drones

MDPI

*Article*

# CNN-Based Temporal Fusion of UAV and Sentinel-2 Imagery for Crop Monitoring

**Alvaro Murguia-Cozar [1], Gilberto de Jesús López-Canteñs [1,*], Antonia Macedo-Cruz [2], Noé Velázquez-López [1] and Ronald Ernesto Ontiveros-Capurata [3]**

[1] Posgrado en Ingeniería Agrícola y Uso Integral del Agua, Universidad Autónoma Chapingo, Carretera México-Texcoco km 38.5, Chapingo 56230, Mexico; al22120251@chapingo.mx (A.M.-C.); nvelazquezl@chapingo.mx (N.V.-L.)
[2] Postgrado de Hidrociencias, Colegio de Postgraduados, Carretera México-Texcoco km. 36.5, Montecillo 56264, Mexico; macedoan@colpos.mx
[3] Coordinación de Seguridad Hídrica, Instituto Mexicano de Tecnología del Agua, Paseo Cuauhnáhuac 8532, Jiutepec 62550, Mexico; ronald.ontiveros@posgrado.imta.edu.mx
* Correspondence: glopezc@chapingo.mx

## Highlights

**What are the main findings?**

- The TVDSR model demonstrated the greatest robustness to temporal variability, maintaining high reconstruction quality, spectral consistency, and vegetation index preservation when evaluated on acquisition dates not included in the training dataset.
- Models incorporating both temporal and spectral information exhibited superior generalization capability compared with architectures relying primarily on spatial features. These results highlight the importance of spatiotemporal modeling for Sentinel-2 image super-resolution using UAV multispectral imagery in agricultural environments.

**What are the implications of the main findings?**

- The incorporation of Bidirectional ConvLSTM-based temporal modeling improves the robustness and reliability of Sentinel-2 super-resolution products across varying phenological stages and radiometric conditions, facilitating their operational use in agricultural monitoring.
- The proposed approach enables the reconstruction of Sentinel-2 imagery with a high degree of preservation of vegetation indices, including NDVI, SAVI, GNDVI, NDWI, and NDVIre. This capability supports its application in precision agriculture, crop monitoring, and the estimation of biophysical variables.

## Abstract

The use of Sentinel-2 satellite imagery for crop monitoring is constrained by its native spatial resolution. To address this limitation, data fusion-based super-resolution techniques have emerged as a promising approach for generating imagery with enhanced spatial detail. In this study, we propose TVDSR, a Sentinel-2 and UAV image fusion model that integrates temporal and spectral information through a Bidirectional ConvLSTM-based architecture. The images used for super-resolution were constructed from spectral compositions of five vegetation indices. Model performance was evaluated using 15 acquisition dates, including 11 Sentinel-2 images acquired in 2024 and 4 acquired in 2025. Two independent dates were reserved for validation using reconstruction quality metrics. The proposed model was compared against three state-of-the-art reference models: $S^3$-ESRGAN, DeepSent, and UAV-Net. UAV-Net achieved the best performance for the 2024 validation image (RMSE = 0.128, MAE = 0.10, PSNR = 24.06 dB, and SSIM = 0.86). However, TVDSR obtained the best results for the 2025 evaluation image (RMSE = 0.149, MAE = 0.115, PSNR = 22.60 dB,

Check for updates

Academic Editors: Weiwei Qin, Zhiqiang Tian, Jianhui Xu, Jing Geng and Bing He

Received: 6 October 2026
Revised: 5 November 2026
Accepted: 10 November 2026
Published: 12 November 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

and SSIM = 0.85) and exhibited the lowest performance degradation across acquisition dates, indicating superior temporal generalization capability. Furthermore, TVDSR more consistently preserved the derived vegetation indices, reflecting a better retention of spectral information associated with crop condition, while producing reconstructions that were visually closer to the reference UAV imagery. The joint integration of temporal and spectral information constitutes an effective strategy for improving the robustness and generalization capability of Sentinel-2 image super-resolution for crop monitoring, precision agriculture, crop condition assessment, and agricultural management applications.

**Keywords:** super-resolution; vegetation indices; spatiotemporal model; ConvLSTM; deep learning

## 1. Introduction

Monitoring the Earth's surface through image processing has become a fundamental task across a wide range of disciplines, particularly in agriculture, where it provides a solid basis for crop monitoring, field characterization, and the prediction of biophysical variables aimed at optimizing resource management [1,2]. In recent years, the availability of image acquisition platforms capable of capturing spectral information across different regions of the electromagnetic spectrum has increased, enabling the collection of multispectral and hyperspectral data with enhanced spatial, temporal, and spectral resolution. This development has facilitated the integration of complementary information from multiple sensors, thereby mitigating the inherent limitations associated with relying on a single data source through data fusion techniques [3].

In this context, various techniques have emerged to improve image quality. Among them, super-resolution has become one of the most prominent approaches. Originally developed in the 1980s within the field of image processing, super-resolution focused on reconstructing high-resolution images from multiple low-resolution observations of the same scene [4,5]. With the advancement of satellite remote sensing, this concept evolved into spatial fusion methods such as pansharpening, where a high-resolution panchromatic image is used to enhance the spatial quality of multispectral imagery [6]. The use of high-resolution images as guidance for improving multispectral data has gained increasing attention with the emergence of deep learning techniques. These approaches have enabled the integration of multiscale information acquired from unmanned aerial vehicles (UAVs) and satellite platforms [1,7,8].

Multi-source super-resolution methods can be broadly categorized into two main approaches: single-image super-resolution (SISR) and multi-image super-resolution (MISR). While MISR exploits multiple images of the same scene to enhance a low-resolution (LR) image, SISR relies on a single LR image for reconstruction [9,10]. Representative SISR models include convolutional neural network-based approaches such as the Super-Resolution Convolutional Neural Network (SRCNN) [11], Very Deep Super-Resolution (VDSR) [12], and Enhanced Deep Super-Resolution (EDSR) [13]; generative adversarial methods, including the Enhanced Super-Resolution Generative Adversarial Network (ESRGAN) [14] and Real-ESRGAN [15]; as well as more recent architectures based on Transformers, such as SwinIR [16], and diffusion-based models, including SinSR [17]. These approaches have demonstrated remarkable performance in reconstructing spatial details and improving perceptual image quality.

Despite the advances achieved by SISR methods, their ability to generalize to unseen LR images remains limited. These limitations arise partly because SISR addresses an ill-posed problem, in which multiple high-resolution solutions may correspond to the same

low-resolution image [18]. Consequently, during training, SISR models must infer missing spatial information from incomplete observations, as they rely on a single input image and do not exploit additional sources of information. As a result, their ability to capture spatial, temporal, and spectral variability is constrained, which is particularly relevant in remote sensing applications where surface conditions and spectral reflectance vary over time [19].

In this context, MISR-based approaches provide a more suitable framework for the integration of UAV and satellite imagery, as they enable the incorporation of complementary information from multiple observations of the same area acquired at different dates [20]. By exploiting temporal information, MISR methods enrich the spatial reconstruction process and recover details that may not be observable in a single image due to sensor limitations, noise, or atmospheric conditions [21]. Furthermore, the integration of temporal sequences enables the modeling of surface dynamics over time, leading to more robust and consistent representations under interannual variations, changing illumination conditions, differences in image acquisition settings, and varying degrees of cloud cover commonly encountered in multi-temporal satellite observations.

In this study, we propose a multitemporal super-resolution model, termed Temporal Very Deep Super-Resolution (TVDSR), for Sentinel-2 and UAV data fusion at the vegetation index level. The proposed approach aims to enhance the spatial resolution of Sentinel-2 imagery for crop monitoring by jointly integrating temporal and spectral information, thereby producing more accurate and robust reconstructions over time. The model learns a nonlinear spatiotemporal mapping from a sequence of satellite-derived vegetation index composites to predict a high-resolution image corresponding to a target acquisition date. Its architecture combines a deep convolutional encoder for spatial feature extraction, a ConvLSTM module for temporal dependency modeling, and temporal and spectral attention mechanisms that dynamically weight feature contributions according to their relevance for image reconstruction.

In addition, the model incorporates a day-of-year embedding that modulates the ConvLSTM gating activations, enabling the explicit representation of the cyclical nature of seasonal processes and facilitating generalization across different years and acquisition conditions. TVDSR further adopts a residual learning formulation, in which the network predicts the difference between the low-resolution input and the target high-resolution image. The proposed model is a unified framework based on structures that have demonstrated strong performance in image super-resolution and remote sensing applications. By combining complementary mechanisms for spatial feature extraction, temporal dependency modeling, seasonal representation, and adaptive feature weighting within a single architecture, TVDSR seeks to leverage their respective strengths during the reconstruction process. This integrated design enables the model to exploit multiple sources of information simultaneously, supporting more robust and consistent reconstructions under varying acquisition conditions and interannual variability.

The next section reviews related work on multitemporal super-resolution and satellite-UAV data fusion, with particular emphasis on approaches that exploit temporal information and multisource remote sensing data for enhancing spatial resolution and monitoring vegetation dynamics.

## 2. Related Work

### *2.1. Multi-Image Super-Resolution Models*

MISR models originated from image formation frameworks in which each low-resolution observation is modeled as a degraded version of an unknown high-resolution image. Within this context, image reconstruction was formulated as a regularized inverse problem, where L2- and L1-based loss functions were employed to estimate the image

that best explained the set of available observations [22,23]. Although these approaches demonstrated the feasibility of multi-image reconstruction, their performance was highly dependent on the accuracy of geometric image registration.

To reduce sensitivity to alignment errors, alternative techniques were developed to exploit spatial and temporal redundancies without relying exclusively on explicit motion estimation. Among these, approaches based on Non-Local Means and multidimensional kernel regression introduced more robust mechanisms for modeling similarities among observations and improving reconstruction under noisy conditions or complex motion patterns [24,25]. Nevertheless, these methods remained constrained by the need to explicitly define degradation models and reconstruction rules.

The advent of deep learning transformed the super-resolution problem by enabling the direct learning of mappings between sequences of low-resolution images and their corresponding high-resolution reconstructions. Early approaches combined conventional alignment algorithms with convolutional neural networks responsible for image fusion and reconstruction, as demonstrated by the methods proposed by Liao et al. [26] and the VSRNet model introduced by Kappeler et al. [27]. Subsequently, architectures such as SPMC integrated motion compensation and upsampling within the network itself, eliminating the need for external modules and enabling end-to-end optimization of the reconstruction process [28].

Recent advances have focused on the joint integration of image alignment, temporal fusion, and reconstruction. Models such as RBPN incorporated recurrent mechanisms to iteratively refine temporal information extracted from multiple observations [29], whereas EDVR combined deformable convolutions with spatial and temporal attention mechanisms to improve the dynamic selection of relevant features during reconstruction [30]. Similarly, HighRes-net demonstrated that the hierarchical fusion of deep representations enables the effective integration of information from multiple observations [31].

More recently, research has shifted toward modeling long-range temporal dependencies. BasicVSR and BasicVSR++ established new benchmarks by employing bidirectional feature propagation and advanced alignment strategies based on optical flow and deformable convolutions, achieving substantial improvements in video super-resolution tasks [32,33]. These developments highlight the importance of explicitly exploiting temporal information to enhance reconstruction quality.

### *2.2. Spatiotemporal Image Super-Resolution Models*

Spatiotemporal super-resolution models emerged as an extension of traditional multi-image super-resolution approaches originally developed for video restoration. Unlike methods that rely solely on the spatial fusion of multiple observations, these architectures explicitly model the temporal dependencies between consecutive images to improve the reconstruction of high-resolution details [28,34]. This evolution enabled a more effective exploitation of the redundant information contained in temporal sequences while reducing the reliance on explicit motion estimation procedures.

Subsequently, research shifted toward architectures capable of jointly integrating alignment, fusion, and reconstruction stages. In this context, DeepSUM and DeepSUM++ exploit correlations among multiple satellite acquisitions through deep convolutional networks and incorporate non-local information mechanisms to capture long-range dependencies between similar image regions [35,36]. In parallel, burst super-resolution models introduced adaptive alignment and spatial attention mechanisms to better exploit subtle displacements among input images, as demonstrated by the works of Bhat et al. [37] and Luo et al. [38].

More recent advances have incorporated spatial attention mechanisms, probabilistic models, and hybrid architectures to improve temporal representation learning. These include the probabilistic reformulation proposed by Bhat et al. [39], pseudo-burst strategies [40], and BSRT, which integrates Transformers, optical flow, and deformable convolutions to combine local and global information during the reconstruction process [38]. Similarly, QMambaBSR introduces state-space models and multiscale fusion mechanisms to more efficiently capture both intra-frame and inter-frame interactions [41].

When temporal consistency is a critical requirement, ConvLSTM-based architectures have proven particularly effective due to their ability to simultaneously model spatial and temporal dependencies. Zhang et al. [42] employed ConvLSTM within an encoder–decoder architecture to generate global representations of temporal sequences, while Li et al. [43] incorporated this structure into a multiscale fusion framework to improve reconstruction through the suppression of redundant information. Subsequent studies extended these architectures to image restoration, multimodal fusion, and multitemporal reconstruction tasks using recurrent mechanisms capable of dynamically weighting the contribution of each observation [21,44,45].

In parallel, Transformers have become one of the dominant research directions in multi-image super-resolution. Models such as TR-MISR, MAST, and EpiMISR employ attention mechanisms to establish correspondences among multiple observations and capture global dependencies that are difficult to represent using conventional convolutional operations [46–48]. More recently, HiRT introduced a hierarchical recurrent architecture operating in the frequency domain to simultaneously exploit spatial, temporal, and multiscale dependencies [49]. Likewise, the hybrid architecture FRCFD-Net combines attention mechanisms, gating modules, and cross-layer fusion strategies to improve the recovery of fine details while reducing redundancy in feature representations [50].

### *2.3. Super-Resolution Models for Sentinel-2 and UAV Data Integration*

The super-resolution of Sentinel-2 imagery has evolved from approaches relying exclusively on information provided by the sensor itself toward strategies that integrate multiple data sources to overcome spatial resolution limitations. Early deep learning-based models exploited the higher-resolution Sentinel-2 bands to enhance lower-resolution bands. Representative examples include DSen2 [51], which leverages spectral correlations through deep convolutional networks, as well as the approaches proposed by Palsson et al. [52] and Gargiulo et al. [53], which focus on the recovery of high-frequency details and the preservation of spatial structures through residual learning and information fusion techniques.

