

Os resultados detalhados neste projeto de investigação demonstram a viabilidade de utilizar métodos de super-resolução para monitorizar a saúde de plantas em extensas áreas agrícolas de forma eficaz. A técnica de super-resolução oferece uma alternativa mais acessível e escalável à monitorização realizada através de drones, viabilizando a aplicação em vastas áreas agrícolas reduzindo significativamente os gastos operacionais, promovendo também práticas agrícolas mais sustentáveis e economicamente viáveis, alinhando-se com os objetivos de uma agricultura mais eficiente e responsável.

Em conclusão, a tese propõe uma nova metodologia que tem o potencial de revolucionar o campo da agricultura de precisão. Os resultados obtidos não só validam o desempenho do método de super-resolução proposto mas também abrem caminhos para futura investigação e aplicações em outros aspectos da deteção remota e gestão ambiental.

# Abstract

The exponential growth of global population and the increasing concerns regarding sustainability have fostered the interest in precision agriculture. This dissertation focuses on developing a super-resolution method that integrates remote sensing and multispectral data to generate high-resolution images for effective monitoring of soil and plant health in crops. The use of drone technology has been exploited during the last years in crop monitoring allowing to determine different metrics from plant and soil health. However, their use requires human intervention increasing their usage cost, being unaffordable for small producers. Also, their limited autonomy restricts the monitoring to large-sized areas. The use of satellite imagery addresses these problems, but, in turn, the resolution of the images when compared to drone-captured data represents a challenge. This work aims to address the problem of low-resolution satellite imagery using innovative image processing techniques and deep learning, specifically in the field of super-resolution.
The methodology involves leveraging satellite imagery paired with UAV-acquired data, to train a model capable of approximating the high resolution of UAV images. In particular, this work proposed a novel clustering-based strategy for improving the attention mechanism of vision transformers. The results obtained on a proprietary and publicly available remote sensing dataset suggest that the proposed strategy is capable of achieving competitive performance with state-of-the-art approaches. More importantly, the results obtained suggest that the use of proposed approach can be used to determine plant health using satellite imagery since the estimated NDVI from the super-resolved data differs by a maximum of 13% from the values derived from UAV-acquired data, where at least half of the values present a difference of less than or equal to 3.6%.

# Keywords

Super-resolution, Remote-Sensing, Precision Agriculture, Multispectral Imagery, Computer Vision

# Contents

# List of Figures

# List of Tables

# Acronyms List

**AID** Aerial Image Dataset

**ART** Attention Retractable Transformer

**BIRCH** Balanced Iterative Reduction and Clustering using Hierarchies

**CAB** Channel-attention convolution block

**CAT** Cross Aggregation Transformer

**CATB** Cross Aggregation Transformer block

**CF Tree** Clustering Feature tree

**CNN** Convolutional Neural Networks

**dB** Decibel

**EVI** Enhanced Vegetation Index

**FFN** Feedforward Neural Network

**GAN** Generative Adversarial Network

**GNDVI** Green Normalized Difference Vegetation Index

**HAB** Hybrid Attention Block

**HAUNet** Hybrid Attention-Based U-Shaped Network

**HAT** Hybrid Attention Transformer

**HR** High Resolution

**ISPA** International Society of Precision Agriculture

**LR** Low Resolution

**MLP** MultiLayer Perceptron

**MSA** Multi-head Self-attention

**MSE** Mean Squared Error

**NDRE** Normalized Difference Red Edge

**NDVI** Normalized Difference Vegetation index

**PSNR** Peak Signal-To-Noise Ratio

**ViT** Visual Transformer

**RGB** Red-Green-Blue

**RSTB** Residual Swin Transformer block

**SAVI** Soil Adjusted Vegetation Index

**SSIM** Structural Similarity Index Measure

**STL** Swin Transformer layer

**UAV** Unmanned Aerial Vehicle

# Chapter 1

# Introduction

## 1.1 Scope and Motivation

The exponential growth of the global population and the pressing concerns regarding sustainability have prompted a growing interest in precision agriculture. Innovative approaches have emerged to monitor and optimize crop health and productivity while sustaining a rapidly expanding population and preserving limited resources.
Several recent proposals have tackled the challenge of crop monitoring using drone technology. These solutions make it possible to detect and evaluate various factors that impact crop health, including moisture levels, the extent of weed growth, and the occurrence of disease outbreaks. Despite these advancements, limitations to autonomy and cost considerations often limit drones to medium-sized areas. Recognizing the need to scale up monitoring efforts to encompass more extensive agricultural landscapes without experiencing prohibitive expenses, the exploration of satellite data has become more popular. However, challenges arise with satellite data, as the lower resolution compromises its applicability compared to drone-captured data. Efforts to bridge this resolution disparity have led to the exploration of innovative techniques in image processing with machine learning algorithms and deep learning, particularly in the field of super-resolution. Super-resolution strategies applied to satellite imagery can revolutionize precision agriculture by providing enhanced spatial resolution for monitoring crop health and environmental conditions.

## 1.2 Objectives

This dissertation aims to address the issue of effectively monitoring soil and plant health in crops by developing a super-resolution method capable of producing high-resolution images. These images will facilitate precise monitoring of agricultural fields. The methodology involves leveraging satellite imagery provided by the European Space Agency, which will be paired with data acquired from UAVs. The final goal is to train a model capable of approximating the resolution of satellite images to that of UAV imagery.
A series of crucial tasks are defined to achieve this objective successfully. Initially, the dataset containing satellite and UAV images sourced from the same geographical locations is subjected to preprocessing. This step ensures that the data is prepared appropriately for subsequent analysis. Following preprocessing, state-of-the-art super-resolution methods are evaluated using the prepared dataset. Through this evaluation, the most effective methods are identified. Subsequently, enhancements will be made to one of these methods to better align with the distinctive characteristics of crop images. Adapting the method to suit the specifics of crop images is expected to significantly improve the overall approach's effective-

ness. Finally, the refined approach will be implemented into a web service to be accessible and usable by all relevant stakeholders. By addressing these objectives, this research aims to advance precision agriculture and remote sensing technologies, offering valuable insights for monitoring and enhancing crop health, thereby contributing to the sustainable management of agricultural resources.

## 1.3 Tasks and Timeline

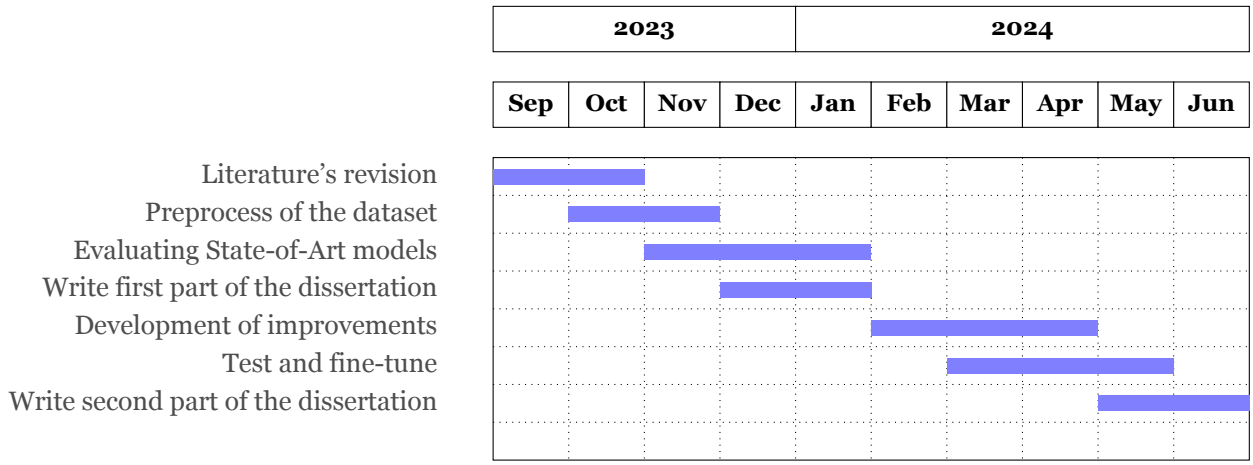



Figure 1.1: Timeline for the development of the tasks

To achieve the objectives of this dissertation, the following tasks have been proposed in the work plan and are presented in the Gaant Chart in 1.1.

### 1.3.1 Review of the literature and related topics

The project's initial phase involves conducting a comprehensive review and analysis of the specialized literature and related works. This phase is dedicated to study the super-resolution task. A thorough investigation is performed into state-of-the-art general super-resolution and remote-sensing super-resolution models. This study explores various algorithms, techniques, and models in this field and includes the examination of different architecture models such as CNN, Generative Adversarial Network (GAN), and Visual Transformer (ViT). Furthermore, this chapter investigates into the specifics of evaluation metrics commonly employed in assessing super-resolution algorithms. Additionally, explores the use of remote sensing images in precision agriculture applications, including how remote sensing technologies, including super-resolution techniques, contribute to optimizing and managing agricultural processes such as crop monitoring, yield prediction, and resource allocation.

### 1.3.2 Preprocess of the dataset

This task involves preprocessing the dataset comprising the satellite and drone images of the same geographical location. The processing starts by creating a structured dataset from the paired satellite and drone images' acquired images, with a defined naming system for each

pair. Next, quality control and data validation are performed, including assessing image quality, identifying artifacts or errors, and ensuring data integrity throughout the dataset. Finally, the dataset is partitioned into training, validation, and testing subsets to train and evaluate the models.

### 1.3.3 Evaluation of state-of-the-art super-resolution methods

From the literature review, some state-of-the-art techniques are chosen for evaluation in this task. This involves adapting the model's implementations to allow a dataset of satellite imagery as input, considering factors such as the image type and number of channels. Performance metrics and evaluation criteria are also established to assess the effectiveness and fidelity of the super-resolution methods. The super-resolution techniques are also compared against each other.

### 1.3.4 Development of improvements to a super-resolution method to enhance its performance to the particularities of the crops

This task conducts an analysis to identify the specific challenges that models presented to perform the super-resolution task using remote sensing data and how the characteristics of these data influence the effectiveness of super-resolution methods. After this, improvements are developed to enhance the performance and adaptability of a method chosen from the previous task to crop-related applications. The development of improvements to the super-resolution method adapted for crop analysis will represent a significant advancement in precision agriculture, enabling more accurate and timely assessment of crop health, productivity, and resource management practices.

### 1.3.5 Test and fine-tune the proposed method

This phase involves rigorous testing and fine-tuning of the method and improvements proposed in the previous task to ensure their effectiveness and efficiency in achieving the desired objectives. A structured testing protocol is defined to evaluate proposed enhancements' performance, including defining test scenarios and evaluation metrics to assess the method's functionality and efficacy. The performance of the proposed method is benchmarked against existing approaches and state-of-the-art techniques in the field.

### 1.3.6 Writing of the master's dissertation, technical documentation, and a journal or conference paper

This task encompasses the culmination of the research project, where the research, methodologies, and results are synthesized into various written documents, including the master's dissertation, technical documentation, and a journal or conference paper. This task will be mainly done in two periods and interleaved with other tasks.

## 1.4 Document Organization

This document is organized into chapters as follows:

- Chapter 1 - Introduction: this chapter delineates the scope and motivation of this dissertation, as well as the objectives, tasks, and associated timelines;
- Chapter 2 - Key Concepts: this chapter introduces fundamental concepts to this dissertation's scope. It elucidates topics such as image super-resolution, encompassing both general and remote sensing applications, and the differences between multispectral and hyperspectral imagery.
- Chapter 3 - State of the Art: this chapter presents the state of the art for general super-resolution and remote sensing super-resolution, highlighting diverse model architectures and presenting the latest advancements in general super-resolution and remote sensing applications. Furthermore, it examines the field of multispectral super-resolution and investigates the techniques utilized in precision agriculture and remote sensing to monitor crop health.
- Chapter 4 - Proposed Method: this chapter provides a detailed explanation of the proposed methodology. It covers the background, the underlying and core concepts, and a precise description of how the proposed model functions.
- Chapter 5 - Experiments and Results: this chapter presents the conducted experiments, including an overview of the dataset and the explanation of the performance evaluation process. The results obtained from these experiments are also discussed in detail.
- Chapter 6 - Conclusion: This chapter provides a comprehensive summary of the research and its outcomes, as well as insights into future work.

# Chapter 2

# Key Concepts

## 2.1 Introduction

Starting with fundamental concepts is essential for a thorough understanding of the topic addressed in this dissertation. Consequently, this chapter provides insight into the field of image super-resolution, encompassing both general and remote sensing applications. It also outlines the distinctions between multispectral and hyperspectral imagery, which are key elements in remote sensing and imaging technology.

## 2.2 Image Super-resolution

Image super-resolution is a research area focused to the reconstruction of high-resolution images from their lower-resolution versions. In single-image super-resolution, deep learning models are trained using pairs of low and high-resolution images to refine image quality. Meanwhile, in multi-image super-resolution, multiple low-resolution images are used to reconstruct a single high-resolution image, allowing for more detailed and accurate improvements.
Image super-resolution finds applications in domains like multi- and hyperspectral image refinement, depth-map enhancement [12], remote sensing, precision agriculture, biomedical imaging, and surveillance [13].
Supervised machine learning strategies for image super-resolution encompass various approaches, each contributing to model performance and adaptability. These include the positioning of upsampling layers with options such as pre-upsampling, post-upsampling, progressive-upsampling, and even iterative up-and-down techniques, as well as the choice of upsampling methods, like bicubic, bilinear, or sub-pixel layers. The choice of network design is the most crucial aspect, with all the existing options incorporating elements such as residual and recursive learning, CNN architectures, and attention mechanisms. The choice of the loss function and the integration of strategies like data augmentation, self-ensemble methods, and multi-task learning [12] further enhance the model's adaptability to specific needs. This flexibility highlights the numerous possibilities for customizing image super-resolution models to achieve optimal performance.

## 2.3 General vs Remote Sensing Super-resolution

General super-resolution refers to the broader concept of increasing the resolution of images across various fields where higher-resolution images are desired for improved interpretation or analysis. This encompasses scenarios including photography, digital imaging, medical

imaging, computer vision, and any context where the goal is to increase the level of detail in an image. The images may vary in characteristics, and the techniques are typically adapted to accommodate differences in content and quality.
Remote sensing super-resolution is a specialized field within image processing that is centered on enhancing the resolution of images acquired through remote sensing devices such as satellites or aerial sensors (like in UAVs). Primarily applied in Earth observation, remote sensing super-resolution is crucial for applications such as precision agriculture, land cover classification, natural resource management, environmental monitoring, urban planning, and other scenarios where detailed and precise spatial information is essential. Remote sensing super-resolution techniques often involve methods adapted to the characteristics of remote sensing data, considering factors like atmospheric conditions, sensor specifications, and the specific needs of the applications.
Both super-resolution tasks use standard Red-Green-Blue (RGB) imagery, multispectral, or hyperspectral imagery, depending on the scenario.

## 2.4 Multispectral vs Hyperspectral Imagery

Multispectral and hyperspectral imagery can be distinguished by the number of bands and the width of each band. Multispectral imagery typically includes 3 to 10 bands, capturing data at specific wavelengths in the visible to infrared regions of the electromagnetic spectrum. Each band is designated with a descriptive title, such as red, green, blue, near-infrared, and short-wave infrared. [14].
In contrast, hyperspectral imagery consists of significantly narrower bands, typically spanning 10-20 nanometers, derived from contiguous sections across both the visible and infrared regions of the electromagnetic spectrum.
Figure 2.1 visually represents the difference between multispectral, with distinct bands that showcase the discrete nature of the captured spectral information, and hyperspectral imaging, which exhibits a continuous spectrum of wavelengths with a larger number of bands.

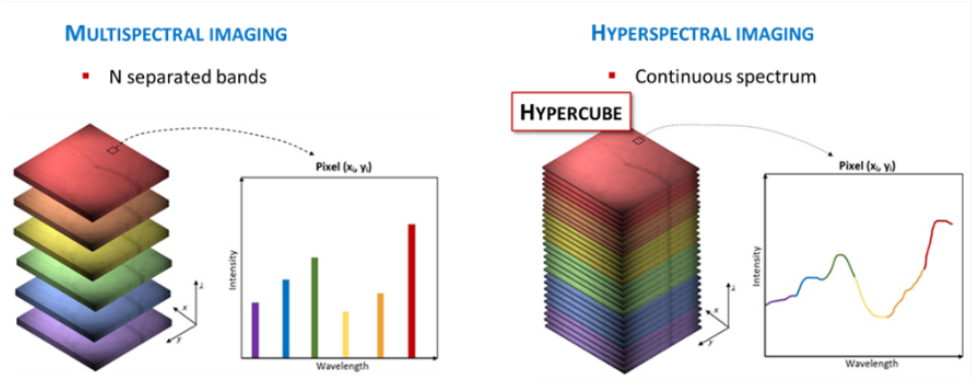



Figure 2.1: Visual difference of multispectral and hyperspectral imagery structure [1]

Several satellites are equipped with multispectral sensors. For example, Landsat-8 [15] features two sensors: the Operational Land Imager (OLI) with nine spectral bands (eight at 30 meters and one panchromatic at 15 meters spatial resolution) and the Thermal Infrared

Sensor (TIRS) with two bands at 100 meters resolution. Another example is the Sentinel-2 satellite, which offers 13 spectral bands: four visible and near-infrared bands at 10 meters, six red-edge/shortwave-infrared bands at 20 meters, and three atmospheric correction bands at 60 meters spatial resolution. These satellite sensors capture a range of bands, including Red, Green, Blue, Near-Infrared, and Short-wave Infrared.
Additionally, there are satellites with hyperspectral sensors, such as the Earth-Observing One (EO-1). This satellite features three instruments, including the hyperspectral instrument Hyperion [16], which provides 220 spectral bands (ranging from 0.38 to 2.58 micrometers) with a 10-nanometer bandwidth and a 30-meter spatial resolution.