Subsequently, several models incorporated external sensors to provide additional spatial information. In this context, Galar et al. [54] employed PlanetScope imagery to generate supervised training datasets using an EDSR-based architecture, whereas Latte & Lejeune [55] proposed a Sentinel-2/PlanetScope fusion strategy based on residual convolutional networks. These studies demonstrated that multisensor integration can overcome the inherent limitations of Sentinel-2 native resolution and significantly improve spatial reconstruction performance.

In parallel, research efforts began to exploit the temporal dimension of remote sensing observations. Jia et al. [56] introduced DL-SDFM, a dual-stream model capable of simultaneously learning spatial and temporal relationships through the adaptive integration of both information sources. More recently, DeepSent incorporated temporal and spectral information within a unified architecture to improve Sentinel-2 image reconstruction at finer metric scales [57]. Likewise, OBSTFM proposed an object-based representation to en-

hance temporal consistency and mitigate noise effects in scenarios characterized by complex spatial changes [58].

Another important research direction involves generative models. RS-ESRGAN [59] demonstrated the potential of generative adversarial networks to enhance the spatial resolution of Sentinel-2 imagery using information from very high-resolution sensors. Subsequently, more advanced architectures such as MAF-GAN [60] and GS-SRGAN [61] incorporated spatial and spectral attention mechanisms to simultaneously improve visual quality and radiometric consistency. More recently, diffusion-based models such as Diff-FuSR [62] have emerged as a promising alternative by explicitly integrating degradation modeling and probabilistic reconstruction processes.

Within the specific context of satellite-UAV integration, recent studies have sought to exploit the complementarity between the high spatial resolution provided by unmanned aerial vehicles and the temporal coverage offered by satellite sensors. RarefyNet utilizes UAV information to refine vegetation indices derived from satellite imagery [63], whereas UAV-Net combines multiscale information from satellite and UAV imagery through deep feature encoding and hierarchical fusion mechanisms [1]. More recently, hybrid approaches integrating super-resolution and spectral harmonization have been proposed. These include the framework presented by Masrur et al. [64], which generates spectrally consistent references using UAV hyperspectral data, and $S^3$-ESRGAN [65], which incorporates spatial attention, channel attention, and spectral preservation mechanisms to maintain radiometric consistency during reconstruction.

Collectively, these advances demonstrate that the integration of UAV and Sentinel-2 data is an effective strategy for combining spatial, temporal, and spectral resolution capabilities [66]. Nevertheless, despite recent progress in deep learning-based fusion and super-resolution techniques, several challenges continue to hinder their widespread adoption. Major limitations include geometric, spectral, and temporal misalignment between sensors caused by differences in spatial resolution, acquisition geometry, observation dates, and spectral response functions; the scarcity of high-quality, co-registered, and multitemporal reference datasets required for effective supervised training; and the substantial computational cost associated with deep architectures that integrate multiple data sources and temporal sequences. Furthermore, temporal integration introduces additional challenges related to scene dynamics, as genuine surface changes may interfere with the learning process, making it difficult to distinguish between legitimate temporal variability and reconstruction artifacts. These limitations highlight the need for more robust, efficient, and adaptive models capable of coherently integrating multiscale, multisensor, and multitemporal information while preserving spectral consistency over time, thereby facilitating their deployment in large-scale operational agricultural and environmental monitoring systems.

## 3. Materials and Methods

### *3.1. Study Area*

This study was conducted at the "San Juan" experimental field of the Autonomous University of Chapingo (Mexico), located at 19°29′35.48″ N, 98°51′35.11″ W, at an elevation of 2211 m above sea level. According to the Köppen climate classification modified by García [67], the study area is characterized as a semi-arid, semi-warm climate with a mean annual temperature below 18 °C, summer rainfall, and a pronounced dry season during winter. Figure 1 shows the location of the study area, which covers 37.9 ha and is divided into 28 maize plots cultivated during the 2024 and 2025 growing seasons.

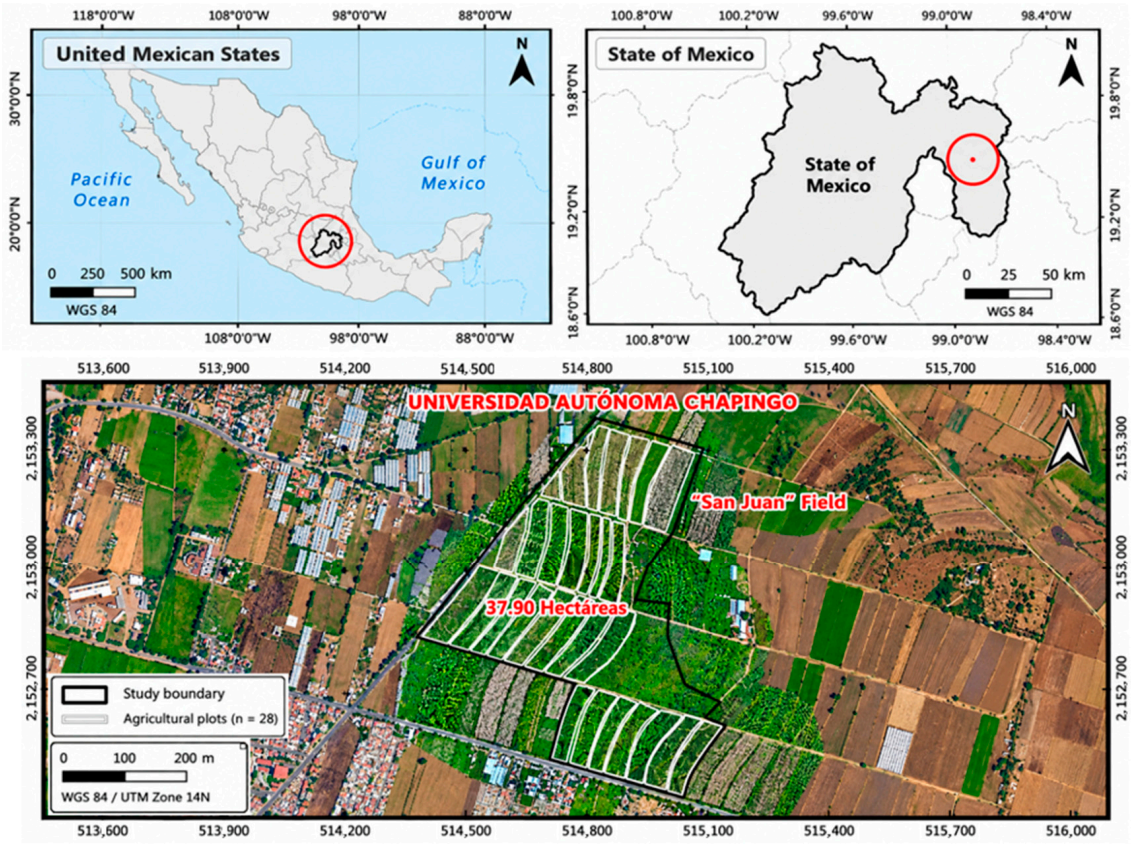



**Figure 1.** Geographic location and delineation of the study area at Universidad Autónoma Chapingo.

### *3.2. Data Acquisition*

#### 3.2.1. UAV Imagery

A DJI Mavic 3M UAV (DJI, Shenzhen, China) was used to conduct image acquisition campaigns during the 2024 and 2025 growing seasons, enabling the collection of high-resolution aerial imagery for vegetation analysis. The platform is equipped with a camera system consisting of one RGB sensor (4/3 CMOS, 20 MP; DJI, Shenzhen, China) and four multispectral sensors (1/2.8-inch CMOS, 5 MP; DJI, Shenzhen, China), including green ($560 \pm 16$ nm), red ($650 \pm 16$ nm), red-edge ($730 \pm 16$ nm), and near-infrared ($860 \pm 26$ nm) bands. The UAV also incorporates a Global Navigation Satellite System (GNSS) module with Real-Time Kinematic (RTK) capabilities, providing centimeter-level positioning accuracy for flight georeferencing.

A total of 11 UAV flights were conducted over the study area between July and November 2024, and four additional flights were conducted between July and September 2025. Table 1 presents information on the temporal acquisition details of the UAV and Sentinel-2 images. All flights were performed at an altitude of 120 m above ground level, with 70% frontal and side overlap. UAV imagery was georeferenced using a D-RTK 2 base station (DJI, Shenzhen, China) installed on a previously established control point within the ITRF2008 reference frame (epoch 2010). The precise coordinates of the control point were determined through static observations acquired with a Trimble R8 receiver (Trimble Inc., Westminster, CO, USA) and processed using differential post-processing in Emlid Studio v.1.9, referenced to the ICMX station of the Mexican National Active Geodetic Network.

**Table 1.** Acquisition details and temporal correspondence between UAV and Sentinel-2 imagery.

| Image Pair | UAV Acquisition Date | Sentinel-2 Acquisition Date | Sentinel-2 Product ID | Time Difference (h) | Cloud Cover (%) | Dataset Assignment |
|---|---|---|---|---|---|---|
| P01 | 14 July 2024 | 14 July 2024 | S2B_MSIL2A_20240714T165849 | 5 | 25 | Training and Testing |
| P02 | 29 July 2024 | 29 July 2024 | S2A_MSIL2A_20240729T165851 | 4 | 8 | Training and Testing |
| P03 | 3 August 2024 | 3 August 2024 | S2B_MSIL2A_20240803T165849 | 6 | 9 | Training and Testing |
| P04 | 8 August 2024 | 8 August 2024 | S2A_MSIL2A_20240808T165851 | 4 | 44 | Training and Testing |
| P05 | 23 August 2024 | 23 August 2024 | S2B_MSIL2A_20240823T165849 | 4 | 51 | Training and Testing |
| P06 | 28 August 2024 | 28 August 2024 | S2A_MSIL2A_20240828T165851 | 3 | 47 | Training and Testing |
| P07 | 12 October 2024 | 12 October 2024 | S2B_MSIL2A_20241012T170149 | 4 | 0 | Training and Testing |
| P08 | 17 October 2024 | 17 October 2024 | S2A_MSIL2A_20241017T170311 | 3 | 4 | Training and Testing |
| P09 | 22 October 2024 | 22 October 2024 | S2B_MSIL2A_20241022T170249 | 4 | 4 | Validation |
| P10 | 1 November 2024 | 1 November 2024 | S2B_MSIL2A_20241101T170349 | 5 | 25 | Training and Testing |
| P11 | 7 November 2024 | 6 November 2024 | S2A_MSIL2A_20241106T170511 | 18 | 19 | Training and Testing |
| P12 | 9 July 2025 | 9 July 2025 | S2B_MSIL2A_20250709T170209 | 1 | 50 | Training and Testing |
| P13 | 4 August 2025 | 3 August 2025 | S2C_MSIL2A_20250803T165911 | 24 | 6 | Validation |
| P14 | 13 August 2025 | 13 August 2025 | S2C_MSIL2A_20250813T165911 | 3 | 12 | Training and Testing |
| P15 | 3 September 2025 | 2 September 2025 | S2C_MSIL2A_20250902T165911 | 22 | 45 | Training and Testing |

Radiometric correction was supported by the onboard sunlight sensor integrated into the DJI Mavic 3 Multispectral, which records solar irradiance during image acquisition and enables compensation for illumination variations during photogrammetric processing. Orthomosaics and reflectance maps were generated using Pix4Dmapper v.4.10 with RTK-based image georeferencing. No ground control points (GCPs), or independent checkpoints were collected during the surveys. Therefore, the geometric consistency of the final products relied on the integrated RTK positioning system and the quality assessment metrics provided by the photogrammetric processing workflow.

Finally, the reflectance maps were resampled through downsampling to a spatial resolution of 2 m by averaging all pixels contained within each 2 m cell. This procedure was intended to generate an intermediate spatial resolution between the native UAV imagery (0.06 m) and the native Sentinel-2 imagery (10 m and 20 m), thereby facilitating the subsequent data fusion process (Figure 2).

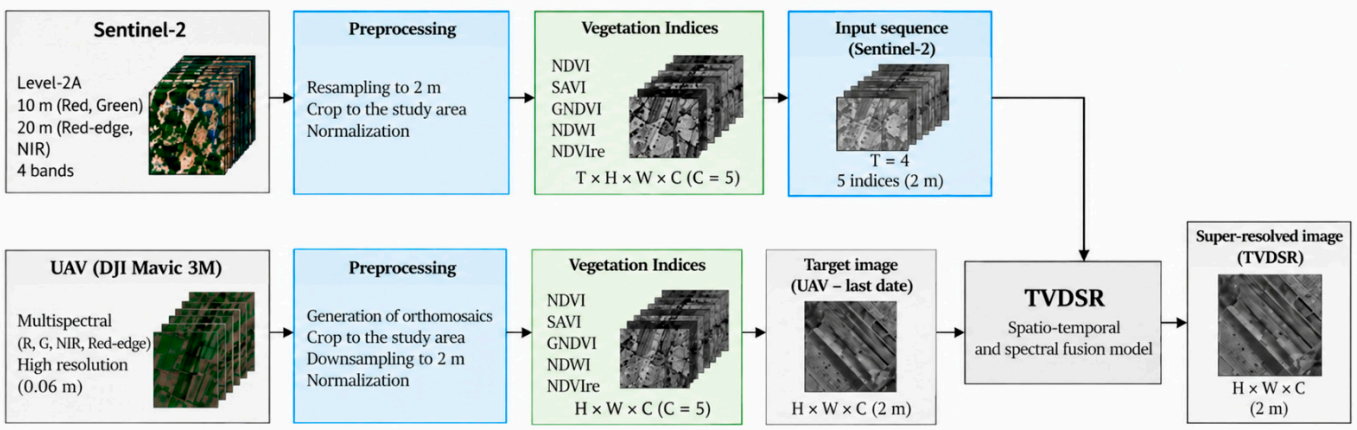



**Figure 2.** Methodological workflow for Sentinel-2 and UAV image fusion using the TVDSR model.

### 3.2.2. Sentinel-2 Imagery

For each UAV acquisition date, Sentinel-2 Level-2A imagery corresponding to the green (B03 560 nm, 10 m), red (B04 665 nm, 10 m), red-edge (B06 740 nm, 20 m), and near-infrared (B8A 865 nm, 20 m) bands were downloaded. These bands were selected because of their relevance for the computation of the vegetation indices used in this study. Subsequently, all Sentinel-2 images were resampled to a 2 m grid using bicubic interpolation. This procedure was employed exclusively to create the spatial grid required by the super-resolution frame-

work and should not be interpreted as an increase in the effective spatial resolution of the Sentinel-2 imagery. Since Sentinel-2 Level-2A products are provided as atmospherically corrected and georeferenced surface-reflectance products, no additional radiometric or geometric corrections were applied prior to resampling and spatial harmonization with the UAV-derived products.

### 3.3. Vegetation Indices

Five vegetation indices were selected because they provide complementary information regarding vegetation vigor, soil-background effects, chlorophyll-related variability, and red-edge response. In addition, these indices share a common numerical range between −1 and 1, which facilitates the harmonization of Sentinel-2 and UAV observations and contributes to a more stable learning process. Furthermore, because absolute radiometric calibration of the UAV imagery was not available, direct fusion at the reflectance level was not considered appropriate. In contrast, vegetation indices provided a more consistent representation between sensors while preserving information relevant for crop monitoring applications.

Since radiometrically calibrated UAV reflectance values were not available, each spectral band was independently normalized to the range $[0, 1]$ prior to vegetation index computation. This preprocessing step was adopted to ensure numerical consistency across sensors and acquisition dates, as the radiometric response of the imagery varied between dates due to differences in illumination conditions and crop phenology. Under these circumstances, defining a single global normalization scheme applicable to all dates and sensors was not feasible without introducing additional distortions in the data distribution. Therefore, min–max normalization (Equation (1)) was employed as a practical preprocessing strategy to harmonize the input data while preserving the relative spectral variability required for vegetation index computation and subsequent image reconstruction.

$$I_{norm} = \frac{I - I_{MIN}}{I_{MAX} - I_{MIN}} \quad (1)$$

where $I_{norm}$ is the normalized band, $I$ is the original pixel reflectance value, $I_{min}$ is the minimum reflectance value within the image, and $I_{max}$ is the maximum reflectance value within the image.

From the normalized images, five vegetation indices were computed for each acquisition date from both Sentinel-2 and UAV imagery. The selected indices were NDVI, SAVI, GNDVI, NDWI, and NDVIre, as they provide complementary information on vegetation vigor, chlorophyll content, vegetation condition, and red-edge spectral response, all of which are widely used for crop monitoring. Although the Green–NIR NDWI proposed by McFeeters was originally developed for open-water detection, previous studies have demonstrated its sensitivity to variations in canopy water content and phenology [67–69]. It should not be confused with NIR–SWIR-based water indices, which are specifically designed to estimate vegetation water content and moisture status.

Subsequently, the five indices were combined into a five-channel multiband composite and clipped to the study area shown in Figure 1. The equations used to compute each vegetation index are presented in Table 2.

**Table 2.** Vegetation indices evaluated in this study.

| Index | Equation |
|---|---|
| Normalized Difference Vegetation Index | $NDVI = \frac{NIR-RED}{NIR+RED}$ |
| Soil Adjusted Vegetation Index | $SAVI = \frac{NIR-RED}{NIR+RED+L}(1+L)$ |
| Green Normalized Difference Vegetation Index | $GNDVI = \frac{NIR-GREEN}{NIR+GREEN}$ |
| Normalized Difference Water Index | $NDWI = \frac{GREEN-NIR}{GREEN+NIR}$ |
| Red-Edge Normalized Difference Vegetation Index | $NDVIre = \frac{RE-RED}{RE+RED}$ |

Note: RED = red band; GREEN = green band; NIR = near-infrared band; RE = red-edge band; L = 0.5 is the soil brightness correction factor used in the SAVI calculation.

### *3.4. Proposed Method*

#### 3.4.1. TVDSR Model Framework

The proposed TVDSR model is a hybrid architecture that integrates several components derived from previous advances in image super-resolution and spatiotemporal modeling (Figure 3). In particular, the model is inspired by the VDSR architecture [12], which is characterized by using deep convolutional networks combined with a global residual learning strategy, enabling effective feature extraction with low computational cost. Moreover, this architecture has demonstrated promising performance in Sentinel-2 and UAV image fusion at the vegetation index level [68]. Since VDSR was originally designed as an SISR model, TVDSR extends its formulation to incorporate the temporal dimension under the MISR framework. To this end, a window-based sequential modeling scheme is introduced, in which each input sample is composed of multiple observations.

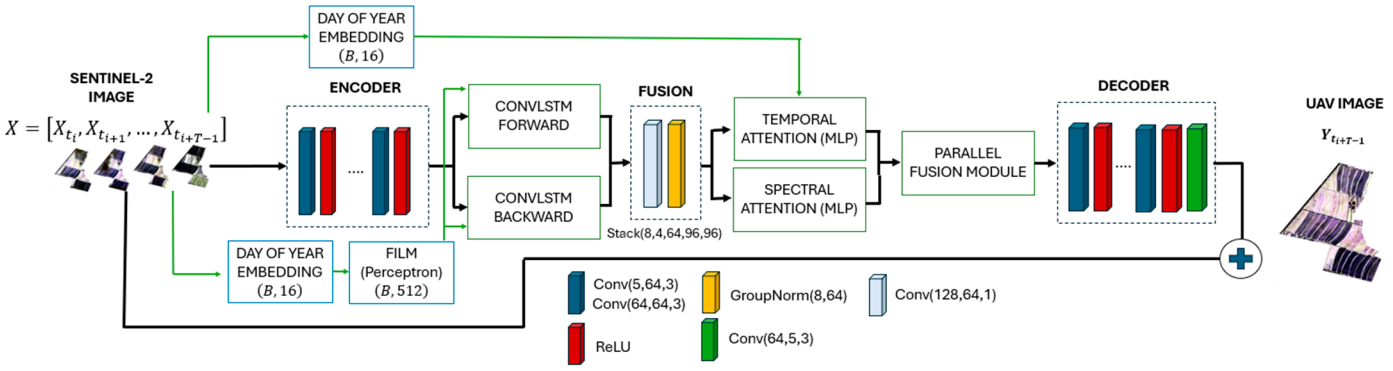



**Figure 3.** Overall architecture of the TVDSR model.

The incorporation of ConvLSTM enables the model to retain temporal information and exploit temporal dependencies among observations, providing contextual information that can support the reconstruction of the target image. Furthermore, day-of-year embedding introduces seasonality information into the temporal modeling process, facilitating the adaptation of the ConvLSTM memory dynamics to crop phenological changes. Temporal and spectral attention mechanisms are incorporated to adaptively weight observations according to their relevance for reconstruction, allowing the model to prioritize images acquired under more favorable temporal and radiometric conditions. Although these components have previously demonstrated effectiveness in image super-resolution and remote sensing applications, their integration within a unified framework for multitemporal Sentinel-2 and UAV vegetation index fusion has not been previously explored.

The problem addressed in this study is formulated as a supervised spatiotemporal super-resolution task, where the objective is to learn a nonlinear mapping function between Sentinel-2 image sequences and a high-resolution UAV image corresponding to the target

acquisition date. Given a set of multitemporal Sentinel-2 observations represented as a chronologically ordered sequence:

$$X = \{X_{t_1}, X_{t_2}, \ldots, X_{t_T}\} \tag{2}$$

where $X_t \in R^{C \times H \times W}$ denotes a five-channel image composed of the vegetation indices, and the objective is to estimate a high-spatial-resolution image acquired by the UAV at the target time step $t_T$:

$$Y_{t_T} = \mathcal{F}(X) \tag{3}$$

where $Y_{t_T} \in R^{C \times H \times W}$ denotes the reference UAV image at the target acquisition date, and $\mathcal{F}(\cdot)$ represents a complex nonlinear function that models the spatial, spectral, and temporal relationships between the input observations and the high-resolution output.

Unlike SISR approaches, the proposed framework explicitly incorporates the temporal dimension, enabling the exploitation of multitemporal redundancy contained within the satellite observations. This redundancy is particularly relevant in agricultural environments, where phenological dynamics generate consistent temporal patterns that can be leveraged to improve spatial reconstruction. Each training sample is constructed using a sliding-window strategy applied to the available image sequences, such that the model receives as input a set of T consecutive Sentinel-2 observations:

$$X = \left[X_{t_i}, X_{t_{i+1}}, \ldots, X_{t_{i+T-1}}\right] \tag{4}$$

and the corresponding UAV image at the final time step of the sequence as the target output $Y_{t_{i+T-1}}$.

To capture spatiotemporal dependencies within the input windows, the proposed model incorporates a ConvLSTM module, which enables temporal information to be integrated while simultaneously preserving the spatial structure of the data [69]. In addition, a bidirectional processing scheme is adopted, whereby the historical observation sequence is analyzed in both forward and backward directions. This strategy enables the model to capture temporal dependencies at different positions within the input window, providing a richer representation of the temporal context available before the target acquisition date [33,49].

Furthermore, the model explicitly incorporates temporal information through sinusoidal embeddings [70] derived from the day of the year, facilitating the representation of periodic patterns associated with the temporal dynamics of the observations. These embeddings are used to modulate the ConvLSTM behavior via a Feature-wise Linear Modulation (FiLM) mechanism [71], which adjusts the internal gating operations through scale and shift parameters conditioned on temporal information.

Subsequently, the model incorporates a temporal attention module that assigns adaptive weights to the hidden states across the sequence based on the sinusoidal embeddings, enabling the network to emphasize the temporal features that are most relevant for the reconstruction task. In parallel, a spectral attention mechanism is employed to estimate channel-wise adaptive weights, highlighting the most informative spectral features. The outputs of both attention modules are then combined through a parallel fusion module, which learns two weighting parameters to achieve an effective integration of temporal and spectral information. Finally, the resulting representation is processed by a convolutional decoder and combined with a residual connection to reconstruct the output image.

#### 3.4.2. Spatial Feature Extraction Module (Encoder)

The spatial feature extraction module transforms the input images into feature tensors that preserve the spatial structure of the data. As illustrated in Figure 4, the encoder consists

of ten 2D convolutional layers with ReLU nonlinear activations. The module begins with an initial convolutional layer that projects the input channels into a higher-dimensional feature space, increasing the number of channels from 5 to 64. This is followed by nine intermediate convolutional layers that maintain the feature dimensionality and progressively extract higher-level spatial representations through successive convolutional operations. Specifically, the encoder is composed of a set of 3 × 3 spatial filters with padding applied to preserve the spatial resolution of the feature maps, thereby facilitating the capture of local spatial patterns. This feature extraction process is independently applied to each element of the temporal window, generating a sequence of feature maps of dimension $B \times C \times H \times W$ por cada instante temporal.

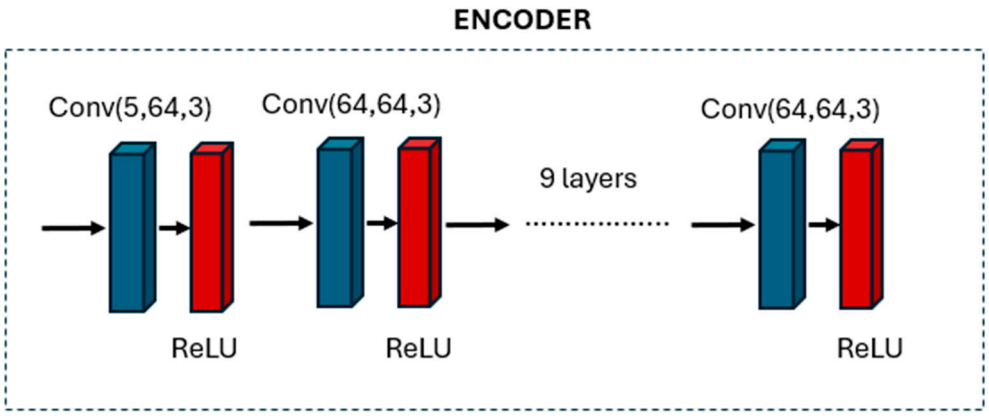



**Figure 4.** Architecture of the TVDSR encoder.

3.4.3. Bidirectional ConvLSTM-Based Temporal Modeling

After extracting spatial features from the Sentinel-2 imagery, the proposed model incorporates a Convolutional Long Short-Term Memory (ConvLSTM) layer. Unlike a conventional LSTM, where gate activations are computed through matrix multiplications, ConvLSTM replaces these operations with two-dimensional convolutions. This modification enables the network to preserve the spatial structure of the feature tensors generated by the encoder while simultaneously modeling temporal dependencies.

Let $F_t$ denote the feature tensor extracted by the encoder at time step $t$. The ConvLSTM updates its hidden and cell states through the input $i_t$, forget $f_t$ y output $o_t$ gates, together with a candidate cell state. The parameter update equations are defined as follows:

$$i_t = \sigma(W_{xi} * F_t + W_{hi} * H_{t-1} + b_i) \quad (5)$$

$$f_t = \sigma\left(W_{xf} * F_t + W_{hf} * H_{t-1} + b_f\right) \quad (6)$$

$$\tilde{C}_t = tanh(W_{xc} * F_t + W_{hc} * H_{t-1} + b_c) \quad (7)$$

$$C_t = f_t \bigodot C_{t-1} + i_t \bigodot \tilde{C}_t \quad (8)$$

$$o_t = \sigma(W_{xo} * F_t + W_{ho} * H_{t-1} + b_o) \quad (9)$$

$$H_t = o_t \bigodot tanh(C_t) \quad (10)$$

where $H_t$ denotes the hidden state, $C_t$ the cell state, $\sigma(\cdot)$ the sigmoid activation function, $\tanh(\cdot)$ the hyperbolic tangent activation function, $W_{xi}$, $W_{xf}$, $W_{xo}$ and $W_{xc}$ represent the convolutional filters applied to the current input; $W_{hi}$, $W_{hf}$, $W_{ho}$ and $W_{hc}$ correspond to the filters associated with the previous hidden state $H_{t-1}$; $b_i$, $b_f$, $b_o$ and $b_c$ denote the bias terms; and $\bigodot$ represents element-wise multiplication.

In the proposed ConvLSTM implementation, the activations of the three gates and the candidate cell state are computed simultaneously through a single 2D convolutional

operation comprising 256 filters of size 3 × 3. To achieve this, the 64-channel feature tensor extracted by the encoder is concatenated with the previous 64-channel hidden state, producing a tensor with 128 channels. After convolution, an output $G_t$ tensor containing 256 feature maps is generated and subsequently partitioned into four groups of 64 channels, corresponding to $i_t$, $f_t$, $o_t$ $y$ $\widetilde{C}_t$, respectively.