## 2.5 Conclusion

In conclusion, super-resolution, the process of reconstructing high-resolution images from low-resolution counterparts, represents a significant advancement in imaging technology with diverse applications. Its ongoing evolution underscores its adaptability, offering various possibilities to tailor image super-resolution models to specific requirements and optimize performance.
Integrating remote sensing data amplifies the scope of technological applications, particularly in fields where detailed spatial information is crucial. Remote sensing enables a range of essential applications, including land cover classification, environmental monitoring, urban planning, natural resource management, and precision agriculture.
Furthermore, multispectral imagery, which captures data at specific bands, and hyperspectral imagery, which captures data across a continuous spectrum, enhance the efficacy of image super-resolution processes. These imaging modalities increase the breadth of available data and improve the depth and precision of analysis, thereby advancing our ability to extract valuable insights from imagery.

# Chapter 3

# State of the Art

## 3.1 Introduction

This chapter explores advancements in image super-resolution within the computer vision and remote sensing domains. It analyzes the main model architectures and the progress in state-of-the-art designs, with detailed explanations provided for some of these models. Additionally, multispectral super-resolution and the application of these techniques in precision agriculture are reviewed.

## 3.2 Model Architectures

### 3.2.1 Convolutional Neural Network

Convolutional Neural Networks, introduced in [17], are neural networks designed for processing data exhibiting a grid-like topology, such as images. A CNN comprises three main categories of layers: convolutional layers, pooling layers, and fully-connected layers. Figure 3.1 illustrates the conventional architecture of a CNN used for image classification.

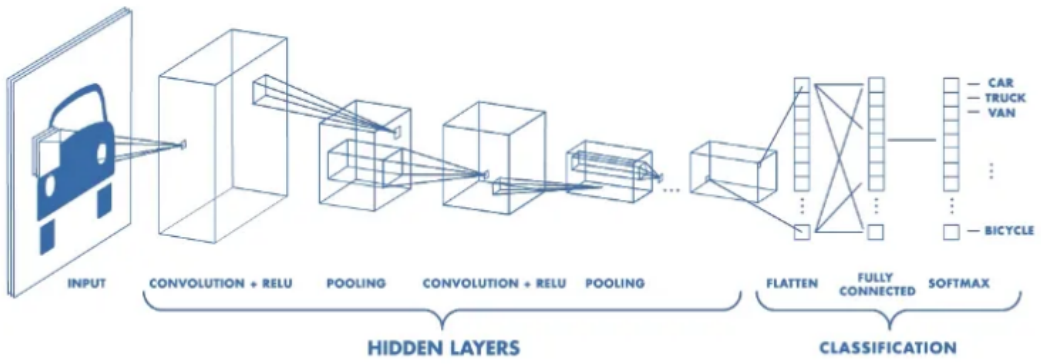



Figure 3.1: CNN architecture for image classification [2]

**Convolutional Layers** use a kernel (a small matrix of weights) to move across the receptive field of an input image, detecting the presence of specific features. The convolutional process involves sliding the kernel across the image. At each position, a dot product is computed between the kernel's weights and the image's pixel values within the receptive field, generating a representation of that specific region. This operation is repeated across the entire image over multiple iterations, producing a set of feature maps. These maps indicate the presence and intensity of various features throughout the image. Multiple convolutional layers are often stacked to identify more complex patterns progressively.

Using convolutional operations, where the kernels's weights are shared across different neurons, enables CNNs to demonstrate the property of equivariance to translation. This property asserts that, for a function to be translation-equivariant, a translation in the input space corresponds to a translation in the output space. In other words, as the filter traverses the input, the CNN recognizes patterns regardless of location, enhancing its capability to comprehend spatial relationships."
Non-linear activation functions, such as Sigmoid or ReLU, are commonly applied after convolution to introduce non-linearity into the model.

**Pooling Layers** follow convolutional layers and aim to reduce the dimensionality of the input data while preserving essential information. Similar to convolutional layers, pooling layers use a sliding two-dimensional filter to summarize the features within the covered region.
The most common pooling functions are "max pooling," which retains the maximum value within a window (determined by the kernel size) and discards the other values, and "average pooling," which calculates the average of the values within the window. Pooling layers help achieve translation invariance, ensuring that the network's recognition of high-level features or objects remains consistent even when these features shift to different positions in the input.

**Fully Connected Layers** integrate features extracted by the previous layers and map them to specific classes or outcomes. Each input from the prior layer connects to every activation unit in the fully connected layer, allowing the CNN to consider all features simultaneously. The training process involves backpropagation and optimizes the weights to achieve accurate predictions.

### 3.2.2 Visual Transformer

The Vision Transformer, usually ViT, was proposed in the paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." [3], is a model designed for tasks related to computer vision that employs a Transformer-like architecture to process image patches through self-attention mechanisms. The functioning of the Vision Transformer involves several steps, each contributing significantly to its overall operation. The original Transformer model, introduced in the paper 'Attention Is All You Need' [18] in the context of natural language processing tasks, laid the foundation for the architectural design employed in ViT. The architecture of the original Vision Transformer, used for image classification in diagram form, is presented in Figure 3.2.
Initially, the input image is divided into fixed-size square patches, referred to as tokens. Each patch is linearly transformed into a vector through a learnable linear projection. This process results in a series of token embeddings, which serve as input tokens for the subsequent layers. Since ViT inherently lacks an understanding of spatial relationships, positional information is added by incorporating positional encodings into the token embeddings.

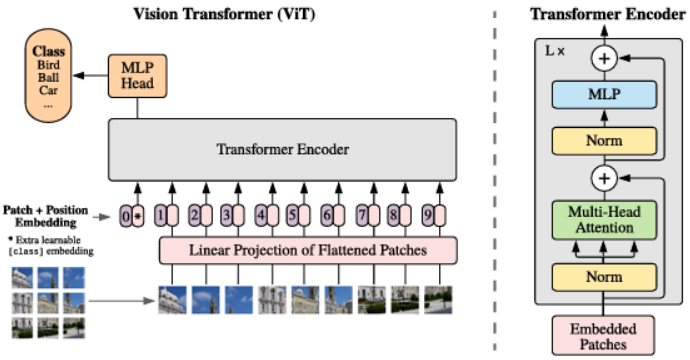



Figure 3.2: Visual Transformer architecture for image classification [3]

These encodings provide essential information for the model to distinguish between different positions within the image and effectively capture spatial relationships. Next, the token embeddings are fed into the Transformer encoder. The encoder is composed of two primary sub-layers: a multi-head self-attention mechanism and a feedforward neural network.

The core component of the first sub-layer is the self-attention mechanism, which captures relationships between different tokens in the input sequence. It involves applying three separate learnable linear transformations to each element, resulting in Query (Q), Key (K), and Value (V) vectors. Each linear transformation is defined by a weight matrix, $W_Q$, $W_K$, and $W_V$, respectively. These transformations are independently applied to each element in the input sequence, as in Equation 3.1, where $X_i$ represents the element in the input sequence's $i^{th}$ position. Consequently, each element possesses its unique set of Query, Key, and Value vectors.

During training, the weight matrices and other learnable parameters are iteratively adjusted to reduce the difference between the model's predictions and the actual target values. This process continuously improves the model's performance on the given task.

$$\begin{aligned} Q_i &= W_Q.X_i, \\ K_i &= W_K.X_i, \\ V_i &= W_V.X_i \end{aligned} \quad (3.1)$$

After obtaining the three vectors, attention scores are calculated for each element in the input by taking the dot product between its Query vector ($Q_i$) and the Key vectors ($K_j$) of all other tokens in the sequence. The softmax function is applied to the resulting attention scores to obtain a set of weights that indicate each token's relative importance in the current token's sequence. The calculation of the attention scores can be written as 3.2 where $Q_i$ is the Query vector for the token $i$, $K_j$ is the Key vector for the token $j$ and $d_K$ is the dimensionality of the Key vectors.

$$AttentionScores_{ij} = softmax(\frac{Q_i \times K_j^T}{\sqrt{d_K}}) \quad (3.2)$$

The weighted sum of the value vectors for the current token is then calculated using the attention scores and the value vectors (V) for all tokens in the sequence, as outlined in Equation 3.3. This weighted sum constitutes the output of the self-attention mechanism. By allowing the model to assess the significance of different elements in the sequence when processing a particular element, the self-attention mechanism effectively captures dependencies and patterns in the data.

$$WeightedSum_i = \sum jAttentionScores_{ij}.V_j \quad (3.3)$$

The multi-head self-attention layer applies the attention mechanism multiple times in parallel, with each head owning its own set of learnable parameters. $X_i$'s self-attention for the $m^{th}$ head can then be determined as3.4 where $Y_i^m$ represents the attention feature of $X_i$ in the $m^{th}$ head, and $Q_i^m, K_i^m, V_i^m \in \mathbb{R}^{C*d}$ indicate the query, key, and value projection matrices for the $m^{th}$ head. 'Dim' specifies the dimension of the channel, while 'B' denotes the learnable relative position encoding implemented in some models. The outputs from different heads are then concatenated or linearly transformed to generate the final output of the multi-head self-attention layer. Multiple attention heads allow the model to capture different patterns in the data, enhancing its representative capacity.

$$\begin{aligned}(Q_i^m, K_i^m, V_i^m) &= (X_i W_m^Q, X_i W_m^K, (X_i W_m^V),\\ Y_i^m = Attention(Q_i^m, K_i^m, V_i^m) &= SoftMax(\frac{Q_i^m (K_i^m)^T}{\sqrt{dim}} + B)V_i^m\end{aligned} \quad (3.4)$$

The output of the multi-head self-attention mechanism is often passed through a feedforward neural network layer, typically an MultiLayer Perceptron (MLP). This layer usually comprises a fully connected layer followed by a non-linear activation function, such as ReLU (Rectified Linear Unit). The feedforward network introduces non-linearity, enabling the model to learn complex relationships between tokens.

Following the self-attention mechanism and the feedforward network are layer normalization and residual connections. Layer normalization helps stabilize and accelerate training by normalizing the inputs to each sub-layer. Residual connections, also known as skip connections, add the original token embeddings to the output of each sub-layer. This assists in gradient flow during training and prevents the vanishing gradient problem. The whole process of the Transformer Encoder can be synthesized as in Equation 3.5, where X represents the input token, Multi-head Self-attention (MSA) represents the multi-layer self-attention mechanism, LN represents layer normalization, an MLP is used for the feedforward network, and Y represents the output for the transformer encoder.

$$\begin{aligned}X' &= MSA(LN(X)) + X\\ Y &= MLP(LN(X') + X'\end{aligned} \quad (3.5)$$

### 3.2.3 Generative Adversarial Networks

The GAN architecture, introduced in the paper ”Generative Adversarial Networks” [19], is a powerful framework for generative modeling. It consists of two sub-models: the generator and the discriminator.
The generator’s goal is to produce artificial data that closely resembles the original dataset. It typically takes random noise as input and transforms it into complex data samples, such as images or sequences. During adversarial training, the generator is continuously refined to improve its ability to produce realistic samples. The discriminator, on the other hand, is a neural network that acts as a classifier, distinguishing between real data from the original dataset and synthetic data created by the generator.
GANs employ an adversarial training paradigm where the generator and discriminator are in a competitive environment. The generator strives to create samples indistinguishable from real data, while the discriminator aims to correctly classify real and generated samples. Their interaction is guided by specific loss functions, such as binary cross-entropy loss, which direct the optimization process. As training progresses, the generator becomes better at producing realistic samples, and the discriminator becomes more proficient at distinguishing between real and generated data.

## 3.3 General Super-resolution

Since the pioneering work of SRCNN [20], which introduced deep CNNs to the image super-resolution task, many other CNN-based models have been proposed, employing various methods and architectural designs. These models include the implementation of residual blocks, as seen in VDSR [21], EDSR [22] and SRGAN [23]; dense blocks, as seen in RDN[24] and ESRGAN [25]; and attention mechanisms inside the CNN framework, RCAN [26] introduces a residual-in-residual structure combined with a channel attention mechanism, while SAN [27] presents a second-order attention network aimed at improving the learning of feature correlations, HAN [28] models the inter-dependencies between different layers, channels, and locations, and NSLA [29] employs non-local attention.
In addition to these approaches, several researchers have explored and developed alternative frameworks. For instance, recursive neural networks such as in DRCN [30] and DRRN [31], and graph neural networks as IGNN [32] have been investigated [31]. Furthermore, GANs have also been implemented to introduce adversarial learning to enhance perceptual quality and generate more realistic results [23], [25].
Although these CNNs have demonstrated outstanding performance in various applications, they frequently encounter two challenges in their architecture. Firstly, the interactions between images and convolutional kernels are content-independent, as the convolutional operations uniformly treat the data based on the learned parameters, and convolutional filters are not designed to adapt their behavior to the individual characteristics of the input content. Also, under the fundamental principle of local processing, the convolutional operations in CNNs are not well-suited for effectively modeling long-range dependencies in the input.
Transformer-based image models have emerged due to their self-attention mechanism. This

mechanism allows the modeling of long-range dependencies, ultimately leading to enhanced performance in super-resolution.
IPT [33] was designed using a pre-trained Transformer and a contrastive learning approach for learning universal features. Following that, SwinIR [5] was proposed based on the Swin Transformer, ART [8] was developed using an attention retractable module to enlarge the receptive field, and CAT [7] designed a new self-attention window to achieve a larger receptive field and a locality complementary module that enables the integration of global and local information. HAT [9] proposes to combine multiple attention mechanisms, such as channel attention, window-based self-attention, and overlapping cross-window attention.

Next, several state-of-the-art models will be explained in depth. These models are built using three key components: the shallow feature extraction module, the deep feature extraction module, and the reconstruction module.
In the shallow feature extraction module, given a low-quality image $I_{LQ} \in \mathbb{R}^{H \times W \times C_{in}}$ where H, W and $C_{in}$ represent height, width and input channel number of the image, a convolution layer extracts a shallow feature $F_0 \in \mathbb{R}^{H \times W \times C}$, where C represents the output feature number. The convolutional layer provides a straightforward method for transforming the input image space into a feature space of higher dimensionality. The resulting feature is then forwarded, via a long skip connection, to both the deep feature extraction module for further processing and the reconstruction module to ensure the preservation of low-frequency information.
In the deep feature extraction module, the deep feature $F_D \in \mathbb{R}^{H \times W \times C}$ is extracted from the received shallow feature $F_0$. During this phase, models typically differ in implementation, but the overall objective remains unchanged: recovering lost high-frequency data while stabilizing training. For each model in the subsequent sections, this stage will explained in greater detail.
Finally, the reconstruction module merges the shallow and deep features to generate a high-quality reconstruction of the input image in a process similar to 3.6, where $H_{REC}$ represents the reconstruction module's function and $I_{RHQ}$ represents the final reconstructed image.

$$I_{RHQ} = H_{REC}(F_0 + F_{DF}) \tag{3.6}$$

The four models explained in the next sections use a sub-pixel convolution layer for this module. This layer employs regular convolutional layers followed by a specialized image-reshaping technique known as phase shift. This phase shift, also called "pixel shuffle", is an operation that rearranges elements in a tensor of shape $(H, W, C \times r^2)$ to a tensor of shape $(rH, rW, C)$. [4] The operation can be visually represented as in the Figure 3.3.
The definition of the loss function is another crucial aspect of implementing super-resolution models. The following models optimize their parameters by minimizing the L1 Pixel Loss.
L1 loss, also known as mean absolute loss or mean absolute error, in this context, calculates the mean element-wise (pixel-wise) absolute value difference between the pixels in the

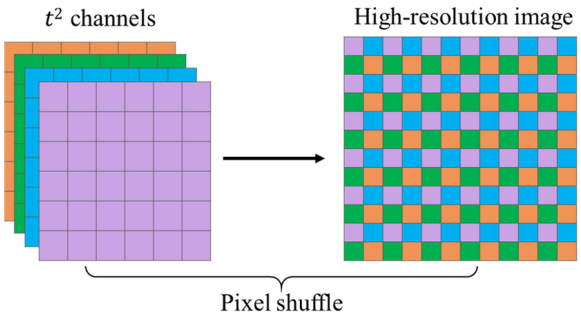



Figure 3.3: Pixel-Shuffle operation representation [4]

super-resolved image and the high-resolution image, as stated in Equation 3.7, where $I_{HQ}$ represents the original high-resolution image and where $I_{RHQ}$ represents the reconstructed high-resolution image, the super-resolved one.

$$L1Loss = MAE(X, Y) = \frac{1}{N}\sum_{n=1}^{N}|I_{RHQ_i} - I_{HQ_i}| \tag{3.7}$$

### 3.3.1 SwinIR: Image Restoration Using Swin Transformer

SwinIR [5] is an image restoration model encompassing super-resolution, based on the Swin Transformer [6]. Figure 3.4 shows the architecture of this model.