To exploit temporal dependencies within the historical observation sequence, a bidirectional architecture composed of two independent ConvLSTM branches was implemented (Figure 3). The first branch processes the input sequences in chronological order, while the second processes the same sequence in reverse order. As a result, two hidden-state representations, $H_t^{fwd}$ and $H_t^{bwd}$ $\epsilon$ $\mathbb{R}^{B\times 64\times H\times W}$, are generated for each time step, corresponding to the forward and backward temporal directions, respectively.

#### 3.4.4. Seasonal Modeling Based on Embeddings and FiLM

The spectral dynamics observed in agricultural systems exhibit a strong seasonal dependence associated with crop development. Consequently, images acquired on nearby dates within the same growing season tend to exhibit similar spectral characteristics, whereas observations separated by longer temporal intervals often display greater spectral variability. To incorporate this temporal context, a day-of-year-based representation was employed. For each Sentinel-2 image, the acquisition date was converted, using Equation (11), into a normalized day-of-year value defined within the interval $[0, 1)$:

$$d_t = \frac{DOY_t - 1}{365} \tag{11}$$

where $DOY_t$ denotes the day of the year corresponding to the acquisition date. In leap years, 29 February is assigned a *DOY* value of 59.5 to maintain continuity in the seasonal representation.

To preserve the cyclical nature of seasonal processes, the normalized values were transformed into sinusoidal vectors using Equation (12). This representation generates a temporal embedding $E_t$ composed of sine and cosine functions at different frequencies, enabling the model to encode periodic temporal patterns and represent temporal proximity between observations in a continuous and smooth manner.

$$E_t = [sin(k2\pi d_t), cos(k2\pi d_t)]_{k=1}^{8} \tag{12}$$

where $k$ represents the different harmonics used to describe annual periodicity.

To incorporate seasonal information into the learning process, a FiLM mechanism was implemented. The temporal embedding vector $E_t$ was projected through a fully connected layer from a 16-dimensional space into a vector containing 512 parameters, as follows:

$$P_t = W_f E_t + b_f \tag{13}$$

where $W_f$ and $b_f$ denote the trainable weight matrix and bias vector of the linear projection layer, respectively.

The resulting vector $P_t$ is subsequently split into two equally sized subsets corresponding to the scaling factors $\gamma_t$ with $\gamma_t$ $\epsilon$ $\mathbb{R}^{256}$ and the shifting factors $\beta_t$ with $\beta_t$ $\epsilon$ $\mathbb{R}^{256}$. These parameters are then used to modulate the activations generated by the internal ConvLSTM convolution through the following operation:

$$G_t' = \gamma_t \bigodot G_t + \beta_t \tag{14}$$

where $G_t$ represents the original activations generated by the recurrent cell.

3.4.5. Temporal Fusion and Attention Mechanisms

To integrate the temporal information extracted by the ConvLSTM layer in both directions, the forward and backward hidden states were concatenated along the channel dimension, as shown in Equation (15).

$$H_t^{cat} = \left[H_t^{fwd},\ H_t^{bwd}\right] con\ H_t^{cat} \in \mathbb{R}^{B\times 128\times H\times W} \tag{15}$$

Subsequently, a state fusion operation was applied using a convolutional layer with 64 filters of size $1 \times 1$, projecting the concatenated representation $H_t^{cat}$ into a lower-dimensional feature space and generating the fused feature tensor $H_t^{fusion} \in \mathbb{R}^{B\times 64\times H\times W}$. A Group Normalization (GroupNorm) layer with eight groups and 64 output channels was then applied. This operation normalizes feature values within channel groups, reducing sensitivity to batch size and improving training stability.

Although the Bidirectional ConvLSTM provides an enriched temporal representation, observations within the sequence may contribute differently to the reconstruction of the target date depending on their temporal relationship with the target acquisition. To account for this effect, a season-aware temporal attention mechanism guided by day-of-year embeddings was incorporated. In parallel, a channel-based spectral attention mechanism was employed to adaptively weight the latent feature channels generated by the network according to their relevance for reconstruction. The outputs of these two attention branches were subsequently integrated through a parallel fusion module, which learns global weighting coefficients to balance the contributions of temporal and spectral information during the reconstruction process.

Temporal Attention Mechanism

Let $E_t$ denote the embedding associated with time step $t$ and $E_T$ the embedding corresponding to the target date. Both vectors are concatenated to form a 32-dimensional temporal representation $S_t$. This representation is processed by a multilayer perceptron (MLP) consisting of a hidden layer with 16 neurons and ReLU activation, followed by a single output neuron. For a sequence of length $T$, the perceptron generates a set of attention scores s:

$$s = [s_1,\ s_2, \ldots,\ s_T] \tag{16}$$

The resulting scores are subsequently normalized using a Softmax function to obtain the attention coefficients, as defined in Equation (17).

$$\alpha_t = \frac{e^{s_t}}{\sum_{j=1}^{T} e^{s_j}}\ con \sum_{t=1}^{T} \alpha_t = 1 \tag{17}$$

Each coefficient $\alpha_t$ represents the relative importance of an observation within the temporal sequence. Finally, the representation used for spatial reconstruction is computed as a weighted aggregation of all fused hidden states, as expressed in Equation (18):

$$H_{final} = \sum_{t=1}^{T} \alpha_t H_t^{fusion} \tag{18}$$

Spectral Attention Mechanism

To enhance the spectral representation learned by the ConvLSTM block, a channel-based spectral attention mechanism was incorporated. This module assigns adaptive weights to each spectral feature learned by the network, emphasizing those that are most relevant for reconstructing the target image while suppressing less informative features.

Let $H_t^{fusion} \in \mathbb{R}^{B\times T\times C\times H\times W}$ denote the reduced output of the ConvLSTM block, where $B$ represents the batch size, $T$ the number of temporal observations, $C$ the number of feature channels, and $H \times W$ the spatial dimensions. In the first stage, a temporally aggregated representation is obtained by averaging the features along the temporal dimension, as defined in Equation (19):

$$H_{mean} = \frac{1}{T}\sum_{t=1}^{T} H_t^{fusion} \tag{19}$$

Ahora $H_{mean} \in \mathbb{R}^{B\times C\times H\times W}$.

Subsequently, two global descriptors are computed for each channel: Global Average Pooling (GAP) and Global Max Pooling (GMP). These descriptors summarize the spatial information contained within each channel, as defined by Equations (20) and (21), respectively.

$$z_{avg} = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} H_{mean}(c,i,j) \tag{20}$$

$$z_{max}(c) = \underset{i,j}{max} H_{mean}(c,i,j) \tag{21}$$

Both descriptors are processed through a shared MLP. This shared architecture reduces the number of trainable parameters while enabling the modeling of global dependencies among spectral channels:

$$f(z) = W_2\delta(W_1 z) \tag{22}$$

where $W_1$ y $W_2$ denote the weight matrices of the MLP, and $\delta(\cdot)$ represents the ReLU activation function. The outputs obtained from the GAP and GMP descriptors are then combined and normalized using a sigmoid activation function to generate spectral attention weights, as defined in Equation (23).

$$A = \sigma(f\left(z_{avg}\right) + f(z_{max})) \tag{23}$$

where $A \in \mathbb{R}^C$ contains the estimated relative importance of each feature channel. Finally, the computed attention weights are applied to the temporally aggregated representation through channel-wise multiplication, as expressed in Equation (24).

$$H_{spec} = A \bigodot H_{mean} \tag{24}$$

where $\bigodot$ denotes element-wise multiplication.

Parallel Fusion Module

To simultaneously exploit the information provided by the temporal and spectral attention modules, an adaptive parallel fusion module was incorporated. Unlike fixed feature combination strategies, this mechanism enables the network to automatically learn the relative contribution of each branch during training.

Let $H_{final}$ denote the representation produced by the temporal attention module and $H_{spec}$ the representation generated by the spectral attention module. Both branches produce feature maps with compatible dimensions, which are integrated through a weighted combination. To this end, a vector of trainable parameters is defined as follows:

$$\theta = [\theta_t, \theta_s] \tag{25}$$

where $\theta_t$ y $\theta_s$ represent the trainable parameters associated with the temporal and spectral branches, respectively. During training, these parameters are optimized through backpropagation and gradient descent. To obtain positive and normalized coefficients, a Softmax function is applied, as expressed in Equation (26):

$$\omega_i = \frac{\exp(\theta_i)}{\sum_{j=1}^{2}\exp\left(\theta_j\right)} \tag{26}$$

yielding the adaptive fusion weights:

$$\omega_t, \omega_s \; con \; \omega_t + \omega_s = 1 \tag{27}$$

The fused representation is then computed as a weighted linear combination of the two attention branches, as defined in Equation (28):

$$F = \omega_t H_{final} + \omega_s H_{spec} \tag{28}$$

3.4.6. Spatial Reconstruction Module (Decoder)

The spatiotemporal representation is processed by a deep convolutional decoder responsible for reconstructing the spatial details absent from the Sentinel-2 composites. As illustrated in Figure 5, the decoder consists of nine convolutional layers with 64 filters of size 3 × 3, each followed by a ReLU activation function. These convolutional operations progressively transform the fused spatiotemporal representation into a spatial residual representation. The final convolutional layer projects the 64-channel latent representation into a five-channel output space through the application of five 3 × 3 filters.

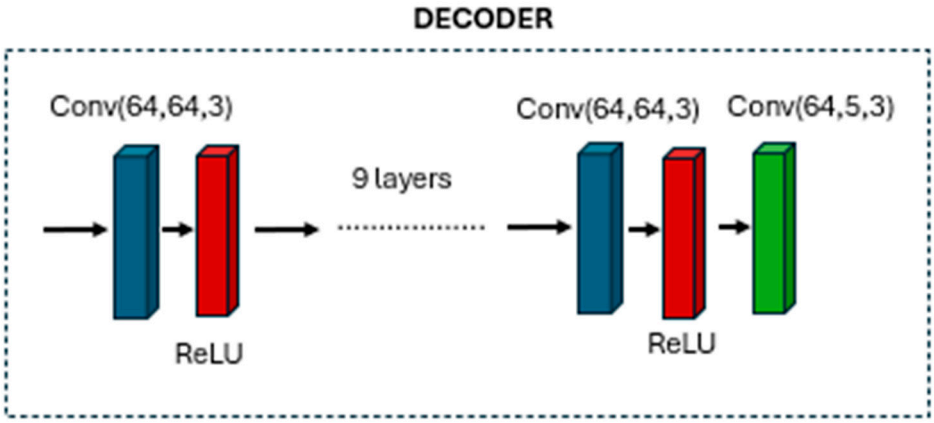



**Figure 5.** TVDSR model decoder architecture.

Finally, the residual learning strategy originally proposed in the VDSR model is applied to generate the final TVDSR output. This strategy employs a skip connection that adds the reconstructed residual to the last Sentinel-2 composite in the temporal sequence, as expressed in Equation (29):

$$Y = X_T + R(F) \tag{29}$$

where $X_T$ denotes the Sentinel-2 image corresponding to the target acquisition date, and $Y$ represents the final prediction generated by TVDSR.

3.4.7. Loss Function and Training Strategy

The TVDSR model was formulated as a supervised regression problem aimed at learning a nonlinear spatiotemporal transformation between a sequence of Sentinel-2 composites and the UAV composite corresponding to the final date of the temporal window. During training, the prediction generated by the network was compared with the reference UAV image using the Mean Squared Error (MSE) loss function, defined as follows:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{Y}_i - Y_i\right)^2 \tag{30}$$

where $\hat{Y}_i$ represents the value predicted by the model, $Y_i$ corresponds to the observed value in the reference UAV image, and N denotes the total number of evaluated pixels.

Minimization of this loss function enabled the joint optimization of all trainable parameters through backpropagation. Consequently, the loss was used not only to update the weights of the convolutional encoder, the ConvLSTM module, and the decoder, but also the parameters of the linear projection employed in the FiLM module and the weights of the multilayer perceptron used in the temporal attention mechanism. As a result, the entire architecture was trained in an end-to-end manner.

Optimization was performed using the Adam algorithm with an initial learning rate of $\eta_0 = 3 \times 10^{-4}$. To promote stable convergence during the final stages of training, a Step Learning Rate Scheduler was employed, reducing the learning rate by a factor of 0.5 every 50 epochs. In addition, a gradient clipping strategy with a maximum norm of 0.5 was applied to prevent numerical instabilities caused by excessively large gradients.

#### 3.4.8. Implementation and Computational Details

The TVDSR model was implemented in Python using the PyTorch library for the development and training of the proposed deep learning architecture. Geospatial data input and output operations were performed using Rasterio v.1.3, while NumPy v.1.24 was employed for the numerical processing of the input data. The architecture was developed following a modular design that enabled the integration of spatial feature extraction, temporal modeling, seasonal modulation, and residual reconstruction components within a single end-to-end trainable network.

Training samples were generated using sliding temporal windows of length $T = 4$ from 13 acquisition dates. This configuration produced 10 temporal windows, including seven windows whose target acquisition belonged to the 2024 growing season and three windows whose target acquisition belonged to the 2025 growing season. Windows were generated from the complete chronological series and therefore could span the temporal gap between acquisition campaigns.

The hyperparameters of TVDSR were selected through an iterative empirical tuning process based on validation performance and computational efficiency. Different configurations of patch size, batch size, learning rate, hidden-state dimensionality, and training duration were evaluated during model development. The final configuration corresponded to the best trade-off between reconstruction accuracy, training stability, and computational cost.

The model was trained for 500 epochs using a batch size of eight samples. Each sample consisted of a sequence of four consecutive Sentinel-2 observations and a corresponding UAV image acquired on the last date of the sequence. To increase the number of training examples, 200 image patches of 96 × 96 pixels were randomly extracted from each temporal window. Patch locations were sampled randomly across the raster extent, allowing overlap between adjacent patches. All observations belonging to the same temporal sample were extracted from the same spatial location to preserve spatiotemporal consistency throughout the sequence. Pixels located outside the valid study area were represented as no data values and excluded during preprocessing. Temporal samples containing only invalid pixels were discarded, whereas patches containing a mixture of valid and invalid pixels were retained.