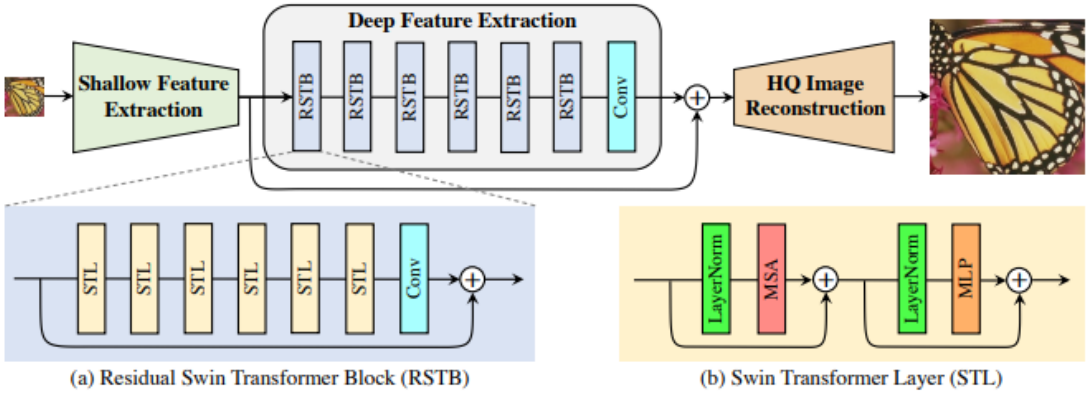



Figure 3.4: SwinIR architecture [5]

#### 3.3.1.1 Deep feature extraction module of SwinIR

The deep feature extraction module of SwinIR comprises a set (K number) of (Residual Swin Transformer blocks (RSTBs)) and a 3×3 convolutional layer. Each RSTB extracts intermediate features $F_1, F_2, ..., F_K$ block-by-block. The input for the first RSTB is the $F_0$ feature obtained from the shallow feature extraction module, and subsequently, the input for each

RSTB is the output of the previous one. The final deep feature $F_{DF}$ is obtained after a final convolution layer. This process is summarized in Equation 3.8 where $RSTB_i(.)$ denotes the $i^{th}$ RSTB and $Conv(.)$ represents the stated convolutional layer.

$$\begin{aligned} F_i &= RSTB_i(F_i - 1), i = 1, 2, ..., K \\ F_{DF} &= Conv(F_K) \end{aligned} \tag{3.8}$$

**Residual Swin Transformer blocks** The RSTB, represented in the subfigure (a) of Figure 3.4, is a residual block with a set of Swin Transformer layers (STLs) and a convolutional layer.

For each input feature ($F_{i,0}$) of the $i^{th}$ RSTB, each STL extracts intermediate features $F_{i,1}, F_{i,2}, ..., F_{i,N}$, where N is the number of STLs. This process is illustrated in Equation 3.9, where $STL_{i,j}$ represents the $j^{th}$ STL in the $i^{th}$ RSTB. Subsequently, after extracting all these features, they are passed through a convolutional layer. Finally, the residual connection adds the input feature of the RSTB to the feature obtained from this process.

$$F_{i,j} = H_{STL_{i,j}}(F_{i,j} - 1), j = 1, 2, ..., L \tag{3.9}$$

**Swin Transformer Layer** The STL is based on the standard multi-head self-attention from the original Transformer. The primary differences are the implementation of local attention and the shifted window mechanism.

Instead of dividing the input image into patches and treating each patch as a token for the calculation of self-attention, Swin Transformer [6] partitions the input of size $H \times W \times C$ into non-overlapping $M \times M$ local windows. This process reshapes this input to a $HW/(M^2) * M * C$ feature. The total number of windows is equal to $HW/M$. The standard self-attention is then calculated independently for every window, in which each pixel is treated as a token, introducing the local attention.

The query, key, and value matrices for a local window feature $X \in R^{M \times M \times C}$ are calculated by 3.1 where, $W_Q$, $W_K$, and $W_V$ are shared among windows. In a local window, the self-attention mechanism calculates the attention matrix as in Equation 3.10, where B represents a learnable relative positional encoding.

$$Attention(Q, K, V) = SoftMax(\frac{Q \times K^T}{\sqrt{d}} + B)V \tag{3.10}$$

For the MSA, the attention function is executed h times in parallel, with the outputs concatenated. As the original process, subsequent feature transformations are performed using a MLP consisting of two fully connected layers with GELU non-linearity between them. The layer normalization is added before the MSA and MLP (sub-figure (b) on Fig 3.4), and for both modules, a residual connection is used. This process is represented by Equation 3.5.

Regular and shifted window partitioning are used alternately to enable cross-window connections, where shifted window partitioning means shifting the feature by $(M/2, M/2)$ pixels before partitioning, as illustrated in Figure 3.5.

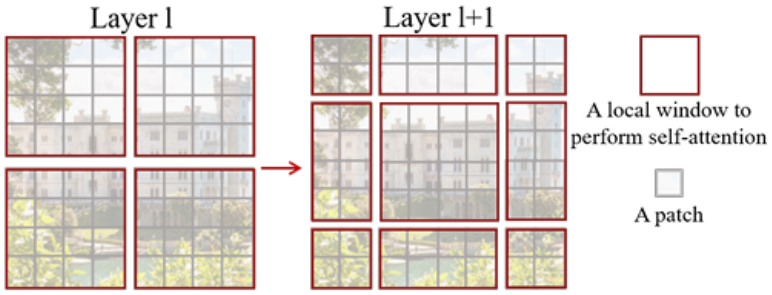



Figure 3.5: Window shift mechanism illustration. In the first layer, a regular window partition scheme is used. In the next layer, there is a shift in the window partition, resulting in different windows [6]

### 3.3.2 Cross Aggregation Transformer for Image Restoration

The CAT [7] introduces a new self-attention mechanism that employs rectangular window self-attention with an axial-shift operation and a locality complementary module to facilitate the integration of global and local information. The architecture of this model is presented in Figure 3.6.

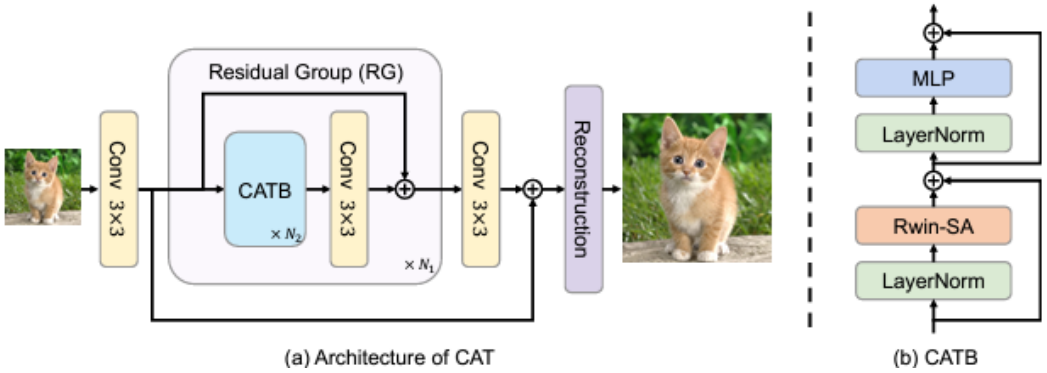



Figure 3.6: CAT architecture [7]

.

#### 3.3.2.1 Deep feature extraction module of CAT

The deep feature extraction module is composed of multiple Residual Groups (RG) and a single convolution layer that combines the previously extracted features from the Residual Groups.

**Residual Group (RG)** A Residual Group is formed by a series of Cross Aggregation Transformer block (CATB) and a convolution layer to add locality and translation invariance to the

Transformer blocks' output. In every Residual Group, a residual strategy is implemented.

**Cross Aggregation Transformer Block** The CATB implements three innovative features: a rectangle-window self-attention mechanism, the axial-shift operation, and the locality complementary module. The CATB, Figure 3.6(b) is a transformer-based encoder with an MLP for the Feedforward Neural Network (FFN), where the window-self attention mechanism is replaced by the new rectangle-window self-attention. The MLP features a GELU non-linearity between its two linear projection layers. The two-layer normalization layers and the residual connections are implemented.

**Rectangle-window Self-attention** Instead of using the usual square window, CAT uses a new window attention mechanism, the Rwin-SA, that employs a rectangle window. This rectangle window is divided into two parts: the horizontal component is referred to as H-Rwin, and the vertical component is termed V-Rwin. These elements are used in parallel with different attention heads.
For each attention head, the input $X \in \mathbb{R}^{H \times W \times C}$ is split into non-overlapping sh $\times$ sw rectangle windows, where sh stands for the window height and sw stands for the window width. The $i^{th}$ rectangle window feature is represented as $X_i \in \mathbb{R}^{(sh \times sw \times C)}$. $X_i$'s self-attention for the $m^{th}$ head can then be determined as 3.4. The attention feature $Y_i^m \in \mathbb{R}^{HxWD}$ is obtained after performing the attention operation on all $X_i$.
The attention operation is identical for both H-Rwin and V-Rwin. The attention heads are split into two sections, where H-Rwin is applied to one section and V-Rwin to the other in parallel. Subsequently, the outputs from these sections are concatenated along the channel dimension.
CAT proposes another variant, the Axial Rectangle Window (Axial-Rwin), where one side's length of the rectangle is fixed to be the size of the height or width of the image, making the window become a strip along the axis.
The H-Rwin and V-Rwin enable the expansion of the attention area and aggregate features across windows without adding computational complexity. The Axial-Rwin, with its larger attention area compared to the other two windows, is able to capture more information, especially in the axial direction.

**Axial-Shift Operation** A new shift operation, called the axial-shift, is introduced, which is meant to increase the amount of information that each pixel can aggregate. It consists of two shift operations: the horizontal shift (H-Shift) for the H-Rwin and the vertical shift (V-Shift) designated for the V-Rwin. The window partition is moved down and left by $\frac{sh}{2}\frac{sw}{2}$ via the axial-shift operation, where sh and sw represent the H-Rwin and V-Rwin window height and width, respectively, as illustrated in Figure 3.7(b). The axial-shift operation is applied on the interval between two successive cross-aggregation transformer blocks.
The window partition is shifted downward and to the left by $\frac{sh}{2}\frac{sw}{2}$ through the axial-shift operation, where sh and sw denote the height and width of the H-Rwin and V-Rwin windows, respectively, as shown in Figure 3.7(b). This axial-shift operation occurs between two consecutive cross-aggregation transformer blocks.

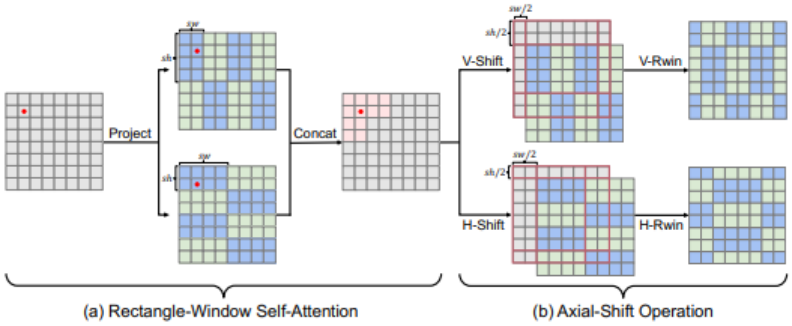



Figure 3.7: Rectangle-Window Self-Attention and Axial-Shift Operation [7]

**Locality Complementary Module** To enhance the Transformer's ability to handle locality and facilitate the integration of global and local information, an independent convolution operation is introduced for use in computing self-attention. This convolution is applied directly to value V as 3.11 where $V \in \mathbb{R}^{H \times W \times C}$ is the value that is directly projected from X without using window partition.

$$Rwin - SA(X) = (Concat(Y_1, Y_2 ... Y_M) + Conv(V))W^P \tag{3.11}$$

### 3.3.3 Attention Retractable Transformer

The Attention Retractable Transformer (ART) proposes the use of sparse attention based on the concept that interactions between tokens from a sparse image region can expand the module's receptive field. Considering this, this model alternates between applying two self-attention blocks to capture local and global receptive fields simultaneously. Figure 3.8 presents the model's architecture.

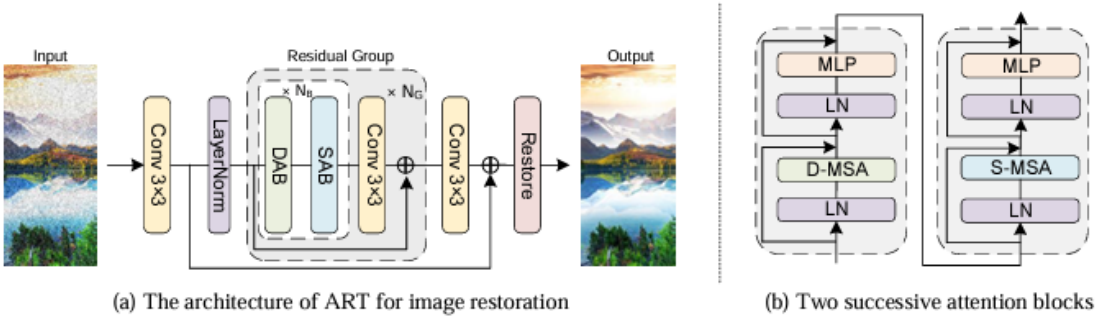



Figure 3.8: ART architecture [8]

#### 3.3.3.1 Deep feature extraction module of ART

A residual in residual structure is implemented for the deep feature extraction module. The shallow feature that is obtained is normalized and then directed into a series of residual groups. A deep feature is extracted and sent through a further 3x3 Convolution to obtain additional feature embeddings ($F_D$). The final feature map is acquired from the element-wise sum of $F_0$ and $F_D$.

**Residual Group** A Residual Group comprises pairs of two new proposed attention blocks, the DAB and the SAB, placed successively, each with its attention modules, the D-MSA and the S-MSA, respectively, and a final 3x3 convolution layer. Between each pair of blocks, a long-distance residual connection is implemented.

**Dense Attention (D-MSA)** The Dense attention, used in standard window self-attention, allows each token to interact with a smaller number of neighboring tokens within a non-overlapping $W \times W$ window. All tokens are divided into groups, each containing $W \times W$ tokens. For each group, self-attention is computed $\frac{h}{W} \times \frac{w}{W}$ times, where $h$ represents the image height, $w$ the image width, and $W$ the size of the window's side.

**Sparse Attention (S-MSA)** The proposed sparse attention mechanism enables each token to interact with a few tokens from sparse positions, defined by an interval size $I$. Subsequently, the updates of all tokens are divided into several groups, with each group containing $\frac{h}{I} \times \frac{w}{I}$ tokens. Self-attention is then computed $I * I$ times.

After computing all groups, the outputs are combined to reconstruct a feature map of the original size. Sparse attention is used to design the sparse attention block (SAB), and dense attention is used to design the dense attention block (SAB).

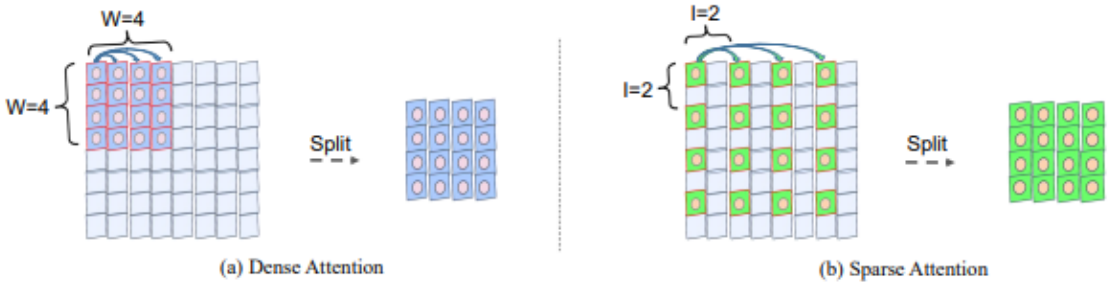



Figure 3.9: Dense and sparse attention [8]

### 3.3.4 Hybrid Attention Transformer

The HAT integrates channel attention with window-based self-attention techniques and features a cross-attention module to improve the interaction between neighboring window fea-

tures. The architecture of HAT is presented in Figure 3.10.

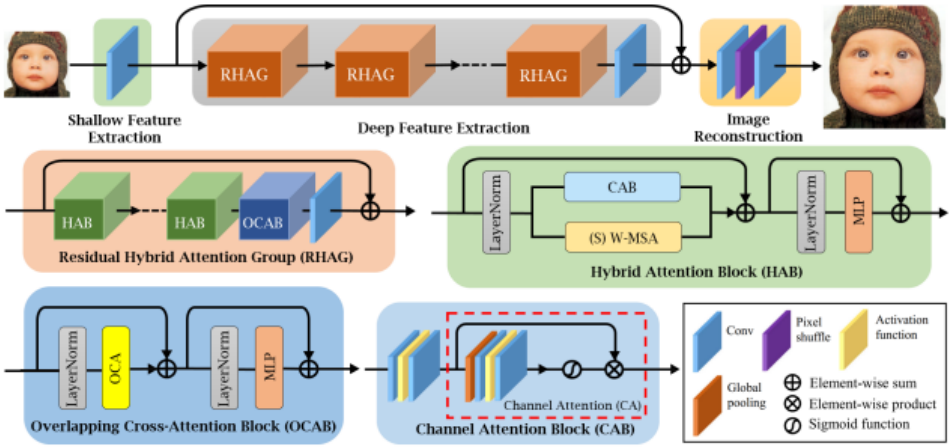



Figure 3.10: HAT architecture [9]

#### 3.3.4.1 Deep Feature Extraction of HAT

The deep feature extraction module consists of a series of residual hybrid attention groups (RHAG) and a single 3x3 convolution layer. Each RHAG includes multiple Hybrid Attention Block (HAB), an overlapping cross-attention block (OCAB), and a 3x3 convolution layer with a residual connection.

**Hybrid Attention Block** This block enhances the model's representation ability by incorporating a Channel-attention convolution block (CAB) into the Swin Transformer block. Following the normalization layer, CAB is inserted in parallel with the window-based multi-head self-attention (W-MSA) module. Consecutive HABs employ the shifted window-based self-attention. To prevent potential conflicts between CAB and MSA in optimization and visual representation, a small constant is multiplied by the output of CAB.
Given an input feature $X$, the entire process of HAB can be formulated as 3.12, where $X_N$, $X_M$ denote intermediate features and Y represents the output of HAB. In this model, the patch size is set to 1, which treats each pixel as a token for embedding.

$$\begin{aligned} X_N &= LN(X); \\ X_M &= (S)W - MSA(X_N) + \alpha CAB(X_N) + X; \\ Y &= MLP(LN(X_M)) + X_M; \end{aligned} \tag{3.12}$$

Then, the window-based self-attention is calculated as in Equation 3.10.
A CAB comprises two standard convolution layers with a GELU activation and a channel attention (CA) module, which includes a global pooling layer and two standard convolutional layers with an activation function.

**Overlapping Cross-Attention Block (OCAB)** The OCAB comprises overlapping cross-attention (OCA) and MLP layers, with the key difference from the standard procedure being that the standard window partition functions as a sliding partition with both the kernel size and stride set to M, while the overlapping window partition uses a kernel size of $M_o$ and a stride of M. The idea is that computing the key/value from a larger field than the standard window-based self-attention allows it to utilize more helpful information for the query.

## 3.4 Remote Sensing Super-resolution

Remote sensing images are used in various applications, including environmental surveys, disaster monitoring, scene analysis, object detection, and natural resource monitoring [10]. Spatial resolution is an essential remote sensing image characteristic that advanced satellite and aircraft researchers are working to improve. Furthermore, due to imaging equipment limitations, factors such as motion blur, atmospheric interference, ultra-long-range imaging, and transmission noise [10][34] can all impact remote sense image quality, causing varying degrees of degradation. The focus on researching methodologies for super-resolution in remote sensing images has intensified due to these factors.

Since deep-learning-based super-resolution algorithms have demonstrated excellent performance in general scene images, many researchers have proposed super-resolution algorithms for remote sensing images using deep-learning techniques.

LGCNet [35], the pioneering CNN-based model for remote sensing image super-resolution, introduces a local–global combined network. This model learns multiscale remote sensing data representations, leveraging local and global representations/features to effectively understand the image residuals between High Resolution (HR) and upscaled Low Resolution (LR) images. Haut et al. [36] introduce a deep compendium model, which integrates components including residual unit, skip connection, and network-in-network structure to extract more informative features. In [37], residual dense back-projection blocks (RDBPN) with two types of modules for up-projection and down-projection were proposed to exploit residual learning in both global and local manners. In DGANet-ISE [38], a gradient-aware loss is designed and combined with L1 loss to preserve more image gradient information and improve the recovered edges of the targets. Some GAN models were additionally presented, with the same objective as in general super-resolution, to enhance the visual outcomes of the super-resolution task. EEGAN [39] integrates an edge-enhancement structure into the traditional GAN framework to reduce artifacts and noise generated by adversarial training by simultaneously optimizing the high-frequency and low-frequency components. The CDGANs [40] introduced three elements in the discriminator: a dual-path network architecture, a random gate, and coupled adversarial loss. The first two aim to enhance enhanced discriminatory capabilities and the last is designed to learn the better correspondence between the discriminative results and the paired inputs. TE-SAGAN [41] integrates the weight normalization, instead of batch normalization, and self-attention mechanism into the GAN. Also, a joint loss is designed to combine content loss, perceptual loss, adversarial loss, and texture loss. Continuous super-resolution entails magnifying an image to arbitrary

scaling levels rather than being limited to specific integer factors. This approach permits greater adaptability in managing diverse levels of magnification. RSI-HFAS[42], SADN[43], and [44] are models that implement this. Diffusion models have also been applied to the super-resolution task, like in EHC-DMSR[45] and EDiffSR[46]. The idea behind using diffusion models in super-resolution is to simulate the diffusion process to refine the image at a higher resolution iteratively. Some models integrated attention mechanisms in their implementations. MHAN[34] applies weights to different levels of convolution in the feature extraction stage to retain more critical information and adds a frequency-aware connection in the feature refinement stage to fuse and refine the features of different depths through the high-order attention module. Using non-local attention, HSENet[47] exploits the hybrid-scale self-similarity information in the remote sensing images. TransENet [48] proposes a multi-stage enhanced transformer that explores features at different scales with self-attention. This structure allows the fusing of multiscale low-/high-dimensional features, capturing the long-term dependencies between them. In HAUNet [10], two kinds of convolutional attention-based single-scale feature extraction modules are built at different levels to emphasize the global-specific context and abstract content information while maintaining detailed local information. GCRDN [49] proposes a non-local sparse residual dense encoder incorporating non-local sparse attention into residual dense networks to capture similar contextual information from a global perspective. DTRN [50] proposes a dual-branch model empowered with one transformer branch to characterize long-distance global spatial correlations and one CNN-based residual branch to extract local features. Features extracted within each branch are progressively fused between branches, which enables more effective global and local feature fusion. In [11], redundant token representation in remote sensing scenarios is considered, and the model adaptively selects the most critical tokens based on the top-k selective mechanism, eliminating the interference of irrelevant tokens and making the long-range modeling more effective and compact.

### 3.4.1 Hybrid Attention-Based U-Shaped Network for Remote Sensing Image Super-Resolution

HAUNet [10] is a hybrid attention-based U-shaped model for remote super-resolution. It proposes two types of feature extraction modules (S-CEM and CEM) that use attention mechanisms at different levels. These modules emphasize global context, abstract content, and local details. Additionally, the architecture includes a Cross-Attention-Based Multiscale Enhancement (CIM) that bridges semantic and resolution gaps between different scale features by collaboratively fusing them, enabling adaptive and context-aware feature integration. The architecture of this model is presented in 3.11

#### 3.4.1.1 Overview

A typical convolutional layer transforms the input from pixel space to low-level feature embeddings. The shallow feature passes through three-level encoders that hierarchically reduce the spatial size and transform features into multiscale features with different resolutions as

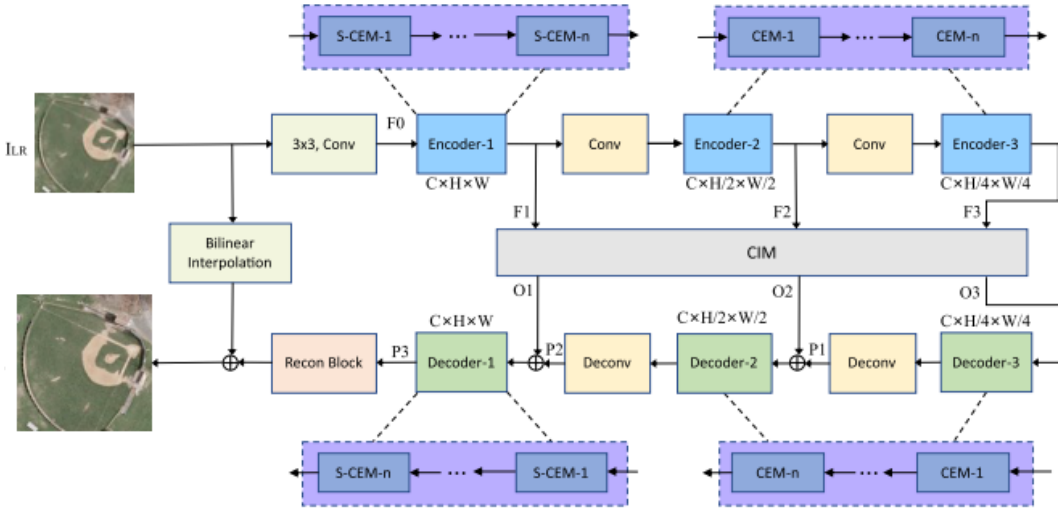



Figure 3.11: HAUNet architecture [10]

3.13, where $F_0$ represents the shallow feature and $Conv_{2\times2}$ represents a 2 × 2 convolutional layer with stride 2 for feature downsampling.

$$\begin{cases} F_1 = Enc_1(F_0) \\ F_i = Enc_i(Conv_{2\times2\downarrow}(F_1), i = 2, 3 \end{cases} \tag{3.13}$$

After computing F3, HAUNet employs the CIM to fuse and bridge the gaps between these three scale features [F1, F2, F3]. The output fused-scale features can be computed as 3.14

$$[O_1, O_2, O_3] = CIM(F_1, F_2, F_3). \tag{3.14}$$

The lowest-resolution output feature of CIM ($O_3$) is the input of the last-level decoder. After that, the sum of the decoder output features with the same same-resolution CIM output features ($P_i$) is input for the following decoders as 3.15. This mechanism is used to recover and enrich the original-scale representations progressively.

$$\begin{cases} P_1 = DeConv \downarrow 2 \times 2(Dec_3(O_3)) \\ P_2 = DeConv \downarrow 2 \times 2(Dec_2(O_2 + P_1)) \\ P_3 = Dec_1(O_1 + P_2) \end{cases} \tag{3.15}$$

The output feature for Decoder-1 ($P_3$) is upsampled through a convolutional layer and pixel-shuffle operations. The final image is the addition of a bilinear interpolation of the original input LR and the previous output.

**Single-Scale Feature Extraction Modules** HAUNet renovates the conventional attention modules into convolutional attention-based single-scale feature extraction modules

(SEM): S-CEM and CEM. These modules employ convolutional attention blocks at the channel and spatial levels, known as channel attention block (CAB) and spatial attention block (SAB). The channel level captures global abstract information, while the spatial level focuses on detailed spatial relationships. To maintain model efficiency, S-CEM incorporates both SAB and CAB for original scale features (Encoder and Decoder 1), while CEM uses only CAB for lower scales (Encoder and Decoder 2 and 3).

From a layer-normalized input feature $X$, both S-CEM and CEM first apply $1 \times 1$ convolutions $W_p()$ to aggregate pixel-wise cross-channel context, followed by $3 \times 3$ depth-wise convolutions $W_d()$ to emphasize channel-wise spatial context. The resulting projections are treated as query, key, and value for attention computation. The attention map is calculated using dot-product interaction between reshaped query and key projections, followed by softmax normalization with a ReLU gating mechanism. This mechanism controls which complementary features should be prioritized and forwarded, enabling subsequent layers to focus on refined image attributes as illustrated in Figure 3.12. The feature feedforward network consists of two $1 \times 1$ convolutions and a non-linear SimpleGate activation function that divides the feature into two parts along the channel dimension and multiplies them, thereby controlling the flow of complementary features.

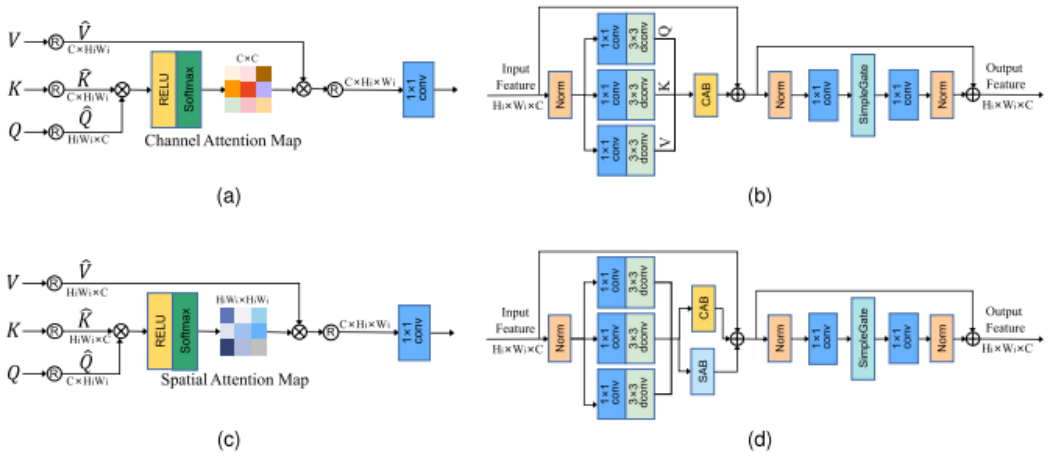



Figure 3.12: The two attention mechanisms (a) CAB and (b) SAB and the two feature extraction modules of HAUNet, (c) CEM, and (d) S-CEM. [10]

**Cross-Attention-Based Multiscale Enhancement** This module aims to improve information flow between encoders and decoders in a multiscale context. It is introduced to address the issue of lost information during downsampling and to facilitate the recovery of fine details. The CIM adaptively fuses encoder features from each scale with those from other scales to address semantic gaps and enhance the perception of multiscale contents.

As illustrated in image 3.13, an upsample operation is performed to match resolutions given the outputs of three scale encoder features. Then, the tokens of all three layers are concatenated and form key and value tensors for multi-head cross-attention. Each token is used as a query to perform multi-head cross-attention from the channel perspective, like in CAB. As stated, the resulting features are processed through multi-head attention and a forward net-

work. The lowest-resolution output features of CIM serve as input for the last-level decoder, while outputs from other resolutions are added to corresponding decoder features.

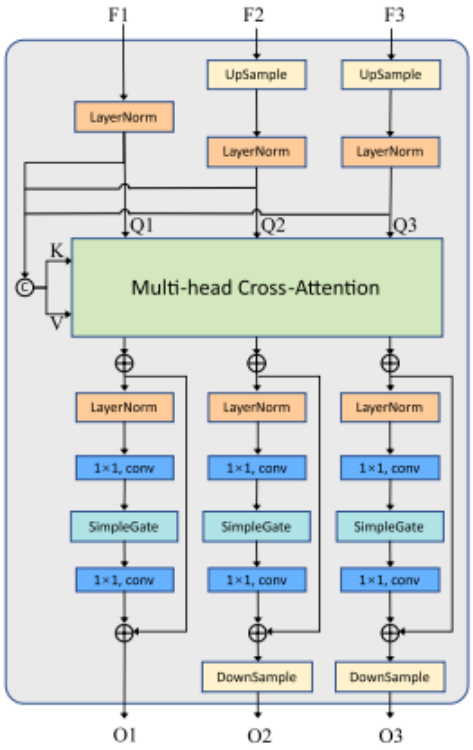



Figure 3.13: CIM [10]

### 3.4.2 TTST: A Top-k Token Selective Transformer for Remote Sensing Image Super-Resolution

Following the standard architecture for super-resolution models, TTST [11] comprises three main components: the shallow feature extraction module, the deep feature extraction module, and the reconstruction module, using pixel-shuffle for the upsampling. The architecture is illustrated in Figure 3.14.

#### 3.4.2.1 Deep feature extraction module of TTST

The deep feature extraction module comprises several Residual Token Selective Groups (RTSGs), each featuring four main components: the Top-k Token Selective Attention (TTSA), the standard window-based self-attention, the Multi-scale Feed-forward Layer (MFL), and an optional Global Context Attention (GCA) module.