To monitor the training process and reduce the risk of overfitting, the available data were divided into training and validation subsets using an 80:20 ratio. Temporal samples were generated from sliding windows constructed from the chronologically ordered acquisition series. Patches used for training and validation were extracted from the temporal windows assigned to their corresponding subset. For the testing dates of 22 October 2024 and 3 August 2025, the input temporal window was constructed using training acquisitions collected prior to the testing date, while the testing image itself was defined as the target acquisition. This strategy was adopted because no additional observations were available to construct a complete temporal context around the target date. Consequently, the model performed inference using only historical information available before the evaluation date.

### *3.5. Validation of the TVDSR Model*

To assess the temporal generalization capability of the TVDSR model within the same study area, a quantitative evaluation was conducted using Sentinel-2 and UAV imagery acquired on 22 October 2024, and 3 August 2025. These acquisition dates were completely excluded from the model development process, including both input temporal sequences and target images, and were used exclusively for testing. This evaluation strategy enabled the assessment of the model under realistic operational conditions, where inference is performed on acquisition dates that were not available during training or validation.

Additionally, the obtained metrics were used to compare the performance of TVDSR against three representative state-of-the-art satellite image super-resolution models: DeepSent, UAV-Net, and $S^3$-ESRGAN. These models were adapted to the characteristics of the datasets used in this study. Their selection was motivated by the fact that they represent distinct reconstruction paradigms. DeepSent integrates temporal and spectral information to enhance Sentinel-2 imagery through deep learning; UAV-Net focuses on Sentinel-2 and UAV image integration using multiscale fusion mechanisms; and $S^3$-ESRGAN employs a generative adversarial architecture incorporating spatial attention and spectral preservation mechanisms for Sentinel-2/UAV fusion. Consequently, the comparison enabled the evaluation of TVDSR against approaches based on temporal integration, multisensor fusion, and generative modeling, respectively. For the comparison models, the hyperparameter settings reported in the original publications were adopted whenever possible, introducing only the modifications required to adapt the models to the experimental framework of this study.

The original UAV-Net architecture was adapted for Sentinel-2 and UAV vegetation index fusion. The main modification consisted of replacing the original reflectance-band inputs and outputs with five vegetation indices, increasing the input channels from 12 to 15 and the output channels from four to five. Additional adjustments included the removal of Batch Normalization layers, the use of smaller training patches ($96 \times 96$ pixels), a reduced batch size (8), an extended training schedule (500 epochs), and input normalization to the range $[-1, 1]$. A bilinear interpolation layer was also incorporated at the end of the decoder to ensure spatial consistency between input and output images.

The original DeepSent framework was adapted to operate on vegetation indices instead of Sentinel-2 reflectance bands. Like the original architecture, the adapted model seeks to jointly exploit temporal and spectral information to improve image reconstruction. However, while the original framework models temporal and spectral relationships directly from Sentinel-2 reflectance bands, the adapted version operates on vegetation indices. Furthermore, the original recursive spectral-temporal fusion strategy was replaced by parallel temporal and spectral branches that jointly process multitemporal Sentinel-2 observations and directly reconstruct the corresponding UAV-derived vegetation index image. Training was performed using fixed temporal windows and UAV vegetation index maps as supervision targets.

The original $S^3$-ESRGAN framework was adapted from RGB Sentinel-2 super-resolution to Sentinel-2 and UAV vegetation index fusion. Like the original architecture, the adapted model preserves the use of RRDB blocks, spatial attention mechanisms, and spectral feature extraction modules to jointly exploit spatial and spectral information. However, instead of operating on RGB reflectance bands and performing a 5× spatial upscaling task, the proposed version processes five vegetation indices represented on a common 2 m spatial grid. Consequently, the network was reformulated to perform direct reconstruction of UAV vegetation index maps from Sentinel-2 vegetation indices, using a scale factor of 1 and UAV-derived vegetation index products as supervision targets. Additionally, the original spectral preservation strategy designed for RGB imagery was replaced by a cosine-similarity-based spectral consistency loss adapted to multi-index representations.

To characterize reconstruction quality, a set of complementary spatial and spectral metrics was employed. These metrics enabled the simultaneous evaluation of spatial detail preservation and the consistency of spectral information associated with vegetation indices. The evaluation metrics used in this study are summarized in Table 3. All evaluation metrics were computed using only valid pixels located within the study area (Figure 1). Pixels outside the study boundary were assigned as invalid and excluded from processing during training, validation, and testing. Although SAM and ERGAS were originally proposed for multispectral reflectance imagery, in this study the five vegetation indices were treated as a multivariate representation of crop condition. Consequently, SAM was used to evaluate the angular similarity between vegetation-index vectors, while ERGAS provided a global measure of reconstruction error across all vegetation indices.

**Table 3.** Quality metrics used to evaluate the performance of the super-resolution models.

| Metric | Equation | Description |
| --- | --- | --- |
| Root Mean Square Error | $RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{N}}$ | A radiometric metric that measures the square root of the average reconstruction error, providing an assessment of the reconstruction fidelity of an image fusion method [72]. |
| Mean Absolute Error | $MAE = \frac{1}{N}\sum_{i=1}^{N}\lvert x_i - y_i\rvert$ | A radiometric metric that quantifies the average absolute difference between the reconstructed image and the reference image values. |
| Peak Signal-to-Noise Ratio | $PSNR = 10log\left(\frac{{max_I}^2}{MSE}\right)$ | A radiometric metric that measures the ratio between the maximum possible signal value and the reconstruction error. It is widely used to evaluate radiometric distortion and reconstruction quality, with higher values indicating greater similarity to the reference image [73,74]. |
| Structural Similarity Index Measure | $SSIM(x,\ y) = \frac{((2\cdot\mu_x\cdot\mu_y + C_1)\cdot\ (2\cdot\sigma_{xy} + C_2))}{(({\mu_x}^2 + {\mu_y}^2 + C_1)\ \cdot\ ({\sigma_x}^2 + {\sigma_y}^2 + C_2))}$ | A structural metric that evaluates image degradation in terms of luminance, contrast, and structural similarity between the reconstructed and reference images. |
| Correlation | $corr = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$ | A radiometric metric that assesses whether the spatial variations and patterns present in the reference image are preserved in the super-resolved image. |
| Spectral Angle Mapper | $SAM = arccos\left(\frac{\sum_{j=1}^{B} y_j x_j}{\sqrt{\sum_{j=1}^{B} y_j^2}\sqrt{\sum_{j=1}^{B} x_j^2}}\right)$ | A global spectral metric that calculates the angle between the reference spectral vector and the reconstructed spectral vector, allowing the preservation of spectral relationships among bands to be assessed. |
| Relative Dimensionless Global Error of Synthesis | $ERGAS = 100 \cdot \frac{h}{l}\sqrt{\frac{1}{B}\sum_{j=1}^{B}\left(\frac{RMSE_j}{\mu_j}\right)^2}$ | A global spectral metric that combines the reconstruction errors of all spectral bands and their mean values to generate an indicator describing the overall reconstruction quality of the super-resolved image [72]. |

$x_i$ denotes the super-resolved image; $y_i$ represents the reference image; $i$ denotes the pixel under analysis; N is the total number of pixels per band; $max_I = 2$ is the dynamic range of the vegetation indices, calculated as the difference between their maximum and minimum possible values; MSE is the mean squared error; $\mu_x$ and $\mu_y$ are the mean values of the super-resolved and reference images, respectively; $\sigma_x^2$ y $\sigma_y^2$ are the variances of the super-resolved and reference images, respectively; $\sigma_{xy}$ denotes the covariance between the pixel values of the two images; $C_1 = (K_1 \cdot L)^2$ and $C_2 = (K_2 \cdot L)^2$ are stabilization constants, with $K_1 = 0.01$, $K_2 = 0.03$ y L = 2; $\overline{x}$ and $\overline{y}$ are the mean pixel values of the super-resolved and reference images, respectively; B = 5 is the number of spectral bands; j denotes the band under analysis; $\frac{h}{l} = 1$ is the ratio between the spatial resolutions of the high-resolution ($h$) and low-resolution ($l$) images; and $\mu_j$ is the mean pixel value of band $j$.

## 4. Results and Discussion

### *4.1. Computational Complexity of the Evaluated Models*

The computational complexity of the DeepSent, TVDSR, UAV-Net, and S$^3$-ESRGAN architectures was evaluated in terms of the number of trainable parameters, training time, inference time, processing throughput, peak memory consumption, and computational cost expressed in GFLOPs. Computational performance was evaluated on an NVIDIA GeForce RTX 4090 using Python 3.10.19, PyTorch 2.5.1, and CUDA 12.1. Inference was performed with a batch size of one and 96 × 96 pixel inputs using FP32 precision. DeepSent received an eight-date sequence, TVDSR a four-date sequence, UAV-Net three five-channel images, and S$^3$-ESRGAN a single five-channel image. Inference time was measured over 100 forward passes after 10 warm-up iterations, with CUDA synchronization before and after each measurement. Reported latency corresponds exclusively to the forward pass and excludes image loading, preprocessing, data transfer, and output writing. Throughput was calculated as the number of output patches processed per second, while peak memory corresponds to the maximum GPU memory allocated during inference. GFLOPs were estimated for a single forward pass using the PyTorch profiler.

The results showed important differences between the models (Table 4). DeepSent presented the most compact architecture and the highest processing throughput. In contrast, UAV-Net presented the largest number of trainable parameters, approximately 21 times more than the TVDSR model. However, UAV-Net exhibited the lowest computational cost and the shortest training time. This behavior is attributed to the downsampling strategy implemented in its encoder module, which progressively reduces the spatial size of the feature maps and decreases the number of operations required during forward propagation. The TVDSR model exhibited an intermediate level of computational complexity, requiring more training time and computational resources than DeepSent and UAV-Net, but less than S$^3$-ESRGAN.

**Table 4.** Performance and Computational Cost of the Models.

| Model | Trainable Parameters | Training Time (min) | Inference Time (s) | Throughput (Patches/s) | Peak Memory (MB) | Computational Cost (GFLOPs) |
|---|---|---|---|---|---|---|
| DeepSent | <u>882,437</u> | <u>32.20</u> | 4.4 | <u>518.62</u> | <u>79.76</u> | <u>25.67</u> |
| TVDSR | 1,288,304 | 71.40 | <u>4.2</u> | 169.58 | 82.72 | 74.99 |
| UAV-Net | **27,001,285** | **30.63** | **4.1** | **313.98** | **147.40** | **10.25** |
| S$^3$-ESRGAN | 9,845,801 | 406.50 | 4.3 | 69.50 | 66.15 | 160.24 |

GFLOPs: Giga Floating-Point Operations; iter: iteration; MB: megabytes; min: minutes; s: seconds. Inference time corresponds to the total time required by the model to process the two complete evaluation images, from data loading and inference to the generation and saving of the output raster file. The bold values indicate the best-performing model, while the underlined values correspond to the second best-performing model.

### *4.2. Performance of the Super-Resolution Models*

Table 5 presents the quantitative performance of the S$^3$-ESRGAN, DeepSent, UAV-Net, and TVDSR models for the evaluation dates. The results showed that UAV-Net achieved the best overall performance for the 2024 image, obtaining the lowest reconstruction errors and the highest spatial reconstruction metrics. Nevertheless, this superiority was not consistent across all evaluated properties. DeepSent achieved the lowest SAM value, indicating better preservation of the relationships among the reconstructed vegetation indices, whereas DeepSent and TVDSR obtained slightly higher correlation coefficients than UAV-Net. These results highlight that superior performance in spatial reconstruction metrics does not necessarily imply an equivalent preservation of the relationships among the reconstructed vegetation indices. This observation is consistent with the findings of Aybar et al. [75],

who reported that super-resolution methods may increase spatial detail at the expense of introducing radiometric distortions or spectral inconsistencies when such properties are not explicitly considered during training.

**Table 5.** Consolidated quantitative results for the two evaluation dates.

| Date | Model | RMSE | MAE | PSNR (dB) | SSIM | Corr | ERGAS | SAM (°) |
|---|---|---|---|---|---|---|---|---|
| 22 October 2024 | $S^3$-ESRGAN | 0.154 | 0.117 | 22.5 | 0.79 | 0.8 | 32.42 | 14.96 |
| 22 October 2024 | DeepSent | 0.152 | 0.128 | 22.45 | 0.85 | 0.88 | 32.05 | **9.86** |
| 22 October 2024 | UAV-Net | **0.128** | **0.1** | **24.06** | **0.86** | **0.87** | **27.06** | 12.89 |
| 22 October 2024 | TVDSR | 0.152 | 0.126 | 22.8 | 0.84 | 0.88 | 32.88 | 17.71 |
| 3 August 2025 | $S^3$-ESRGAN | 0.203 | 0.147 | 19.94 | 0.69 | 0.73 | 47.79 | 26.23 |
| 3 August 2025 | DeepSent | 0.182 | 0.129 | 20.83 | 0.78 | 0.8 | 43.13 | 19.58 |
| 3 August 2025 | UAV-Net | 0.181 | 0.139 | 20.92 | 0.75 | 0.78 | 41.76 | 23.29 |
| 3 August 2025 | TVDSR | **0.149** | **0.115** | **22.6** | **0.85** | **0.87** | **34.94** | **18.19** |

Corr: Correlation, dB: decibels, °: degrees. RMSE, MAE, PSNR, SSIM, and Corr were computed independently for each vegetation index, and the values reported correspond to the arithmetic mean across the five indices. Conversely, SAM and ERGAS were computed globally at the image level considering the complete five-index representation simultaneously. The bold values correspond to the model that achieved the best performance for each evaluated date.

The behavior observed on 3 August 2025 differed from that of the first evaluation date. TVDSR achieved the best overall performance, simultaneously obtaining the lowest reconstruction errors and the highest levels of structural similarity. These results suggest that the explicit incorporation of temporal dependencies through Bidirectional ConvLSTM improves reconstruction performance under conditions different from those observed during training. In this regard, Chang et al. [76] reported improvements in reconstruction quality through the incorporation of BiConvLSTM structures capable of integrating complementary information from multiple observations. Similarly, Michel et al. [77] demonstrated that the joint integration of multitemporal and multiresolution information can outperform approaches that independently exploit spatial, spectral, or temporal dimensions.

The change in performance between the two evaluation dates represents one of the most relevant findings of this study. In Sentinel-2 and UAV image fusion scenarios, where phenological dynamics, changes in vegetation cover, and radiometric differences among acquisitions can alter the spectral response of the surface, the ability to simultaneously exploit temporal dependencies from the historical observation sequence constitutes an important advantage for high-resolution image reconstruction. Overall, the results reveal a trade-off among spatial reconstruction accuracy, vegetation-index consistency, and temporal generalization capability. UAV-Net achieved the best spatial reconstruction performance on the first evaluation date, DeepSent exhibited superior preservation of vegetation-index consistency, and TVDSR demonstrated the most stable performance across dates and the best overall results in the second evaluation.