**Top-k Token Selective Attention** The "Top-k Token Selective Attention Module" aims to minimize the interference caused by noisy tokens during self-attention calculation. It achieves this by identifying and selecting the top $k$ tokens with the highest relevance from the channel-wise attention matrix based on their significance to the query. With the query, key, and value matrices, each having dimensions denoted by $d \times H \times W$, where d represents

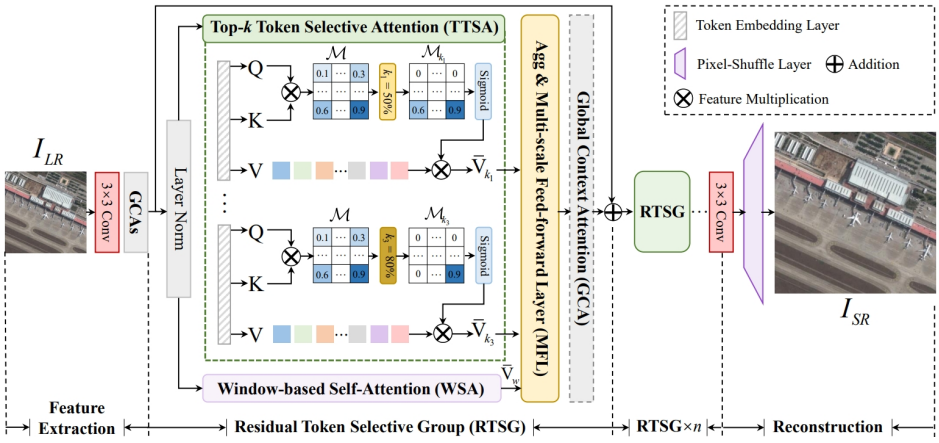



Figure 3.14: TTST's architecture [11]

the input dimension, the attention matrix is computed channel-wise through a standard dot-product of Q and the transposed K across channels. Once the channel-wise attention matrix is generated, an adaptive strategy masks lower attention values, preserving only the top-k elements with higher attention values. For instance, if $k$ is set to 1/4, only the top 25% of elements are activated, while the rest are masked to zero. This approach is made to work dynamically, as $k$ is set to a range of values, enabling a selective process that ranges from sparse to dense. After deriving attention matrices for the various $k$ values, each matrix is processed through an activation function and multiplication with the value matrix. The final output of this attention module is obtained by averaging these resulting matrices.

**Multi-Scale Feed-Forward Layer (MFL)** The authors of TTST have introduced the MFL module to explore multi-scale properties in remote sensing imagery. This multi-scale feed-forward layer is designed to enhance feature generation, replacing the conventional MLP layer and its linear projection method for feature propagation.
Figure 3.15 demonstrates the MFL process. After going through a layer normalization, the normalized feature is fed into three parallel branches. These branches are designed to explore multi-scale representations using 3×3, 5×5, and 7×7 depth-wise convolutions (DW-Conv), respectively. The multi-scale representations are divided into three segments along the channel dimension using a chunk operation to enhance the interaction among multi-scale localities further. These segments are then concatenated after being activated by ReLU activation. This approach is aimed at optimizing the integration and processing of features at various scales.

**Global Context Attention** Scenes often exhibit self-similarity and redundancy in large-scale remote sensing imagery, which serves as valuable prior knowledge for image restoration. To effectively utilize this information, the approach involves generating diverse global context features from extensive receptive fields using an adaptive selection process. This is achieved by strategically decomposing a large-scale convolutional kernel into a series of Depth-Wise Convolutions (DW-Conv) with varied kernel sizes. The kernel-decomposition strategy uses kernels of different sizes to gather important information while keeping the

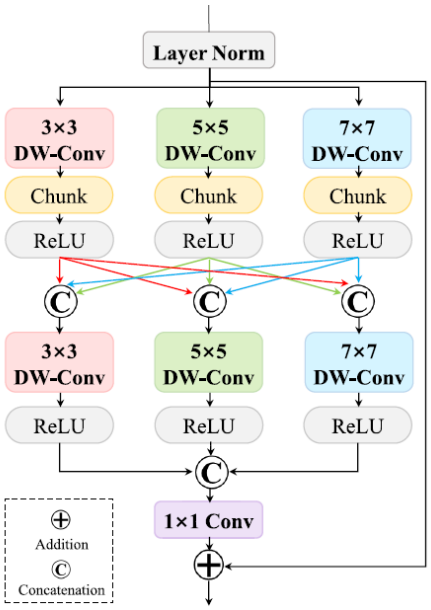



Figure 3.15: Multi-Scale Feed-Forward Layer [11]

model light and efficient. This approach is more resource-saving than using bigger kernels. It also improves the way images are restored by better-analyzing information from various scales and focusing on the most relevant details, leading to clearer and more accurate images.

## 3.5 Multispectral Super-resolution

Multispectral super-resolution refers to enhancing the spatial resolution of images captured in multiple spectral bands. Although commonly used for standard RGB images, the previous models can easily be modified to allow multispectral inputs. Another option is first to enhance the spatial resolution of each spectral band and then combine the high-resolution versions of each spectral band to create a multi-spectral super-resolved image through a fusion method.

Still, some researchers develop models specifically for multi-band remote sensing image super-resolution. MSAGAN[51] introduces the multiscale structure and attention mechanisms into a GAN network. The multiscale structure extracts features at different scales, and the attention mechanism motivates the model to pay attention to crucial high-frequency information. Also, a real multispectral super-resolution dataset is constructed from Landsat-8[15] and Sentinel-2[52] images. Several widely used satellite imagers record spectral bands with different spatial resolutions (like Sentinel-2 [52]), as such, some researchers focus on super-resolving lower-resolution bands to the exact resolution as the higher-spatial-resolution band by capturing the correlations between bands, as [53] and [54].

In hyperspectral imaging, enhancing the spectral resolution is an important aspect; on the contrary, the spectral super-resolution of multispectral images is usually not a primary concern as this imagery focuses on capturing information on critical spectral bands relevant to specific applications.

## 3.6 Precision Agriculture and the use of UAV and Satellite Image for Crop Health Monitoring

Precision Agriculture is defined by International Society of Precision Agriculture (ISPA) as the ”management strategy that gathers, processes and analyzes temporal, spatial and individual plant and animal data and combines it with other information to support management decisions according to estimated variability for improved resource use efficiency, productivity, quality, profitability and sustainability of agricultural production” [55].

As stated in the definition, one of the primary goals of this methodology is to enhance resource efficiency and minimize decision uncertainty in managing farm variability. Numerous technologies are employed for this approach to be successful, such as the Global Positioning System (GPS), which allows the gathering of data with accurate location information in real-time, Geographic Information Systems (GIS), computer-based tools used to store, visualize, analyze, and interpret geographic data, and remote sensing data.

This methodology is a cyclic procedure comprising three sequential stages: gathering information about variability, manipulating and analyzing data to evaluate the significance of the variability, and implementing changes needed in the management of inputs. [56].

Remote sensing technology enables the non-destructive acquisition of information concerning the Earth’s surface, potentially facilitating the implementation of Precision Agriculture. Through remote sensing instruments such as cameras, laser scanners, and sensors, it is possible to identify and distinguish various aspects of current crop conditions, including maturity period, and detect crop stresses such as nutrient and water stress, disease, pest, and weed infestations. [57]

To acquire remote sensing data, UAVs have gained significant popularity in recent years due to their ability to provide an adaptable and cost-effective way of obtaining high-resolution (centimeters scale) images necessary for precision agriculture applications. On the other hand, while data obtained via satellite encounters restrictions, such as limitations imposed by cloud cover and restricted flexibility in on-demand imaging solutions, notable enhancements in satellite sensors’ spatial, spectral, and temporal resolution have arisen. Additionally, the high volume of satellite data has prompted researchers to explore cutting-edge techniques for processing and storing data, such as cloud computing and machine learning.

Vegetation indices offer valuable insights into a wide range of precision agriculture techniques. They provide quantitative information on crop growth and health and can be calculated using information in multispectral bands obtained by remote sensing instruments.

Some of these indices include the NDVI, typically employed to quantify vegetation greenness and health; the Enhanced Vegetation Index (EVI), which extends beyond NDVI by addressing atmospheric interference and soil background reflectance issues; the Soil Adjusted Vegetation Index (SAVI), which adjusts NDVI by integrating a soil brightness correction factor to enhance accuracy, particularly in regions with diverse soil reflectance properties; the Normalized Difference Red Edge (NDRE) for measuring the chlorophyll content and the Green Normalized Difference Vegetation Index (GNDVI), functioning similarly to NDVI but utilizing the green band instead of the red band, proving beneficial in areas with intensely

Table 3.1: Vegetation Indexes

| Index | Abbreviation | Equation |
|---|---|---|
| Normalized difference vegetation index | NDVI | $\frac{NIR-RED}{NIR+RED}$ |
| Enhanced vegetation index | EVI | $\frac{2.5\times(NIR-RED)}{NIR+6\times RED-7.5\times BLUE+1}$ |
| Soil-adjusted vegetation index | SAVI | $\frac{(NIR-RED)(1+L)}{NIR+RED+L}$ |
| Normalized difference red-edge index | NDRE | $\frac{NIR-REDEDGE}{NIR+REDEDGE}$ |
| Green-normalized difference vegetation index | GNDVI | $\frac{NIR-GREEN}{NIR+GREEN}$ |

green vegetation. The detailed formulations are provided in Table 3.1, where L denotes the soil brightness correction factor in the SAVI formula.

## 3.7 Summary

In the early stages, CNNs have rapidly advanced the field of image super-resolution. Models based on visual transformers have emerged to surpass the limitations of CNNs.
Precision agriculture, characterized by its objective to optimize resource efficiency and reduce decision uncertainty in farm management, benefits immensely from these technological innovations. By harnessing the combined power of image super-resolution and multispectral imagery, precision agriculture gains access to a wealth of valuable insights. Multispectral imagery facilitates the assessment of indices such as NDVI, EVI, and NDRE, which serve as quantitative indicators of crop growth and health. These metrics are the backbone of a wide range of precision agriculture techniques, facilitating informed decision-making processes across various aspects of farm management. From crop monitoring and irrigation management to pest and disease control, precision agriculture practitioners rely on these insights to drive efficient farming practices and sustainable agricultural endeavors. Precision agriculture strives to optimize resource allocation, enhance productivity, and foster environmentally sustainable farming practices by integrating image super-resolution and multispectral imagery.

# Chapter 4

# Proposed Method

## 4.1 Introduction

The revision of the state-of-the-art provided in Chapter 3 reveals that traditional super-resolution techniques have evolved significantly with the advances in deep learning, particularly with Convolutional Neural Networks (CNNs) and, more recently, Transformer-based models. While CNNs are capable of enhancing image resolution through their ability to capture local features, they often struggle with modelling long-range dependencies and contextual information, which are crucial for accurately reconstructing high-resolution images from low-resolution inputs. Transformers address this problem through the use of the self-attention mechanism, which allows transformers to capture global dependencies across the entire image, making them more effective in understanding and reconstructing intricate details that span wide areas. The capability to process and integrate information from different parts of an image gives transformers a distinct advantage over CNNs.

Despite these advances, original Visual Transformers fail to capture fine details at lower levels, abstract concepts at higher levels, and understand the image's structure. To address these problems, the Swin Transformer introduces the concept of non-overlapping local windows, where self-attention is calculated within each window. This window-based approach enables the model to maintain a balance between capturing fine-grained details and preserving computational efficiency. By shifting the windows between layers, the Swin [5] Transformer ensures that cross-window interactions are also modelled, enhancing the overall contextual understanding and improving the quality of the super-resolved images.

Among the window attention transformers, the Top-k Token Selective Transformer [11] stands out among transformer-based models for its innovative approach to handle remote sensing data. TTST[11] employs a selective attention mechanism that focuses on the most relevant parts of the input data, effectively reducing redundancy and enhancing computational efficiency. This selective focus ensures that critical features and textures necessary for accurate image reconstruction are prioritized, making TTST[11] particularly well-suited for super-resolution tasks in remote sensing applications.

Considering this, we build on TTST to propose a novel approach for improving the image reconstruction quality in multispectral satellite data of agricultural fields. This chapter provides the details of the proposed approach.

## 4.2 Methodology

### 4.2.1 Background

Super-resolution techniques have advanced significantly with the advent of transformers, which utilize attention mechanisms to enhance image quality by evaluating pixel similarities to generate higher-resolution images. A key component of this approach is the attention window, which determines the pixels that can contribute to the attention calculation for each pixel's new representation. Current state-of-the-art methods investigate various window definitions and attention calculation methods: SwinIR [5] employs a square window, ART [8] uses a fixed sparse window, CAT [7] integrates rectangular windows, and HAT [9] implements channel-wise attention, which is further refined in TTST [11] with a top-k attention selection method. These methods rely on pre-defined attention windows, lacking input data context. The proposed method aims to use content-aware attention windows, reducing the number of pixels involved in the attention calculation for each pixel. Although having the same objective of masking out irrelevant data as TTST[11], which uses the top-k channel-wise technique after computing dense attention, the proposed method clusters pixels spatially first to allow for a more content-sensitive attention calculation.
In contrast to attention maps derived from predefined spatial windows, which consider all pixels within the window, this approach aims to provide a more meaningful representation of the data. By clustering similar pixels together, it selectively enables those in the same group to contribute to new pixel representations, discarding those that lack similarity.
Moreover, this method allows pixels to be grouped based on similarity rather than spatial proximity. This means that even if two pixels are far apart in the image, they can still influence each other's representation if they share similar characteristics. This contrasts with traditional methods that only consider pixels within a predefined spatial window, thereby ignoring distant but similar pixels.

### 4.2.2 Proposed Methodology

Intending to refine the attention mechanism's effectiveness, the proposed method redefines how attention windows are created, prioritizing pixel similarity to enhance the relevance of considered information. This involves two primary steps: clustering pixels based on similarities and applying attention mechanisms to these clusters.
In the first step, all pixels from the image are input to a clustering algorithm. Based on pixel similarity, the algorithm outputs a mask with the same width and height as the input image, defining the groups to which the pixels belong. In the following step, attention is calculated. The attention module will receive the input tokens and the mask from the first step, exclusively computing attention between tokens within the same group.
The proposed method comprises two versions. In the first version, illustrated by the diagram in Figure 4.1, the input image, with the shape of $H \times W \times C$, where $H$ represents the image height, $W$ the image width and $C$ the number of channels, is input to the clustering and convolution modules. In the clustering module, a clustering algorithm categorizes the pixels of the image into groups based on a similarity measure. It then generates a mask where each

position corresponds to the cluster assigned to the pixel at the same position in the image. In parallel, following the standard procedure of transformer-based models, the image is processed by a convolutional layer to be transformed into a higher-dimensional representation (represented by $C_{dim}$). The transformed image and mask are then passed to the attention module, where the attention mechanism is adapted to perform intra-cluster attention. This means that only pixels within the same cluster contribute to the new representations of the pixels in that cluster. The new representations of the images will go through the attention module $N$ times, producing more refined versions of themselves. Finally, the last representations of the image will be input to the traditional upsampling module and will be upsampled with pixel-shuffle [4]. The final image will have the shape of $(H \times scale) \times (W \times scale) \times C$, where scale defines the upsampling value.

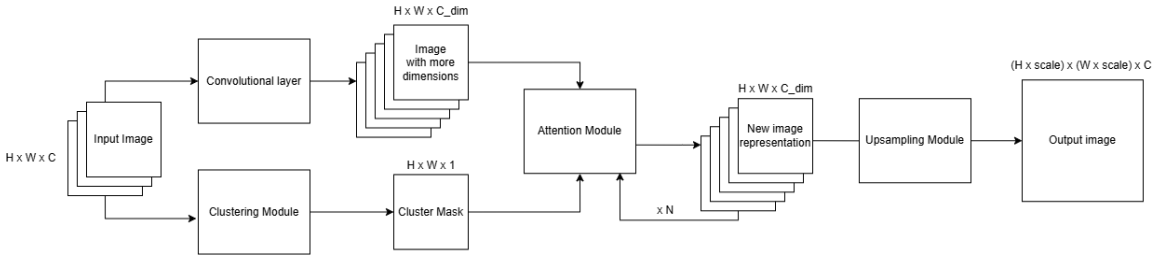



Figure 4.1: First version of the proposed method

The second version leverages the idea that transforming the input image into a higher-dimensional representation allows the model to extract richer and more detailed features, thereby enhancing its ability to capture complex patterns and relationships within the image. As shown in Figure 4.2, instead of directly inputting the original image into the clustering algorithm, the image first passes through a convolutional layer to increase its dimensionality before entering the clustering algorithm. The rest of the model remains unchanged.

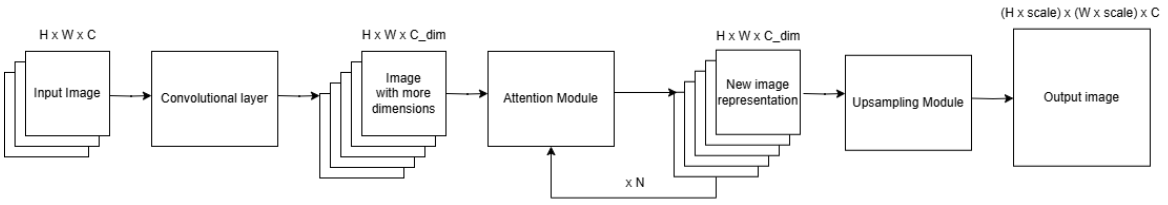



Figure 4.2: Second version of the proposed method

## 4.3 Conclusion

While the state-of-the-art in super-resolution continues to advance, certain strategies, despite their benefits, may impose limitations on model performance. By designing windows based on content rather than exclusively on spatial aspects, more pertinent information can be incorporated into the attention modules, facilitating the capture of long-range dependencies. The proposed method introduces a different approach to defining attention windows, with the objective of optimizing the efficiency of the attention mechanism. By prioritizing

content-driven window design, the proposed method aims to reveal new potential for super-resolution models, paving the way for more effective and sophisticated image improvement methodologies.