### *4.3. Temporal Generalization Capability*

Table 6 presents the variation in the evaluated metrics between the two validation dates. While $S^3$-ESRGAN, DeepSent, and UAV-Net experienced substantial degradation in several metrics, TVDSR maintained constant performance across both dates. In general, the increases observed in RMSE, MAE, ERGAS, and SAM, together with the reductions in PSNR, SSIM, and correlation, indicate that the August 2025 date represented a more challenging scenario for the super-resolution process. Since both evaluations correspond to different time points within the same temporal series, the results suggest that the changes occurring at the surface had a direct impact on the reconstruction capability of the models.

**Table 6.** Percentage variation in performance between the evaluation dates (2024–2025).

| Model | ΔRMSE (%) | ΔMAE (%) | ΔPSNR (%) | ΔSSIM (%) | ΔCorr (%) | ΔERGAS (%) | ΔSAM (%) |
|---|---|---|---|---|---|---|---|
| $S^3$-ESRGAN | 31.8 | 25.6 | −11.4 | −12.7 | −8.8 | 47.4 | 75.3 |
| DeepSent | 19.7 | 0.8 | −7.2 | −8.2 | −9.1 | 34.6 | 98.6 |
| UAV-Net | 41.4 | 39 | −13.1 | −12.8 | −10.3 | 54.3 | 80.7 |
| TVDSR | **−2** | **−8.7** | **−0.9** | **1.2** | **−1.1** | **6.3** | **2.7** |

Δ: variation; Corr: Correlation. The bold values correspond to the model that exhibited the lowest variation in performance across the two evaluation dates.

The results suggest important differences in the temporal generalization capability of the evaluated models. $S^3$-ESRGAN exhibited the largest performance degradation between evaluation dates, with substantial increases in both reconstruction errors and inter-index inconsistencies, suggesting that the incorporation of spatial information alone was insufficient to adequately capture the temporal variability of the scene. DeepSent showed a more stable behavior, particularly in terms of MAE, although it exhibited considerable increases in RMSE, ERGAS, and SAM, indicating a noticeable degradation in the reconstruction of the vegetation-index representation under different temporal conditions. UAV-Net, despite achieving the best overall performance in the 2024 evaluation, experienced substantial degradation across all metrics in the 2025 evaluation, demonstrating that excellent performance on a specific date does not necessarily guarantee strong temporal generalization capability. In contrast, TVDSR exhibited the smallest performance variations between dates, maintaining stable error, structural similarity, and vegetation-index reconstruction metrics. In particular, the changes observed in RMSE, PSNR, SSIM, correlation, ERGAS, and SAM were significantly smaller than those recorded for the other models, suggesting a greater capacity to handle temporal variability within the evaluated conditions.

Generalization capability remains one of the main challenges for super-resolution models applied to remote sensing. Xiong et al. [78] demonstrated that model performance may vary when models are transferred beyond the conditions used during training, particularly under changes in geographic location and sensor characteristics. The authors indicated that deep learning-based methods depend on the representativeness of the training data and often encounter difficulties when exposed to acquisition conditions different from those observed during the learning phase.

The temporal dimension is particularly important in agricultural environments because the spectral and structural characteristics of the surface evolve continuously because of phenological, environmental, and management-related processes. Xiao et al. [79] reported that spatiotemporal relationships must be adequately captured to maintain model performance under changing conditions. They further noted that an insufficient representation of these relationships limits the generalization capability of the model. Wang et al. [80] identified generalization and adaptation to real-world scenarios as key challenges in satellite image super-resolution. Similarly, Chen et al. [81] argued that a significant gap persists between training environments and real-world conditions due to the complexity and diversity of degradations encountered in practical applications. Consequently, high performance on a particular date does not necessarily imply an equivalent reconstruction capability when temporal changes occur in the observed scene.

From this perspective, the degradation observed in $S^3$-ESRGAN, DeepSent, and UAV-Net between the two evaluation dates can be interpreted as an indication of a limited ability to cope with unseen data distributions. The increases observed in RMSE, MAE, ERGAS, and SAM, together with the decreases in PSNR, SSIM, and correlation, suggest that the 2025 evaluation date represented a scenario further from the data distribution observed during training, making it more challenging to maintain reconstruction accuracy, structural

similarity, and consistency across the reconstructed vegetation indices. In contrast, TVDSR exhibited stable behavior, with minimal variations across all evaluated metrics. This stability suggests the presence of representations that are less dependent on the specific conditions observed during training and more resilient to temporal changes in the scene.

### *4.4. Performance of the Super-Resolution Models by Vegetation Index*

The effect of super-resolution on reflectance-derived variables was evaluated because numerous products used in agricultural monitoring, biomass estimation, crop water status assessment, and chlorophyll content analysis are computed from spectral indices. Therefore, the preservation of these indices constitutes a complementary indicator of the ability of the models to retain relevant biophysical information after the reconstruction process. To this end, five vegetation indices widely used in agriculture and remote sensing were considered: NDVI, SAVI, GNDVI, NDWI, and NDVIre. To summarize model performance across multiple evaluation metrics, a composite score was computed and defined as *Composite score* $= (1 - RMSE) + SSIM + Corr$. Figure 6 summarizes the results obtained for the two evaluation dates using heat maps.

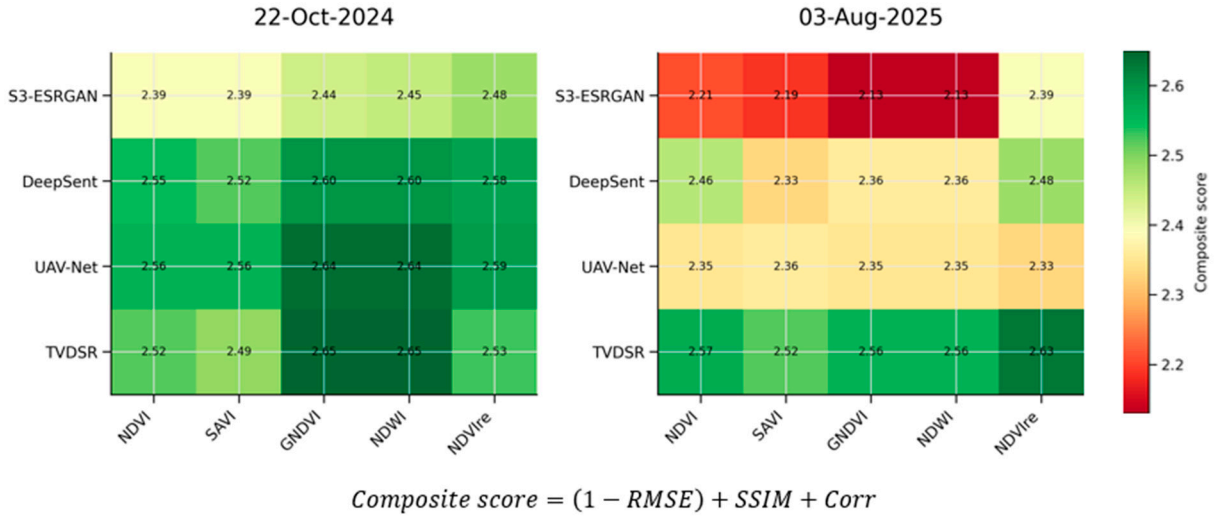



**Figure 6.** Heat map of the performance of the super-resolution models by vegetation index reconstruction.

The 2024 evaluation showed that UAV-Net achieved the best overall performance for most of the analyzed vegetation indices. The highest composite scores were obtained for NDVI, SAVI, and NDVIre, whereas for GNDVI and NDWI, UAV-Net produced results like those obtained by TVDSR. These findings indicate a strong ability to preserve the relationships among vegetation indices required for the accurate estimation of crop-condition descriptors under temporal conditions like those observed during training. DeepSent exhibited intermediate performance, maintaining consistent scores across all evaluated indices, with values ranging from 2.52 to 2.60. In contrast, $S^3$-ESRGAN achieved the lowest scores for all indices, particularly for NDVI and SAVI, reflecting a lower ability to preserve the relationships required for the estimation of biophysical variables.

The situation changed during the 2025 evaluation. On this date, TVDSR achieved the best performance across all five analyzed indices. The model obtained the highest composite scores, with the largest differences relative to the other models observed for NDVI and NDVIre. These results demonstrate a strong ability to preserve biophysical information even when the temporal conditions of the scene differ from those observed during training.

DeepSent ranked as the second-best model during this evaluation. Although its scores were lower than those of TVDSR, it maintained stable performance across all indices, particularly for NDVI and NDVIre. In contrast, UAV-Net experienced a substantial degradation relative to the first evaluation date. Its scores decreased for all indices, especially for NDVIre, where the score dropped from 2.59 to 2.33. A similar trend was observed for GNDVI and NDWI, whose scores decreased by approximately 11%. The most severe degradation was observed for $S^3$-ESRGAN, whose scores consistently decreased across all vegetation indices. The lowest values were recorded for GNDVI and NDWI, indicating a considerable loss in their ability to preserve the spectral relationships associated with chlorophyll content and vegetation water status.

Figure 7 presents the absolute error maps obtained by comparing the TVDSR reconstructions with the corresponding UAV reference images for each vegetation index. For the 2024 evaluation date, the largest errors are concentrated along field boundaries, agricultural roads, and transition zones between adjacent parcels. NDVI, SAVI, and NDVIre exhibit the highest error levels, which is consistent with the composite scores reported in Figure 6. In contrast, GNDVI and NDWI show lower reconstruction errors across most of the study area.

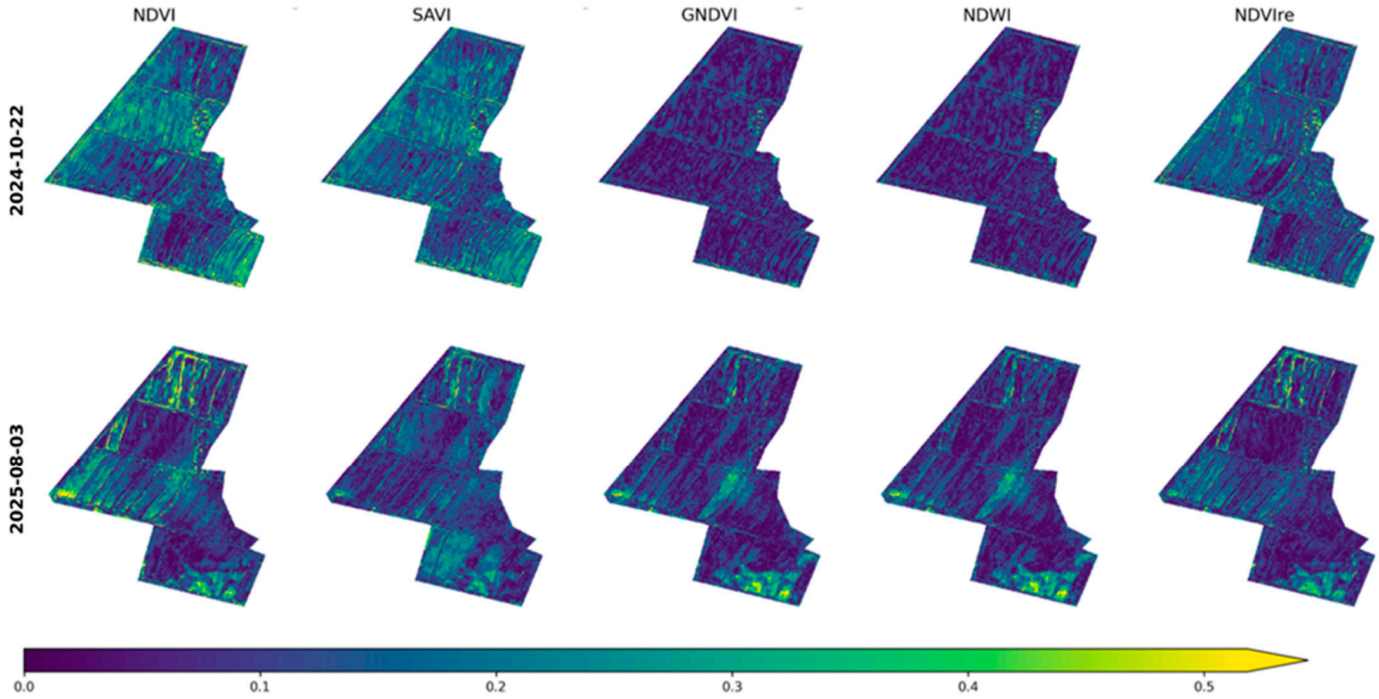



**Figure 7.** Absolute error maps by vegetation index of TVDSR model.

A similar spatial pattern is observed for the 2025 evaluation date, where the largest errors remain concentrated near roads, parcel boundaries, and other heterogeneous regions. Once again, GNDVI and NDWI exhibit comparatively lower error levels, whereas NDVI, SAVI, and NDVIre show greater sensitivity to local variations. The error distribution is not random; rather, it follows spatially structured patterns that are consistently associated with landscape discontinuities and areas characterized by higher spatial complexity. Conversely, large homogeneous regions located within individual agricultural parcels exhibit lower error values, suggesting that the dominant vegetation-index patterns are reproduced more consistently in these areas.

Several localized regions with increased error can also be observed consistently across the five vegetation indices, particularly in the 2025 evaluation date. The recurrence of these patterns among different indices suggests the presence of common sources of uncertainty,

potentially related to acquisition conditions, cross-sensor discrepancies, or residual cloud-related radiometric effects.

### 4.5. Visual Evaluation of the Reconstructed Images

Figure 8 presents a visual comparison between the reference UAV image, the original Sentinel-2 image, and the reconstructions generated by $S^3$-ESRGAN, DeepSent, UAV-Net, and TVDSR for the 3 August 2025 evaluation date. The highlighted regions allow the analysis of differences in the recovery of field boundaries, internal crop patterns, and areas affected by degradations present in the input image. In general, all models improved the visual quality relative to the original Sentinel-2 image. The reconstructed images recovered part of the spatial structure observed in the UAV reference image. However, reconstruction quality differed among the evaluated architectures, particularly in terms of field boundary definition and the preservation of vegetation-index variability within agricultural plots.