# Chapter 5

# Experiments and Results

## 5.1 Introduction

In this chapter, state-of-the-art super-resolution methods are comprehensively evaluated using two distinct datasets. The first dataset is derived from a real-world scenario comprising satellite and UAV imagery, while the second dataset is widely used in remote-sensing super-resolution but only comprises data acquired from satellites.
Additionally, this chapter details the experiments conducted to evaluate the efficacy of the proposed super-resolution model. A series of experiments were designed and executed to validate the efficiency of this new approach. These experiments assess not only the proposed method's performance using common super-resolution metrics, like PSNR and SSIM, but also its applicability in real-world agricultural scenarios. Specifically, the model's ability to monitor crop health using the NDVI is examined, highlighting the practical implications of deploying super-resolution methods in such settings.

## 5.2 Datasets

This section provides a detailed explanation of the two datasets employed in the experiments. The first dataset features images from a real-world scenario, where both high-resolution and low-resolution images are captured using remote-sensing devices. The second dataset, a standard and widely recognized dataset, is commonly used in remote-sensing super-resolution tasks.

### 5.2.1 Real-world Scenario Dataset

The first dataset used in the experiments comprises data obtained in a real-world scenario. It combines low-resolution satellite images from Sentinel-2 with high-resolution images captured by a drone. Both types of images are taken from the same geographical location and on the same days. The dataset consists of multispectral "geotiff" orthophotomap pairs containing embedded geographical information, capturing various growth stages of a vineyard. Each image comprises five channels: Red, Green, Blue, Near Infrared, and Red Edge.
The dataset features images captured on six different dates spread over six months: April 4th, May 16th, August 10th and 17th, and September 7th and 25th. This timeline corresponds to six distinct pairs of orthophotomaps, each representing a specific moment in time of the vineyard's development.
These orthophotomaps are cropped into patches for training. The resulting dataset, derived from the orthophotomaps, includes both low-resolution and high-resolution patches, each

characterized by distinct spatial resolutions and sizes. The low-resolution patches, with a spatial resolution of 10 meters per pixel and a size of 16x16 pixels, provide a significantly degraded view of the vineyard. In contrast, the high-resolution patches offer a more detailed perspective with a spatial resolution of 0.625 meters per pixel and a size of 256x256 pixels. This difference in resolution defines a super-resolution task with a scaling factor of 16, highlighting the need for significant improvement in image detail.
Following the cropping process, the dataset includes 7,590 pairs of images for each date. Each patch is systematically named using a column and row system that indicates its spatial position on the orthophotomap, with rows numbered from 0 to 114 and columns from 0 to 65. Figure 5.1 illustrates the RGB versions of the image pairs from each date, displaying the high-resolution images on the left and the low-resolution versions on the right.

(a) April 4th (d) May 16th

(b) August 10th (e) August 17

(c) September 7th (f) September 25th

Figure 5.1: Patches from the same geographical place on different dates.

### 5.2.2 Crafted Resolution Dataset

Multi-scenario datasets are commonly used in remote sensing tasks. However, there is a lack of state-of-the-art datasets specifically tailored for agricultural crops. One widely used

dataset for general remote-sensing applications is the Aerial Image Dataset (AID), compiled from Google Earth imagery. AIDs's diverse collection of images from various remote sensing devices makes it a challenging multi-source dataset for computer vision tasks. This dataset includes 10,000 RGB images categorized into 30 distinct aerial scene types: airports, farmland, urban areas, forests, and other scenarios. For the specific task of super-resolution in remote sensing, high-resolution images from datasets like AID are typically used. These images are processed using MATLAB's bicubic interpolation to generate corresponding low-resolution pairs.
Given the absence of specialized datasets for evaluating super-resolution in agricultural contexts, AID, as a widely recognized dataset, was also used for some experiments to give an idea of the performance of state-of-the-art models. Figure 5.2 illustrates examples from some AID classes.

(a) Farmland (b) Forest (c) Sparse Residencial

Figure 5.2: Representation of different classes of AID

## 5.3 Performance Evaluation

Super-resolution model performance evaluation is generally done through the use of performance metrics such as the PSNR and the SSIM [58].
PSNR measures the ratio between a signal's maximum power and the power of the signal's noise in Decibel (dB). PSNR is defined by equation 5.1, where MAX denotes the maximum possible pixel value of the image (255 if pixels are represented with 8 bits per sample) and Mean Squared Error (MSE) represents the mean-squared-error function. The PSNR can range from 0 to infinity. It reaches infinity when there is zero mean squared error between the original and reconstructed images, indicating no noise and perfect signal reconstruction.

$$PSNR = 10.\log_{10}\frac{MAX^2}{MSE} \quad (5.1)$$

SSIM was introduced in [58] as an index for measuring the similarity between two images, considering three aspects of image quality: luminance, contrast, and structural similarity. For images with more than one channel, the SSIM is calculated per channel and then averaged to obtain the final value for the entire image. SSIM ranges from -1 to 1, with 1 indicating the highest level of image similarity.

- Luminance represents the average brightness (intensity) of the image of the pixels in the image. For two images $x$ and $y$, the luminance comparison is defined as in Equation 5.2 where $\mu$ is the average brightness value for the corresponding image.

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (5.2)$$

- Contrast quantifies the difference in intensity between pixels (how much the intensities spread in an image), which is determined using the standard deviation of pixel intensities. The contrast comparison of $x$ and $y$ is given by Equation 5.3, where $\sigma$ represents the standard deviation of intensities (square root of variance) over the corresponding image.

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (5.3)$$

- Structural similarity represents the correlation between the pixel intensities of the two images. For its calculation, the covariance of pixel intensities is used. The equation for the structural similarity comparison calculation is 5.4 where $\sigma_{xy}$ represents the covariance of intensities between the two images:

$$s(x,y) = \frac{2\sigma_{xy} + C_3}{\sigma_x + \sigma_y + C_3} \quad (5.4)$$

  C1, C2, and C3 are constants defined to avoid instability when the denominator is close to 0.

The final SSIM is calculated as the product of the three components, and the final simplified equation is given by 5.5.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)((2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (5.5)$$

Given that this task concerns remote-sensing super-resolution, with a specific focus on crop health monitoring, an additional performance evaluation step in the experiments employs one of the most widely used indexes for this purpose: the NDVI, given by Equation 5.6). The NDVI uses values from channel red and channel near-infrared (NIR) from an image. To assess performance, the additional evaluation will compare the NDVI values from the original images with those derived from the super-resolved images.

$$\frac{NIR - RED}{NIR + RED} \quad (5.6)$$

## 5.4 Experiments

The experimental study was structured into three distinct phases:

- The first phase evaluates state-of-the-art models on a real-world scenario dataset and different methods of splitting the vineyard dataset into training, validation, and test sets. Four experiments were conducted for this purpose.
- The second phase evaluates state-of-the-art models using a widely recognized remote-sensing dataset.
- The third phase evaluates the proposed method, involving five experiments.

### 5.4.1 First Experimental Study

This first phase of the experimental study aims to identify the state-of-the-art models that perform best in real-world conditions and determine which dataset characteristics are most effective for training, validation, and testing in agricultural imaging.
Four experiments were conducted with this objective.

- The first experiment involves splitting the train, validation and test sets by considering the spatial characteristics of the images. In this stage, images from April 4th, May 16th, August 10th, August 17th, and September 7th were used, and images from September 25th were excluded. Considering the naming system explained in the previous section, images from columns 20 to 30 were used for validation, images from columns 55 to 65 were assigned to the test set, and the remaining were used for training. This strategy was designed to test the model's performance on previously unseen land. In this experiment, all crop growth stages are represented in the three sets. However, spatially, the datasets consist of distinct, non-overlapping areas of land;
- The second experiment is designed to evaluate the performance of the model developed in the first experiment on a new test set representing a different crop growth stage. Since the images from September 25th were not used during the first experiment, the test set for this experiment was composed of these images;
- The third experiment aims to assess how certain crop growth stages in the training set affect the model's ability to generalize to unseen growth stages. To achieve this, the dataset was first narrowed to focus on agricultural land by removing areas primarily containing trees, roads, and ground. After this procedure, only rows 10 to 70 and columns 20 to 54 were retained, resulting in 2,415 images per date. The dataset was then divided by dates: images from April 4th, May 16th, September 7th, and September 25th were assigned to training, images from August 10th to validation, and images from August 17th to testing. This experiment determines whether using training images of crop growth stages that are further apart enables the model to generalize to the intermediate crop growth stage;

| | Datasets | PSNR | SSIM |
|---|---|---|---|
| SwinIR | Experiment 1 | 28.196 | 0.748 |
| | Experiment 2 | 23.955 | 0.510 |
| | Experiment 3 | 25.066 | 0.488 |
| | Experiment 4 | 24.494 | 0.529 |
| ART | Experiment 1 | 29.177 | 0.730 |
| | Experiment 2 | 24.232 | 0.498 |
| | Experiment 3 | 25.424 | 0.542 |
| | Experiment 4 | 24.719 | 0.565 |
| HAT | Experiment 1 | 28.198 | 0.736 |
| | Experiment 2 | 24.697 | 0.501 |
| | Experiment 3 | 25.159 | 0.484 |
| | Experiment 4 | 24.160 | 0.519 |
| CAT-R | Experiment 1 | 28.341 | 0.761 |
| | Experiment 2 | 24.030 | 0.498 |
| | Experiment 3 | 24.742 | 0.469 |
| | Experiment 4 | 24.323 | 0.504 |
| CAT-A | Experiment 1 | 28.465 | 0.753 |
| | Experiment 2 | 24.091 | 0.492 |
| | Experiment 3 | 25.148 | 0.536 |
| | Experiment 4 | 24.761 | 0.561 |
| HAUNET | Experiment 1 | 24.857 | 0.516 |
| | Experiment 2 | 25.787 | 0.575 |
| | Experiment 3 | 24.740 | 0.411 |
| | Experiment 4 | 24.410 | 0.501 |
| TTST | Experiment 1 | 28.458 | 0.743 |
| | Experiment 2 | 24.608 | 0.524 |
| | Experiment 3 | 25.556 | 0.587 |
| | Experiment 4 | 24.744 | 0.567 |

Table 5.1: Results of PSNR and SSIM from the first experimental study

- The fourth experiment shares the same objective as the third but with a modification in the division of the datasets. This experiment explores whether training images from dates closer together are more beneficial than images representing a wider range of crop growth stages. In this experiment, images from April 4th, May 16th, August 17th, and September 25th are used for training, images from August 10th are used in validation, and images from September 7th are used in testing.

Figure 5.3 visually represents the different data splits from each experiment's train, validation, and test datasets.
The difference between the geographical area used in Experiments 1 and 2 and the reduced version in Experiments 3 and 4 is presented in Figure 5.4, where both the low-resolution and high-resolution othophotomaps are presented on the date of August 10th.

For each experiment, state-of-the-art models SwinIR [5], ART[8], HAT[9], CAT[7] (with its two versions CAT-A and CAT-R), HAUNET[10], and TTST[11], were trained and evaluated on the corresponding test datasets using two metrics, the overall mean PSNR and SSIM reported on the test datasets, which is the standard evaluation procedure of super-resolution methods. These values are reported in 5.1.

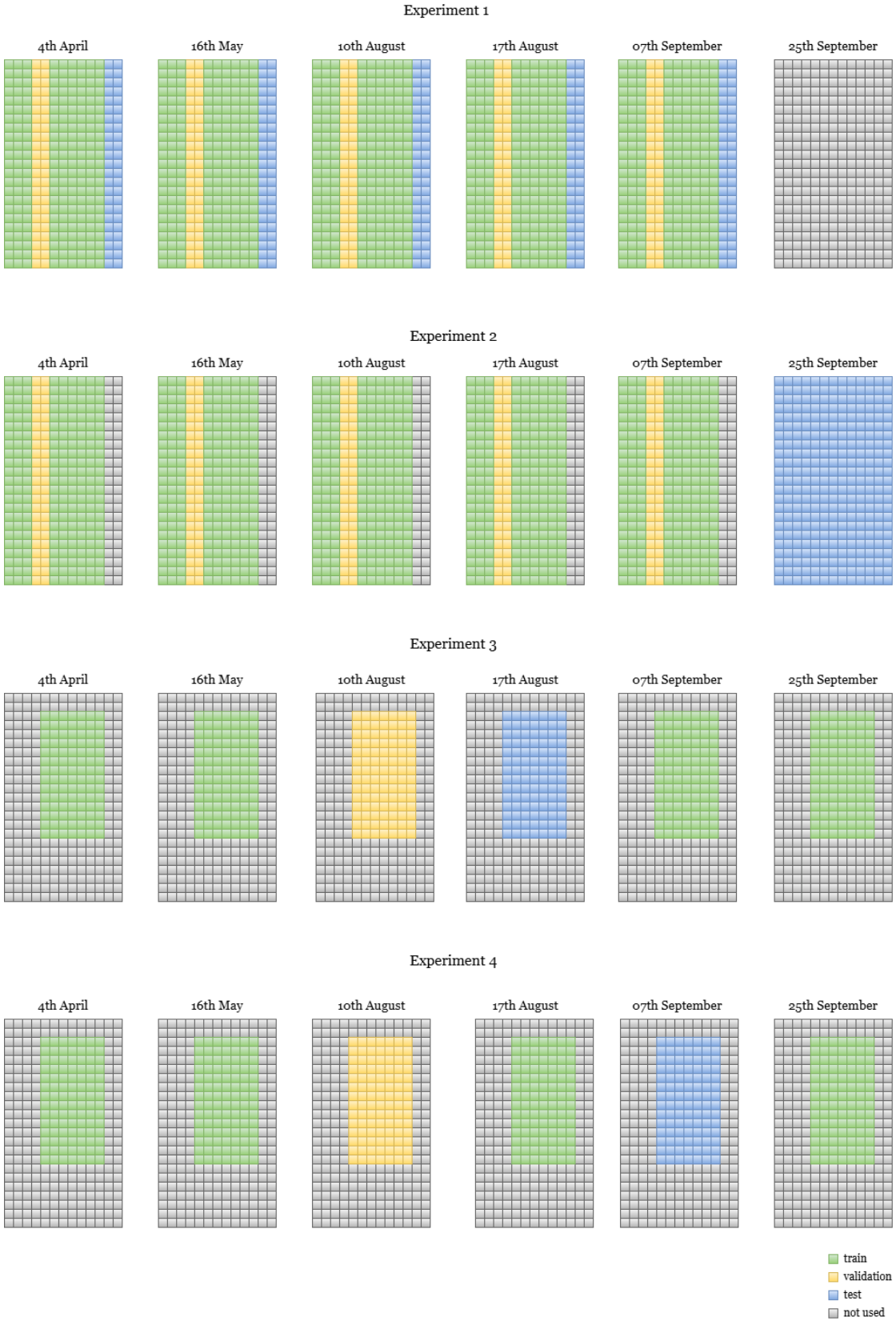
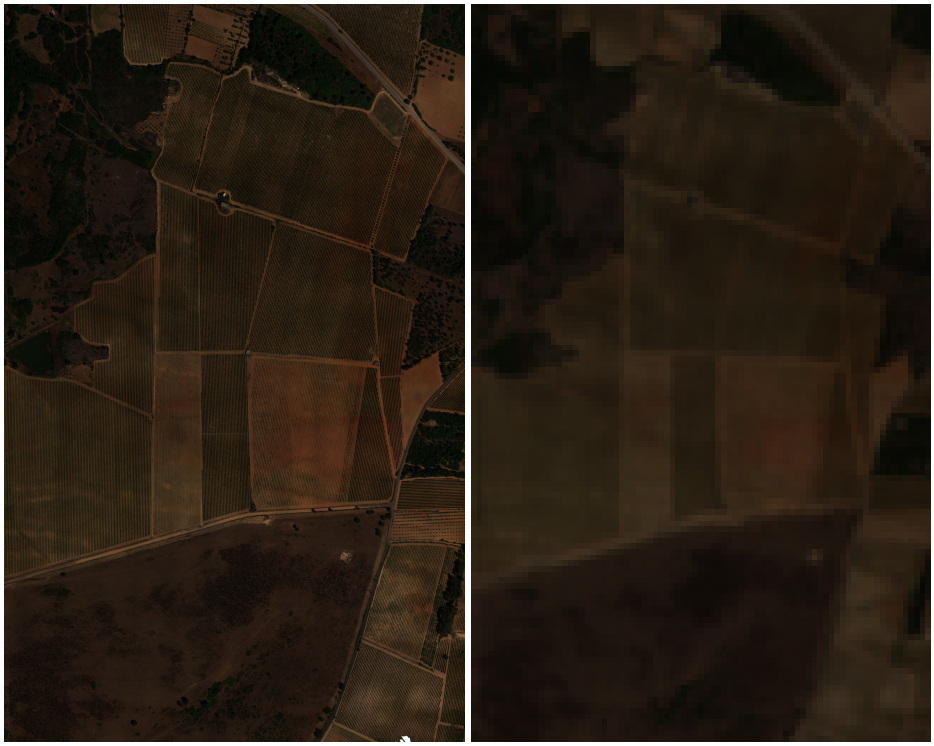



Figure 5.3: Dataset splits of the different experiments. Each grid represents the orthophotomap for a specific date. The cells within each grid indicate the cropped patches used for training (green), validation (yellow), and testing (blue). The patches not used are represented in gray.



(a) Complete orthophotomap (Experiment 1 and 2)

(b) Reduced orthophotomap (Experiment 3 and 4)

Figure 5.4: Orthomaps from the vineyard

| | Datasets | PSNR | SSIM |
|---|---|---|---|
| SwinIR | Experiment 1 | 28.196 | 0.748 |
| | Experiment 2 | 23.955 | 0.510 |
| | Experiment 3 | 25.066 | 0.488 |
| | Experiment 4 | 24.494 | 0.529 |
| ART | Experiment 1 | 29.177 | 0.730 |
| | Experiment 2 | 24.232 | 0.498 |
| | Experiment 3 | 25.424 | 0.542 |
| | Experiment 4 | 24.719 | 0.565 |
| HAT | Experiment 1 | 28.198 | 0.736 |
| | Experiment 2 | 24.697 | 0.501 |
| | Experiment 3 | 25.159 | 0.484 |
| | Experiment 4 | 24.160 | 0.519 |
| CAT-R | Experiment 1 | 28.341 | 0.761 |
| | Experiment 2 | 24.030 | 0.498 |
| | Experiment 3 | 24.742 | 0.469 |
| | Experiment 4 | 24.323 | 0.504 |
| CAT-A | Experiment 1 | 28.465 | 0.753 |
| | Experiment 2 | 24.091 | 0.492 |
| | Experiment 3 | 25.148 | 0.536 |
| | Experiment 4 | 24.761 | 0.561 |
| HAUNET | Experiment 1 | 24.857 | 0.516 |
| | Experiment 2 | 25.787 | 0.575 |
| | Experiment 3 | 24.740 | 0.411 |
| | Experiment 4 | 24.410 | 0.501 |
| TTST | Experiment 1 | 28.458 | 0.743 |
| | Experiment 2 | 24.608 | 0.524 |
| | Experiment 3 | 25.556 | 0.587 |
| | Experiment 4 | 24.744 | 0.567 |

Table 5.2: Results of PSNR and SSIM from the first experimental study

Several can be derived from the results provided in Table 5.1. First, models have performed significantly better in the first experiment. Second, their performance drops drastically in experiment two, which regards the evaluation of the model on a new dataset with an unseen crop growth stage. In this context, the models transition from being the best-performing across all experiments to being worst in 5 out of 7 models in terms of PSNR and the worst in 4 out of 7 models in terms of SSIM.
It transitions from being one of the best-performing models across all experiments to the worst in 5 out of 7 models in terms of PSNR and the worst in 4 out of 7 models in terms of SSIM.
The results from the first two experiments indicate that using data from all growth stages in the training sets while only spatially splitting the data results in a biased model that cannot be generalized to unseen crop growth stages.
Regarding the data division based on dates, experiments 3 and 4 present divergent results in terms of PSNR and SSIM. Experiment 3 consistently shows higher PSNR across all tests, while experiment 4 exhibits the opposite trend, with better SSIM but lower PSNR.

Given that this study's overall objective is to evaluate the feasibility of using super-resolution models for crop health monitoring, achieving the closest possible pixel values between the high-resolution image and the super-resolution image is crucial. Therefore, the PSNR is a

| | PSNR | | SSIM | |
|---|---|---|---|---|
| | All Classes | Farmland | All Classes | Farmland |
| SWINIR | 29.264 | 33.697 | 0.784 | 0.849 |
| ART | 29.211 | 33.632 | 0.787 | 0.847 |
| HAT | 29.315 | 33.775 | 0.787 | 0.852 |
| CAT-R | 29.257 | 33.712 | 0.785 | 0.850 |
| CAT-A | 29.299 | 33.745 | 0.786 | 0.851 |
| HAUNET | 22.185 | 23.538 | 0.695 | 0.783 |
| TTST | 29.155 | 33.616 | 0.784 | 0.850 |

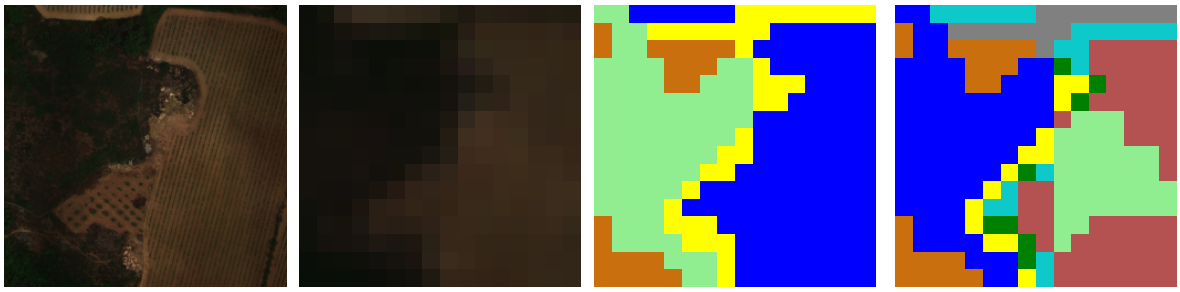
Table 5.3: Results for the AID Dataset

more critical metric than the SSIM for this study. Based on this criterion, the dataset split used in Experiment 3 is the most suitable for this problem. Consequently, this dataset will be utilized in the proposed method experiments.

ART[8], CAT-A[7], and TTST[11] emerged among the evaluated models as the top performers, validating the proposed hypothesis. The ART model employs a sparse attention window with fixed positions, which allows it to cover a wide spatial area and enhance its attention mechanism. CAT-A, on the other hand, uses an attention window that is shaped like a strip along the axial direction by fixing one side at the image size. This configuration enables it to model long-range dependencies effectively. Although the TTST[11] model does not employ a specialized attention window, it stands out by applying channel-wise attention across the entire image and using an algorithm to ignore less relevant tokens based on their attention scores selectively. These results reinforce the idea that using a content-aware attention window based on token similarity can significantly increase model performance, especially in this remote-sensing super-resolution task.
Since the TTST[11] model achieved the highest performance, it will be used as the baseline model for experiments involving the proposed model.

### 5.4.2 Second Experimental Study

As previously stated, the second experimental study seeks to evaluate the state-of-the-art models on a widely known remote-sensing super-resolution dataset. The AID dataset was used in this experiment, from which 3000 images were randomly chosen from training (100 from each class), 300 for validation (10 from each class), and 900 for testing (30 from each class). The same metrics as before were used. The results are presented in Table 5.3, which reports the overall PSNR and SSIM scores for the test set, along with the specific metrics for the Farmland class.
The evaluation of state-of-the-art models on the AID dataset performs better than the vineyard dataset. This outcome is expected, as the vineyard dataset presents a more challenging task as it represents a real-world scenario with a significant scale of downsampling (x16) between high-resolution and low-resolution images. In contrast, the AID dataset has a lower scale factor (x4), where the low-resolution images are generated using MATLAB's bicubic interpolation function, which will typically retain more image quality.



(a) HR (b) LR (c) 4 clusters (d) 8 clusters

Figure 5.5: Clustering performed by the BIRCH algorithm, when taking the lower resolution image as input

In terms of PSNR, the best-performing models overall were HAT[9] and CAT[7] for the overall classes as well as the Farmland class specifically. This suggests that the channel attention module in HAT[9] can positively impact training with certain datasets. Additionally, all classes benefit from the different windows definitions of the two CAT model versions.
However, in terms of SSIM, the difference between the models is not significant, except for HAUNET[10], which performs poorly. This lack of significant difference may be explained by the fact that images might have less variability in structural content, making it harder to distinguish performance variations among the better-performing models.

### 5.4.3 Third Experimental Study

The third experimental study assesses the performance of the proposed methodology. Considering that the proposed approach depends on the clustering of image feature maps, an efficient and fast clustering algorithm capable of handling high-dimensional, large-scale data was selected. After considering different clustering methods, the BIRCH was chosen as the most appropriate method. BIRCH is a hierarchical clustering algorithm that efficiently handles large datasets by first generating a small, compact summary that retains as much information as possible. This summary is then used for clustering instead of the entire dataset. BIRCH constructs a hierarchical tree structure known as the Clustering Feature tree (CF Tree) during a single data scan, allowing it to handle large datasets efficiently. The CF Tree incrementally and dynamically clusters incoming multi-dimensional data points. This means that as new data points are added, the CF Tree structure is updated dynamically without reprocessing the entire dataset. This incremental approach ensures that BIRCH can handle large-scale datasets efficiently, providing high scalability. The hierarchical approach makes BIRCH more efficient and scalable than many traditional clustering methods. For example, KMeans clusters the entire dataset directly and often requires multiple passes over the data, making it less efficient for very large datasets.
Figure 5.5 shows an example of how BIRCH clusters pixels from a low-resolution image from the vineyard dataset 5.5b. Although this image is depicted in the RGB format, it actually has 5 channels, as stated in the dataset section. All channels were used for clustering.
Figure 5.5c shows the result of the clustering algorithm when the number of clusters is set to four. We can observe that most of the farmland is clustered together (in blue), with a separate cluster for areas likely shared between farmland and ground (in yellow). The remaining parts

| | TTST reduced | | 4 clusters | | | | | | 8 clusters | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 5-channel | | Increase dim. | | Filtered attention matrix | | Increase dim. | |
| | % | C.% | % | C.% | % | C.% | % | C.% | % | C.% |
| [0.0,0.1[ | 66.25 | 66.25 | 63.11 | 63.11 | 64.66 | 64.66 | 63.52 | 63.52 | 65.34 | 65.34 |
| [0.1,0.2[ | 24.27 | 90.52 | 26.21 | 89.32 | 25.14 | 89.80 | 25.99 | 89.51 | 24.16 | 89.50 |
| [0.2,0.3[ | 7.21 | 97.73 | 7.42 | 96.74 | 7.30 | 97.10 | 7.67 | 97.18 | 7.28 | 96.78 |
| [0.3,0.4[ | 1.95 | 99.68 | 2.91 | 99.65 | 2.61 | 99.71 | 2.49 | 99.67 | 2.87 | 99.65 |
| [0.4,0.5[ | 0.26 | 99.94 | 0.34 | 99.99 | 0.28 | 99.99 | 0.30 | 99.97 | 0.33 | 99.98 |

Table 5.4: Class distribution of NDVI error percentages across experimental models

of the image, corresponding to vegetation outside the farmland, are divided into two groups, reflecting their different characteristics. This clustering division seems logical and aligns with the expected distribution.

Figure 5.5d shows the result of the clustering algorithm defining 8 clusters. Although the two clusters associated with the area outside the farmland remain mostly unchanged, the area corresponding to the farmland is subdivided into 6 clusters. This means that the clustering algorithm found different similarity groups in this area.

As the objective is to create groups of pixels that share similarities, this strategy effectively highlights the varying characteristics within the patches of the dataset, allowing for a more detailed analysis of the different regions.

Five experiments were conducted in this phase. The architecture of the super-resolution models is composed of several layers, each containing all the attention modules suggested by the methods, iterated multiple times. To enable performing multiple experiments, the number of layers in the backbone architecture was reduced from the typical 36 to 6, significantly reducing computational complexity by a factor of six. Consequently, the best-performing model on the vineyard dataset experiments, TTST [11], was trained again with this reduced number of layers. This adjustment allows for effective comparative analysis across different experiments involving the proposed architecture. As stated previously, the dataset used in this section is the same as the one of Experiment 3 regarding the first experimental study, visually represented in Figure 5.3.

The following list details the three experiments conducted in this phase.

- The first experiment involves training and evaluating the TTST[11] model in a configuration with a reduced number of layers, as previously described;

- The second experiment evaluates the proposed methodology using the BIRCH algorithm for clustering. The two versions of the method are tested: the one that clusters the original input data directly and the second that processes the input through a convolutional layer to increase its dimensionality before clustering. Different dimensionality values were defined for the input of the clustering module and attention modules, with the first set to 30 and the second set to 180 as standard;

- The last experiment studies whether applying an attention mechanism, followed by filtering out lower values from the attention matrix, is more or less effective than employing clustering algorithms. This process is similar to the TTST[11] method, but it is performed spatially.

| | | PSNR | SSIM |
|---|---|---|---|
| TTST reduced baseline | | 25.853 | 0.520 |
| 4 clusters | 5-channel clustering | 25.899 | 0.478 |
| | Clustering after increase in dimensionality | 26.015 | 0.492 |
| | Filter the attention matrix | 25.700 | 0.494 |
| 8 clusters | Clustering after increase in dimensionality | 26.119 | 0.488 |

Table 5.5: Experiments for the evaluation of the proposed method

| | Q1 | Q2 | Q3 | 100% (without outliers) | 100% |
|---|---|---|---|---|---|
| TTST reduced | 0.0259 | 0.0670 | 0.1235 | 0.2699 | 0.9377 |
| 8 clusters with dimensionality inscrease | 0.0276 | 0.0725 | 0.1232 | 0.2666 | 0.9275 |
| 4 clusters with dimensionality increase | 0.0290 | 0.0720 | 0.1254 | 0.2700 | 0.9487 |
| 4 clusters on 5-channels | 0.0300 | 0.0751 | 0.1276 | 0.2740 | 0.9375 |
| Attention matrix filtered | 0.0271 | 0.0729 | 0.1292 | 0.2825 | 0.9345 |

Table 5.6: Summarized information about the distribution of the absolute differences between the values of NDVI produced by the different experimental models

In the second experiment, the number of clusters is defined as four, which corresponds to filtering out three-quarters of the data from the third experiment. Afterward, the configuration is adjusted to eight clusters for the best-performing version.

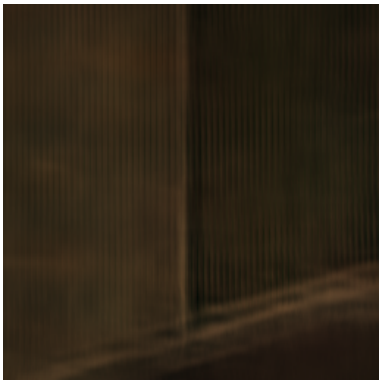
The results for the overall PSNR and SSIM reported in the test set of this experiment are synthesized in Table 5.5, where the first row regards the evaluation of the TTST[11] with the reduced number of layers, the second, third and fifth rows correspond to the proposed method and the fourth row assesses the study of the filtering of the spatial attention matrix based on attention scores.

In terms of PSNR, the experiment that achieved the highest value utilized clustering after increasing the dimensionality of the input. When this setup was trained again with more clusters, its performance improved even further. However, when evaluating SSIM, the score is not better than the reduced version of TTST[11]. This result can make sense, as clustering data with higher similarity prior to attention calculation means that the final attention matrix will only consider pixels within the same cluster. This localized attention can limit the model's ability to capture global structural similarities across different clusters, which is essential for maintaining higher SSIM scores. Figure 5.6 shows an output image from the different model experiments, along with the PSNR and SSIM scores for each.

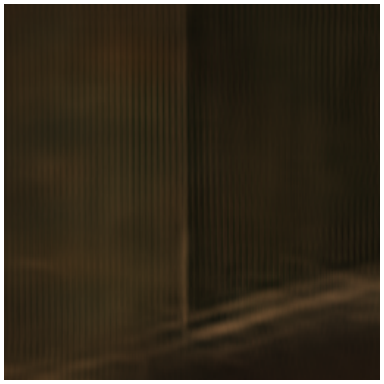
Given that the objective of this research is to determine if super-resolution models can be utilized for crop health monitoring via images obtained from remote-sensing devices, an additional evaluation procedure was conducted.

The NDVI was calculated for the pixels of all images in the test set. Then, the absolute differences between the NDVI values of the high-resolution pixels and the super-resolution pixels were determined. These difference values were categorized into classes (e.g., [0 - 0.1[, [0.1 - 0.2[) to assess which models produced a higher percentage of pixels with smaller differences. This analysis helps identify the models that best preserve the pixel values used to calculate the crop health index. The results are presented in Table 5.4. The "%" columns show the percentage of pixels for which the absolute difference in NDVI values falls within the corresponding

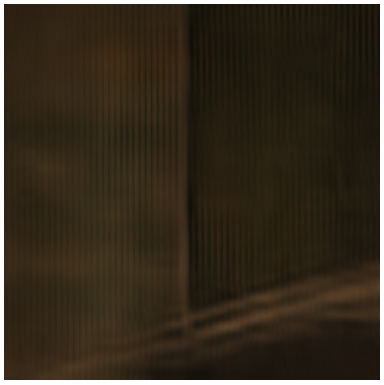
(a) Original Low-resolution image



(b) Model with the attention matrix filtered
PSNR: 28.215
SSIM: 0.670



(c) Model with clustering on 5 channels
PSNR: 28.475
SSIM: 0.712



(d) Clustering model on a higher dim. with 4 clusters
PSNR: 28.248
SSIM: 28.248

(e) Clustering model on a higher dim. with 8 clusters
PSNR: 28.750
SSIM: 0.681

(f) TTST model in the reduced version
PSNR: 28.082
SSIM: 0.712

(g) Original high-resolution image

Figure 5.6: Visual results of the proposed model.