$S^3$-ESRGAN exhibited the weakest visual performance. Although the model increased the apparent spatial resolution of the scene, the reconstructions showed excessive smoothing, loss of fine details, and reduced definition of internal crop patterns. This behavior is particularly evident in Regions 2 and 3, where field boundaries appear less defined than in the UAV reference image. In addition, chromatic artifacts and anomalies can be observed in Region 4, suggesting greater sensitivity to the degradations present in the original image.

DeepSent achieved a more consistent recovery of the agricultural structure. The model reproduced field boundaries more faithfully, preserved field occupancy patterns, and maintained part of the internal heterogeneity of the crops. Regions 1 and 2 exhibit better spatial definition than those reconstructed by $S^3$-ESRGAN, while the gradients within the fields show a distribution more like that observed in the UAV image. In Region 3, field roads and parcel boundaries are clearly defined. UAV-Net, in turn, produced visually detailed reconstructions with an adequate representation of the overall scene structure. Field boundaries are clearly identifiable, although less defined than those generated by DeepSent. The reconstructed vegetation-index values remain in good agreement with the reference image. However, some agricultural textures appear more homogeneous than in the UAV image, and part of the degradation observed in Region 4 remains visible in the reconstruction.

The most favorable results were obtained with TVDSR, whose reconstruction exhibited the highest visual similarity to the UAV reference image. In Regions 1, 2, and 3, improved recovery of field boundaries, greater spatial continuity of agricultural structures, and a more detailed representation of crop variability can be observed. Furthermore, the spatial distribution of spectral patterns is more consistent with that observed in the high-resolution reference image. Notably, in Region 1, the reconstruction preserved field boundaries and access roads with a degree of spatial definition comparable to that achieved by DeepSent.

A particularly relevant observation can be made in Region 4, where the Sentinel-2 image contains an area affected by cloud-related or other radiometric effects. This anomaly remains clearly visible in the reconstructions generated by $S^3$-ESRGAN and UAV-Net, both of which preserve part of the degraded information present in the input image. In contrast, DeepSent and TVDSR show a partial reduction in the visual impact of the anomaly, generating vegetation-index patterns and tonal variations that are more consistent with those observed in the reference UAV image. However, this qualitative observation alone does not allow verification of the accurate recovery of the underlying crop signal in the affected area (Figure 7). Therefore, the observed improvements should be interpreted as an increase in visual consistency with the reference image rather than as definitive evidence of complete reconstruction of the original vegetation condition.

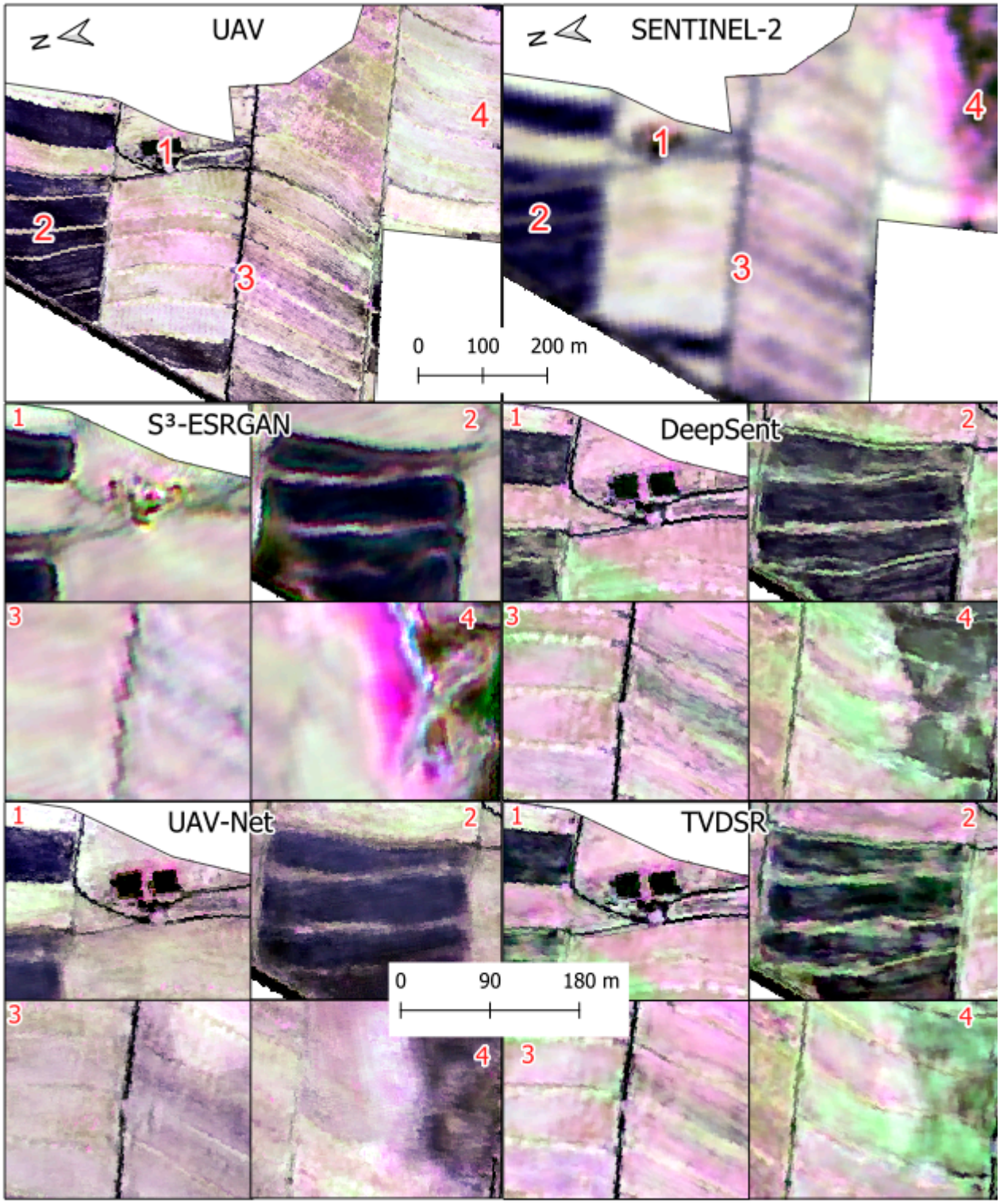



**Figure 8.** Visual comparison of the images reconstructed by the super-resolution models for 3 August 2025, evaluation date. False-color composites were generated using NDVI (R), SAVI (G), and GNDVI (B). All images were displayed using the same min–max stretch to ensure a consistent visual comparison among models. Regions 1–4 indicate areas of interest used for qualitative assessment.

Overall, the visual evaluation confirms the trends previously observed through the image quality metrics and vegetation index analyses. While all models improved the spatial resolution of the Sentinel-2 imagery, TVDSR and DeepSent demonstrated a greater ability to simultaneously preserve spatial structure and vegetation-index consistency under more

complex temporal conditions. The ability of TVDSR to recover fine details and partially reduce the effects of local degradations reinforces the quantitative results presented in the previous sections, where this model exhibited the most consistent performance across the evaluated observation dates.

The architectural components incorporated into TVDSR were selected based on previous evidence reported in the literature. Although the contribution of each component was not evaluated separately in the present study, their combined integration resulted in improved performance relative to the compared approaches. A systematic analysis of the individual contribution of temporal modeling, bidirectional processing, seasonal FiLM conditioning, and attention mechanisms would provide valuable insight into their respective impact on reconstruction performance. Such an analysis could also guide future architectural refinement and contribute to the development of more efficient and effective versions of TVDSR.

Although the obtained results demonstrate the potential of the evaluated models for Sentinel-2 image super-resolution using UAV information, there remain aspects that could be improved in future research. One of the main limitations of this study is the availability of training data, as only 13 UAV images were available, ten acquired during 2024 and three during 2025. Although this dataset was sufficient to successfully train the models and generate high-quality reconstructions, the limited temporal representation may have reduced their ability to fully capture the phenological and radiometric variability present within the study area. This limitation may have contributed, at least partially, to the performance degradation observed in some models during the August 2025 evaluation.

Nevertheless, although TVDSR exhibited the most consistent performance across the evaluated dates and showed the lowest performance degradation among the compared models, the available results do not yet allow a definitive conclusion that the architecture is inherently robust to temporal variability. The evaluation was conducted using only two observation dates, which limits the ability to adequately characterize its behavior under phenological, climatic, and management-related changes throughout complete agricultural cycles. Consequently, future studies should incorporate a larger number of acquisition dates distributed across different phenological stages and growing seasons to determine whether the observed stability represents a consistent property of the model or a behavior specific to the conditions evaluated in this study.

Furthermore, the transferability of the proposed framework to other crops, regions, soil types, and climatic zones remains to be evaluated. The present study was conducted using data from a single agricultural area; therefore, the spatial generalization capability of the evaluated models still requires assessment under more diverse conditions. Differences in crop type, management practices, climatic conditions, soil properties, and phenological patterns may influence the performance of super-resolution architectures when applied to other regions. Future studies should also incorporate additional acquisition dates, repeated training runs, and spatially independent validation schemes to enable more rigorous evaluation of model generalization and uncertainty. In addition, the impact of residual cloud contamination and cloud-shadow effects on vegetation-index reconstruction should be investigated, particularly in scenarios where cloud-affected pixels are retained rather than masked during preprocessing. Comparative analyses between cloud-screened and non-cloud-screened datasets would help quantify the extent to which cloud-related distortions affect reconstruction accuracy and model generalization. Finally, future work should evaluate the scalability of the proposed framework under operational conditions, including processing times for large agricultural areas, different hardware configurations, and near-real-time monitoring scenarios.

## 5. Conclusions

This study demonstrated the feasibility of using deep learning-based super-resolution models to enhance the spatial resolution of Sentinel-2 imagery using UAV information. The results confirm that these approaches can generate improved vegetation-index products for crop-monitoring applications. Furthermore, the study showed that the evaluation of super-resolution models should not be limited to image quality metrics alone. The preservation of vegetation-index information constitutes a complementary evaluation criterion when the generated products are intended for crop monitoring applications, as it allows the assessment of whether the reconstructed vegetation-index representation remains suitable for subsequent crop-monitoring analyses.

The results highlight the importance of considering the temporal dimension when evaluating super-resolution models for agricultural remote sensing. The differences observed between evaluation dates suggest that model performance is influenced by complex spatiotemporal factors, including crop phenological development, interannual variability, changing environmental conditions, and acquisition-related effects. Consequently, conventional validation procedures based on a limited set of dates may not fully capture model behavior under operational conditions. In this regard, the incorporation of temporal validation protocols provides a more realistic and comprehensive assessment of the practical applicability and generalization capability of these methodologies.

TVDSR exhibited consistent performance throughout the conducted evaluations, suggesting that the explicit incorporation of temporal and attention mechanisms represents a promising strategy for the development of super-resolution models intended for vegetation-index reconstruction and crop-monitoring applications. However, the evidence obtained should be considered preliminary, as the temporal evaluation was performed using only two acquisition dates. Therefore, a larger set of observations distributed across different phenological stages, growing seasons, and environmental conditions is required to determine the model's temporal generalization capability more reliably.

This work provides evidence supporting the need to integrate reconstruction, structural, vegetation-index, and temporal criteria within unified evaluation frameworks. Such an approach may contribute to the development of more robust super-resolution models and facilitate their adoption in multitemporal crop monitoring applications based on Sentinel-2 and UAV imagery.

**Author Contributions:** Conceptualization, A.M.-C. (Alvaro Murguia-Cozar), A.M.-C. (Antonia Macedo-Cruz) and G.d.J.L.-C.; methodology, A.M.-C. (Alvaro Murguia-Cozar), A.M.-C. (Antonia Macedo-Cruz), G.d.J.L.-C. and N.V.-L.; software, A.M.-C. (Alvaro Murguia-Cozar); validation, A.M.-C. (Alvaro Murguia-Cozar), A.M.-C. (Antonia Macedo-Cruz), G.d.J.L.-C., N.V.-L. and R.E.O.-C.; formal analysis, A.M.-C. (Alvaro Murguia-Cozar), A.M.-C. (Antonia Macedo-Cruz), G.d.J.L.-C., N.V.-L. and R.E.O.-C.; investigation, A.M.-C. (Alvaro Murguia-Cozar); resources, A.M.-C. (Alvaro Murguia-Cozar); data curation, A.M.-C. (Alvaro Murguia-Cozar); writing—original draft preparation, A.M.-C. (Alvaro Murguia-Cozar); writing—review and editing, A.M.-C. (Antonia Macedo-Cruz), G.d.J.L.-C., N.V.-L. and R.E.O.-C.; visualization, A.M.-C. (Alvaro Murguia-Cozar); supervision, A.M.-C. (Antonia Macedo-Cruz), G.d.J.L.-C., N.V.-L. and R.E.O.-C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Universidad Autónoma Chapingo, under conventional research projects 26209-C-65 and 25069-C-65.

**Data Availability Statement:** The datasets used and the models evaluated are available at: https://github.com/AlvaroMC89/CNN-Based-Temporal-Fusion-of-UAV-and-Sentinel-2-Imagery-.git (accessed on 9 November 2026).