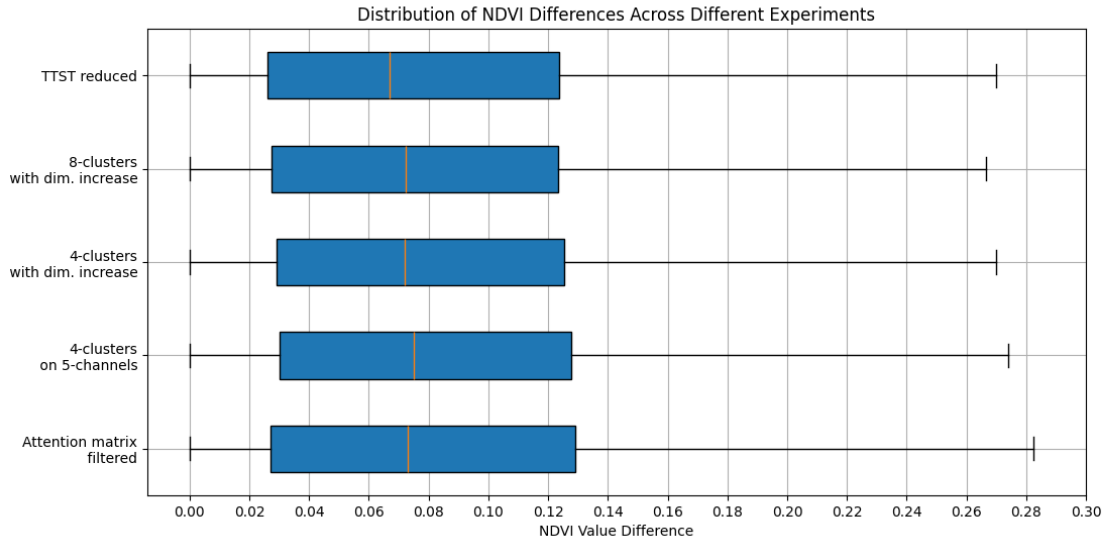



Figure 5.7: Boxplots of the absolute differences of NDVI values produced by the different experimental models

range. The ”C. %” column refers to the cumulative percentage. The boxplot regarding the distribution of the absolute differences of NDVI values was also plotted and presented in Figure 5.7. The median, quartiles, and maximum data values are summarized in Table 5.6. In the reduced TTST[11] model version, at least 50% of the data have an absolute difference of less than or equal to 0.0670 compared to the actual NDVI values. In contrast, in the best version of the proposed model (which clusters the data into eight clusters after increasing the dimensionality of the input data), this percentage has an absolute difference of less than or equal to 0.0725. Additionally, in the best version of the proposed model, at least 75% of the data have an error of less than or equal to 0.1232, whereas in the TTST version, the error is less than or equal to 0.1235. Furthermore, the maximum error is lower in the proposed model version, corresponding to about a 13% difference between the actual and predicted values of NDVI, indicating that this model is more reliable.

## 5.5 Conclusions

This chapter details the experimental process and evaluation of various state-of-the-art super-resolution models and the proposed model applied to remote-sensing datasets, with particular focus on agricultural applications. These experiments were strategically designed not only to benchmark the performance of established super-resolution approaches but also to assess their practical implications in monitoring crop health through NDVI analysis. The results obtained show that the recent advances in super-resolution methods assured notable improvements in the visual quality of super-resolved images. The TTST[11] model emerged as the best among state-of-the-art super-resolution models, demonstrating robustness specifically on the real-world condition dataset.

Furthermore, the effectiveness of the proposed model was particularly notable when eight clusters were used on the feature maps rather than directly on the 5-channel data. This ap-

proach proved more effective in several ways. Firstly, it achieved higher values of PSNR, indicating better image reconstruction quality compared to the experiments. Additionally, it demonstrated a lower maximum error in predicting the NDVI. Specifically, the maximum discrepancy between the actual NDVI values and those predicted by the model is about 13%, where at least half of the values present a difference of NDVI of less than or equal to 3.6%. This smaller error percentage underlines the model's improved accuracy in applications where precise crop health monitoring is critical.

# Chapter 6

# Conclusion and Future Work

## 6.1 Main Conclusions

This dissertation has explored the use of super-resolution techniques and their application in precision agriculture. Beginning with the fundamentals of image super-resolution, this dissertation then progresses to defining and evaluating state-of-the-art models in this field. The highlight of the study is the development of a customized super-resolution method designed specifically for agricultural use. This research was inspired by the fast growth of the global population and the urgent need for innovative agricultural practices to sustainably meet food demands while safeguarding our resources. Precision agriculture, supported by advanced super-resolution techniques, is a promising approach to effectively monitor and improve crop health and yield. The integration of multispectral imagery significantly improves the efficacy of the techniques by providing a way of assessing different metrics of crop health and provide richer data for analysis.

The primary objective of this research was to develop a super-resolution method capable of transforming low-resolution satellite images into high-resolution counterparts, comparable to UAV imagery. The thesis investigated numerous model architectures, with a focus on Transformer-based models and CNNs. This focus aligns with the emerging trend in state-of-the-art super-resolution technologies, where Transformers are increasingly recognized as a breakthrough innovation. A standout was the Top-k Token Selective Transformer (TTST)[11] model, which demonstrated potential in processing remote sensing data. By concentrating on the most relevant parts of the input, by masking out lower attention score values, this model increased both computational efficiency and accuracy. Extensive experiments confirmed the effectiveness of this approach, based on metrics such as the PSNR, SSIM and the NDVI, an index that measures crop health, to assess performance

Furthermore, this work builds on TTST to introduce a novel approach that relies on a clustering algorithm to constrain the attention mechanism to be carried solely within the pixels/features of each cluster.Among the different variants of the proposed approach, creating the cluster mask from feature maps attained the best performance, as it allows for better detail capture and improvement in image super-resolution quality. This approach leverages the strengths of clustering to enhance the attention mechanism, effectively filtering out less relevant data and focusing computational resources on the most critical parts of the image. By increasing the input dimensionality first, the model is better equipped to identify and preserve intricate details, which translates to higher quality super-resolved images. This clustering-based method not only improves the visual quality of the images but also improves the overall accuracy of crop health monitoring, measured by the NDVI, providing a more reliable tool for precision agriculture.

The results obtained suggest that the proposed approach can effectively determine plant health using satellite imagery, as the estimated NDVI from the super-resolved data differs only by a maximum of 13% from the values derived from UAV-acquired data, where at least , where at least half of the values present a difference of NDVI of less than or equal to 3.6%. By reliably improving the resolution of satellite images through super-resolution, this research has made several notable contributions to the field of precision agriculture. The developed super-resolution method enables more accurate monitoring of crop health, aiding in the early detection of issues such as disease outbreaks or inadequate irrigation. The detailed spatial information provided by super-resolution images facilitates better resource management, allowing for the precise application of water, fertilizers, and pesticides, thus promoting sustainable agricultural practices. Additionally, the method's reliance on satellite data ensures scalability, making it feasible to monitor large agricultural areas cost-effectively, which is crucial for widespread adoption in various agricultural regions.

## 6.2 Future Work

Building on the promising results and contributions outlined in this dissertation, several directions for future work can be identified to further enhance and expand the application of super-resolution techniques in precision agriculture.

While the experiments conducted in this research demonstrated impressive performance, an essential next step is to evaluate the impact of these super-resolution results on the practical aspects of vineyard management, specifically focusing on phytosanitary (plant health) conditions. Although the current experiments provided a strong foundation in terms of image quality and technical metrics, it is crucial to assess how these super-resolved images translate into actionable insights for farmers.

Future work should include comprehensive field trials to evaluate the real-world effectiveness of the super-resolution method in improving plant health monitoring. This involves assessing how well the images obtained from super-resolution help in identifying and managing diseases, pests, and other stress factors in vineyards.

Another significant area for future work is the implementation of the super-resolution models as a web-based service. Developing a scalable cloud platform will greatly enhance accessibility for vineyard managers and agricultural stakeholders globally. This web service would allow users to upload satellite imagery and receive high-resolution outputs without the need for extensive local computational resources.

Sustainability assessment remains a vital component of future research. Evaluating the environmental impact, particularly regarding resource savings such as water and fertilizer use, and yield improvements, will be crucial. Additionally, conducting comprehensive cost-benefit analyses will determine the economic feasibility and return on investment for vineyard managers adopting these technologies. Understanding both the environmental and economic implications will provide a comprehensive view of the sustainability of these methods, encouraging widespread adoption and ensuring long-term benefits.

By concentrating on these key areas, future research can further refine and expand the ap-

plication of super-resolution techniques in precision agriculture. This work will contribute to sustainable food production and resource management on a global scale, addressing the urgent need for innovative agricultural practices to meet the growing demands of the global population.

# Bibliography

[1] What is hyperspectral imaging? [Online]. Available: https://www.nireos.com/hyperspectral-imaging/ xiii, 6

[2] M. Mishra. (2020) Convolutional neural networks, explained. [Online]. Available: https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939 xiii, 9

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy xiii, 10, 11

[4] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient subpixel convolutional neural network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1874–1883. xiii, 14, 15, 33

[5] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2021, pp. 1833–1844. xiii, 14, 15, 31, 32, 40

[6] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. xiii, 15, 16, 17

[7] Z. Chen, Y. Zhang, J. Gu, Y. Zhang, L. Kong, and X. Yuan, "Cross aggregation transformer for image restoration," in *Conference on Neural Information Processing Systems*, 2022, pp. 25 478–25 490. xiii, 14, 17, 19, 32, 40, 44, 45

[8] Q. Zhu, P. Li, and Q. Li, "Attention retractable frequency fusion transformer for image super resolution," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 1756–1763. xiii, 14, 19, 20, 32, 40, 44

[9] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 22 367–22 377. xiii, 14, 21, 32, 40, 45

[10] J. Wang, B. Wang, X. Wang, Y. Zhao, and T. Long, "Hybrid attention-based u-shaped network for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023. xiii, 22, 23, 24, 25, 26, 40, 45

[11] Y. Xiao, Q. Yuan, K. Jiang, J. He, C.-W. Lin, and L. Zhang, "Ttst: A top-k token selective transformer for remote sensing image super-resolution," *IEEE Transactions on Image Processing*, vol. 33, pp. 738–752, 2024. xiii, 23, 26, 27, 28, 31, 32, 40, 44, 46, 47, 49, 51

[12] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3365–3387, 2021. 5

[13] L. Yue, H. Shen, J. Li, Q. Yuan, H. Zhang, and L. Zhang, "Image super-resolution: The techniques, applications, and future," *Signal Processing*, vol. 128, pp. 389–408, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168416300536 5

[14] GISGeography. (2023) Multispectral vs hyperspectral imagery explained. [Online]. Available: https://gisgeography.com/multispectral-vs-hyperspectral-imagery-explained/ 6

[15] T. E. S. Agency. About landsat-8. [Online]. Available: https://earth.esa.int/eogateway/missions/landsat-8 6, 28

[16] Eo-1 hyperion. [Online]. Available: https://cmr.earthdata.nasa.gov/search/concepts/C1220567951-USGS_LTA.html 7

[17] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. 9

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Neural Information Processing Systems*, 2017, pp. 6000–6010. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need 10

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 3, 2014. 13

[20] C. Dong and K. H. . X. T. Chen Change Loy, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*, 2014, p. 184–199. 13

[21] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Conference on Computer Vision and Pattern Recognition*, 2016, p. 1646–1654. 13

[22] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Conference on Computer Vision and Pattern Recognition Workshops*, 2017, p. 136–144. 13

[23] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Conference on Computer Vision and Pattern Recognition*, 2017, p. 105–114. 13

[24] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2472–2481. 13

[25] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *European Conference on Computer Vision Workshops*, 2018, p. 63–79. 13

[26] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *European Conference on Computer Vision*, 2018, p. 294–310. 13

[27] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 057–11 066. 13

[28] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, "Single image super-resolution via a holistic attention network," in *IEEE International Conference on Computer Vision*, 2020. 13

[29] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3516–3525. 13

[30] C. Dong, K. H. Chen Change Loy, and X. Tang, "Deeply-recursive convolutional network for image super-resolution," in *Conference on Computer Vision and Pattern Recognition*, 2016, p. 1637–1645. 13

[31] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2790–2798. 13

[32] S. Zhou, J. Zhang, W. Zuo, and C. C. Loy, "Cross-scale internal graph neural network for image super-resolution," in *Advances in Neural Information Processing Systems*, 2020. 13

[33] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 299–12 310. 14

[34] D. Zhang, J. Shao, X. Li, and H. T. Shen, "Remote sensing image super-resolution via mixed high-order attention network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 6, 2021. 22, 23

[35] S. Lei, Z. Shi, and Z. Zou, "Super-resolution for remote sensing images via local–global combined network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 8, pp. 1243–1247, 2017. 22

[36] J. M. Haut, M. E. Paoletti, R. Fernandez-Beltran, J. Plaza, A. Plaza, and J. Li, "Remote sensing single-image superresolution based on a deep compendium model," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 9, pp. 1432–1436, 2019. 22

[37] Z. Pan, W. Ma, J. Guo, and B. Lei, "Super-resolution of single remote sensing image based on residual dense backprojection networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 7918–7933, 2019. 22

[38] M. Qin, S. Mavromatis, L. Hu, F. Zhang, R. Liu, J. Sequeira, and Z. Du, "Remote sensing single-image resolution improvement using a deep gradient-aware network with image-specific enhancement," *Remote Sensing*, vol. 12, 2020. 22

[39] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced gan for remote sensing image superresolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5799–5812, 2019. 22

[40] S. Lei, Z. Shi, and Z. Zou, "Coupled adversarial training for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3633–3643, 2020. 22

[41] Y. Xu, L. Wei, A. Hu, Z. Xie, X. Xie, and L. Tao, "Te-sagan: An improved generative adversarial network for remote sensing super-resolution images," *Remote Sensing*, vol. 14, p. 2425, 2022. 22

[42] N. Ni, H. Wu, and L. Zhang, "Hierarchical feature aggregation and self-learning network for remote sensing image continuous-scale super-resolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022. 23

[43] H. Wu, N. Ni, and L. Zhang, "Learning dynamic scale awareness and global implicit functions for continuous-scale super-resolution of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023. 23

[44] K. Chen, W. Li, S. Lei, J. Chen, X. Jiang, Z. Zou, and Z. Shi, "Continuous remote sensing image super-resolution based on context interaction in implicit function space," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023. 23

[45] L. Han, Y. Zhao, H. Lv, Y. Zhang, H. Liu, G. Bi, and Q. Han, "Enhancing remote sensing image super-resolution with efficient hybrid conditional diffusion model," *Remote Sensing*, vol. 15, p. 3452, 2023. 23

[46] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, "Ediffsr: An efficient diffusion probabilistic model for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024. 23

[47] S. Lei and Z. Shi, "Hybrid-scale self-similarity exploitation for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2022. 23

[48] S. Lei, Z. Shi, and W. Mo, "Transformer-based multistage enhancement for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022. 23

[49] J. Sui, X. Ma, Z. Xiaokang, and M.-O. Pun, "Gcrdn: Global context-driven residual dense network for remote sensing image superresolution," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. PP, pp. 1–13, 2023. 23

[50] J. Sui, X. Ma, X. Zhang, and M.-O. Pun, "Dtrn: Dual transformer residual network for remote sensing super-resolution," in *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, 2023, pp. 6041–6044. 23

[51] C. Wang, X. Zhang, W. Yang, X. Li, B. Lu, and J. Wang, "Msagan: A new super-resolution algorithm for multispectral remote sensing image based on a multiscale attention gan network," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023. 28

[52] T. E. S. Agency. sentinel-2. [Online]. Available: https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-2 28

[53] C. Lanaras, J. Bioucas-Dias, S. Galliani, and E. Baltsavias, "Super-resolution of sentinel-2 images: Learning a globally applicable deep neural network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 146, 2018. 28

[54] V. Vasilescu, M. Datcu, and D. Faur, "Sentinel-2 60-m band super-resolution using hybrid cnn-gpr model," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023. 28

[55] I. S. of Precision Agriculture. (2024) Precision ag definition. [Online]. Available: https://www.ispag.org/about/definition 29

[56] S. Liaghat and S. K. Balasundram, "A review: The role of remote sensing in precision agriculture," *American Journal of Agricultural and Biological Sciences*, pp. 50–55, 2010. 29

[57] R. P. Sishodia, R. L. Ray, and S. K. Singh, "Applications of remote sensing in precision agriculture: A review," *Remote Sensing*, vol. 12, 2020. 29

[58] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. 37