**Acknowledgments:** The authors would like to thank the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) for the scholarship provided. We also gratefully acknowledge the Universidad Autónoma Chapingo for providing the facilities and the UAV equipment used in this research.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Xiao, J.; Aggarwal, A.K.; Rage, U.K.; Katiyar, V.; Avtar, R. Deep Learning-Based Spatiotemporal Fusion of Unmanned Aerial Vehicle and Satellite Reflectance Images for Crop Monitoring. *IEEE Access* **2023**, *11*, 85600–85614. [CrossRef]
2. Allu, A.R.; Mesapam, S. Fusion of Satellite and UAV Imagery for Crop Monitoring. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2025**, *X-G-2025*, 71–79. [CrossRef]
3. Swain, R.; Paul, A.; Behera, M.D. Spatio-Temporal Fusion Methods for Spectral Remote Sensing: A Comprehensive Technical Review and Comparative Analysis. *Trop. Ecol.* **2024**, *65*, 356–375. [CrossRef]
4. Irani, M.; Peleg, S. Super Resolution from Image Sequences. In Proceedings of the 10th International Conference on Pattern Recognition (ICPR), Atlantic City, NJ, USA, 16–21 June 1990; Volume 2, pp. 115–120.
5. Elad, M.; Feuer, A. Restoration of a Single Superresolution Image from Several Blurred, Noisy, and Undersampled Measured Images. *IEEE Trans. Image Process.* **1997**, *6*, 1646–1658. [CrossRef] [PubMed]
6. Pohl, C.; Van Genderen, J.L. Review Article Multisensor Image Fusion in Remote Sensing: Concepts, Methods and Applications. *Int. J. Remote Sens.* **1998**, *19*, 823–854. [CrossRef]
7. Maimaitijiang, M.; Sagan, V.; Sidike, P.; Daloye, A.M.; Erkbol, H.; Fritschi, F.B. Crop Monitoring Using Satellite/UAV Data Fusion and Machine Learning. *Remote Sens.* **2020**, *12*, 1357. [CrossRef]
8. Bazrafkan, A.; Igathinathane, C.; Bandillo, N.; Flores, P. Optimizing Integration Techniques for UAS and Satellite Image Data in Precision Agriculture—A Review. *Front. Remote Sens.* **2025**, *6*, 1622884. [CrossRef]
9. Fernandez-Beltran, R.; Latorre-Carmona, P.; Pla, F. Single-Frame Super-Resolution in Remote Sensing: A Practical Overview. *Int. J. Remote Sens.* **2017**, *38*, 314–354. [CrossRef]
10. Xie, C.; Xing, F.; Zhu, Y.; Ye, S. Multi-Scale Partial Convolution Residual Distillation Network for Efficient Image Super-Resolution. *Signal Image Video Process.* **2026**, *20*, 360. [CrossRef]
11. Dong, C.; Loy, C.C.; He, K.; Tang, X. *Learning a Deep Convolutional Network for Image Super-Resolution*; Springer: Cham, Switzerland, 2014; pp. 184–199.
12. Kim, J.; Lee, J.K.; Lee, K.M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2016; pp. 1646–1654.
13. Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M. Enhanced Deep Residual Networks for Single Image Super-Resolution. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: Piscataway, NJ, USA, 2017; pp. 1132–1140.
14. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Loy, C.C. *ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks*; Springer: Cham, Switzerland, 2019; pp. 63–79.
15. Wang, X.; Xie, L.; Dong, C.; Shan, Y. Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*; IEEE: Piscataway, NJ, USA, 2021; pp. 1905–1914.
16. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. SwinIR: Image Restoration Using Swin Transformer. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*; IEEE: Piscataway, NJ, USA, 2021; pp. 1833–1844.
17. Wang, Y.; Yang, W.; Chen, X.; Wang, Y.; Guo, L.; Chau, L.-P.; Liu, Z.; Qiao, Y.; Kot, A.C.; Wen, B. SinSR: Diffusion-Based Image Super-Resolution in a Single Step. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024; pp. 25796–25805.
18. Zhang, J.; Xu, T.; Li, J.; Jiang, S.; Zhang, Y. Single-Image Super Resolution of Remote Sensing Images with Real-World Degradation Modeling. *Remote Sens.* **2022**, *14*, 2895. [CrossRef]
19. Cao, H.; Tan, Z.; Zhu, B.; Xiong, H.; Lv, Q. A Review on Super-Resolution Reconstruction of Single-Frame Remote Sensing Images via Diffusion Models. *Remote Sens.* **2026**, *18*, 1702. [CrossRef]
20. Ding, X.; Wang, W.; Zhang, B.; Zhang, J. Attention-Based Multi-Image Super-Resolution Reconstruction for Remote Sensing. *Signal Image Video Process.* **2025**, *19*, 455. [CrossRef]
21. Liu, W.; Che, S.; Wang, W.; Du, Y.; Tuo, Y.; Zhang, Z. Research on Remote Sensing Multi-Image Super-Resolution Based on ESTF2N. *Sci. Rep.* **2025**, *15*, 9501. [CrossRef] [PubMed]
22. Elad, M.; Feuer, A. Super-Resolution Reconstruction of Image Sequences. *IEEE Trans. Pattern Anal. Mach. Intell.* **1999**, *21*, 817–834. [CrossRef]

23. Farsiu, S.; Robinson, M.D.; Elad, M.; Milanfar, P. Fast and Robust Multiframe Super Resolution. *IEEE Trans. Image Process.* **2004**, *13*, 1327–1344. [CrossRef] [PubMed]
24. Protter, M.; Elad, M.; Takeda, H.; Milanfar, P. Generalizing the Nonlocal-Means to Super-Resolution Reconstruction. *IEEE Trans. Image Process.* **2009**, *18*, 36–51. [CrossRef] [PubMed]
25. Takeda, H.; Milanfar, P.; Protter, M.; Elad, M. Super-Resolution without Explicit Subpixel Motion Estimation. *IEEE Trans. Image Process.* **2009**, *18*, 1958–1975. [CrossRef] [PubMed]
26. Liao, R.; Tao, X.; Li, R.; Ma, Z.; Jia, J. Video Super-Resolution via Deep Draft-Ensemble Learning. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2015; pp. 531–539.
27. Kappeler, A.; Yoo, S.; Dai, Q.; Katsaggelos, A.K. Video Super-Resolution With Convolutional Neural Networks. *IEEE Trans. Comput. Imaging* **2016**, *2*, 109–122. [CrossRef]
28. Tao, X.; Gao, H.; Liao, R.; Wang, J.; Jia, J. Detail-Revealing Deep Video Super-Resolution. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2017; pp. 4482–4490.
29. Haris, M.; Shakhnarovich, G.; Ukita, N. Recurrent Back-Projection Network for Video Super-Resolution. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2019; pp. 3892–3901.
30. Wang, X.; Chan, K.C.K.; Yu, K.; Dong, C.; Loy, C.C. EDVR: Video Restoration With Enhanced Deformable Convolutional Networks. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: Piscataway, NJ, USA, 2019; pp. 1954–1963.
31. Deudon, M.; Kalaitzis, A.; Goytom, I.; Arefin, M.R.; Lin, Z.; Sankaran, K.; Michalski, V.; Kahou, S.E.; Cornebise, J.; Bengio, Y. HighRes-Net: Recursive Fusion for Multi-Frame Super-Resolution of Satellite Imagery. *arXiv* **2020**, arXiv:2002.06460. [CrossRef]
32. Chan, K.C.K.; Wang, X.; Yu, K.; Dong, C.; Loy, C.C. BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2021; pp. 4945–4954.
33. Chan, K.C.K.; Zhou, S.; Xu, X.; Loy, C.C. BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2022; pp. 5962–5971.
34. Jo, Y.; Oh, S.W.; Kang, J.; Kim, S.J. Deep Video Super-Resolution Network Using Dynamic Upsampling Filters Without Explicit Motion Compensation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 3224–3232.
35. Bordone Molini, A.; Valsesia, D.; Fracastoro, G.; Magli, E. DeepSUM: Deep Neural Network for Super-Resolution of Unregistered Multitemporal Images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 3644–3656. [CrossRef]
36. Molini, A.B.; Valsesia, D.; Fracastoro, G.; Magli, E. Deepsum++: Non-Local Deep Neural Network for Super-Resolution of Unregistered Multitemporal Images. In *Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium*; IEEE: Piscataway, NJ, USA, 2020; pp. 609–612.
37. Bhat, G.; Danelljan, M.; Van Gool, L.; Timofte, R. Deep Burst Super-Resolution. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2021; pp. 9205–9214.
38. Luo, Z.; Yu, L.; Mo, X.; Li, Y.; Jia, L.; Fan, H.; Sun, J.; Liu, S. EBSR: Feature Enhanced Burst Super-Resolution with Deformable Alignment. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*; IEEE: Piscataway, NJ, USA; pp. 471–478.
39. Bhat, G.; Danelljan, M.; Yu, F.; Van Gool, L.; Timofte, R. Deep Reparametrization of Multi-Frame Super-Resolution and Denoising. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2021; pp. 2440–2450.
40. Dudhane, A.; Zamir, S.W.; Khan, S.; Khan, F.S.; Yang, M.-H. Burst Image Restoration and Enhancement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 9454–9467. [CrossRef] [PubMed]
41. Di, X.; Peng, L.; Xia, P.; Li, W.; Pei, R.; Cao, Y.; Wang, Y.; Zha, Z.-J. QMambaBSR: Burst Image Super-Resolution with Query State Space Model. In *Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2025; pp. 23080–23090.
42. Zhang, Y.; Chen, Z.; Liu, S. Video Super Resolution Using Temporal Encoding ConvLSTM and Multi-Stage Fusion. In *Proceedings of the 2020 IEEE International Conference on Visual Communications and Image Processing, VCIP 2020*; IEEE: Piscataway, NJ, USA, 2020; pp. 298–301.
43. Li, C.; Song, S.; Wang, X.; Liu, Y.; Bi, L. RCFNC: A Resolution and Contrast Fusion Network with ConvLSTM for Low-Light Image Enhancement. *Vis. Comput.* **2024**, *40*, 2793–2806. [CrossRef]
44. Heng, H.; Ye, H.; Huang, R. Defocus Blur Detection by Fusing Multiscale Deep Features with Conv-LSTM. *IEEE Access* **2020**, *8*, 97279–97288. [CrossRef]
45. Fu, C.; Yuan, H.; Jiang, S.; Zhang, G.; Shen, L.; Hamzaoui, R. Global Spatial-Temporal Information-Based Residual ConvLSTM for Video Space-Time Super-Resolution. *IEEE Trans. Multimed.* **2025**, *27*, 5212–5224. [CrossRef]

46. An, T.; Zhang, X.; Huo, C.; Xue, B.; Wang, L.; Pan, C. TR-MISR: Multiimage Super-Resolution Based on Feature Fusion With Transformers. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1373–1388. [CrossRef]
47. Li, J.; Lv, Q.; Zhang, W.; Zhu, B.; Zhang, G.; Tan, Z. Multi-Attention Multi-Image Super-Resolution Transformer (MAST) for Remote Sensing. *Remote Sens.* **2023**, *15*, 4183. [CrossRef]
48. Aira, L.S.; Valsesia, D.; Molini, A.B.; Fracastoro, G.; Magli, E.; Mirabile, A. Deep 3D World Models for Multi-Image Super-Resolution Beyond Optical Flow. *IEEE Access* **2024**, *12*, 188902–188913. [CrossRef]
49. Choi, Y.J.; Kim, B.G. Hierarchical Recurrent Transformer Network for Video Super-Resolution. *Eng. Appl. Artif. Intell.* **2026**, *166*, 113714. [CrossRef]
50. Wang, X.; Wang, D.; Song, C.; Zhang, X.; Liu, L.; Sun, M.; Cai, S.; Han, S.; Liu, L.; Zhang, Y. FRCFD-Net: Multi-Frame Super-Resolution through Two-Branch Feature Refinement, Cross-Layer Fusion and Difference Enhancement. *Neurocomputing* **2026**, *681*, 133412. [CrossRef]
51. Lanaras, C.; Bioucas-Dias, J.; Galliani, S.; Baltsavias, E.; Schindler, K. Super-Resolution of Sentinel-2 Images: Learning a Globally Applicable Deep Neural Network. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 305–319. [CrossRef]
52. Palsson, F.; Sveinsson, J.R.; Ulfarsson, M.O. Sentinel-2 Image Fusion Using a Deep Residual Network. *Remote Sens.* **2018**, *10*, 1290. [CrossRef]
53. Gargiulo, M.; Mazza, A.; Gaetano, R.; Ruello, G.; Scarpa, G. Fast Super-Resolution of 20 m Sentinel-2 Bands Using Convolutional Neural Networks. *Remote Sens.* **2019**, *11*, 2635. [CrossRef]
54. Galar, M.; Sesma, R.; Ayala, C.; Albizua, L.; Aranda, C. Super-Resolution of Sentinel-2 Images Using Convolutional Neural Networks and Real Ground Truth Data. *Remote Sens.* **2020**, *12*, 2941. [CrossRef]
55. Latte, N.; Lejeune, P. PlanetScope Radiometric Normalization and Sentinel-2 Super-Resolution (2.5 m): A Straightforward Spectral-Spatial Fusion of Multi-Satellite Multi-Sensor Images Using Residual Convolutional Neural Networks. *Remote Sens.* **2020**, *12*, 2366. [CrossRef]
56. Jia, D.; Song, C.; Cheng, C.; Shen, S.; Ning, L.; Hui, C. A Novel Deep Learning-Based Spatiotemporal Fusion Method for Combining Satellite Images with Different Resolutions Using a Two-Stream Convolutional Neural Network. *Remote Sens.* **2020**, *12*, 698. [CrossRef]
57. Tarasiewicz, T.; Nalepa, J.; Farrugia, R.A.; Valentino, G.; Chen, M.; Briffa, J.A.; Kawulok, M. Multitemporal and Multispectral Data Fusion for Super-Resolution of Sentinel-2 Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4404514. [CrossRef]
58. Zhang, H.; Sun, Y.; Shi, W.; Guo, D.; Zheng, N. An Object-Based Spatiotemporal Fusion Model for Remote Sensing Images. *Eur. J. Remote Sens.* **2021**, *54*, 86–101. [CrossRef]
59. Salgueiro Romero, L.; Marcello, J.; Vilaplana, V. Super-Resolution of Sentinel-2 Imagery Using Generative Adversarial Networks. *Remote Sens.* **2020**, *12*, 2424. [CrossRef]
60. Wang, Z.; Tan, H.; Wang, Z.; Ci, J.; Zhai, H. MAF-GAN: A Multi-Attention Fusion Generative Adversarial Network for Remote Sensing Image Super-Resolution. *Remote Sens.* **2025**, *17*, 3959. [CrossRef]
61. Farah, S.; Saadaoui, H.; Rhinane, H. Enhancing Spatial Resolution of Sentinel-2 Imagery through Deep Learning and Generative Adversarial Networks: GS-SRGAN. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2026**, *XLVIII-4/W17-2025*, 123–134. [CrossRef]
62. Sarmad, M.; Kampffmeyer, M.; Salberg, A.-B. DiffFuSR: Super-Resolution of All Sentinel-2 Multispectral Bands Using Diffusion Models. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4709813. [CrossRef]
63. Mazzia, V.; Comba, L.; Khaliq, A.; Chiaberge, M.; Gay, P. UAV and Machine Learning Based Refinement of a Satellite-Driven Vegetation Index for Precision Agriculture. *Sensors* **2020**, *20*, 2530. [CrossRef] [PubMed]
64. Masrur, A.; Olsen, P.; Jackson, C.; Adler, P. Learning to See More: A Spectral Extension Super-Resolution Framework for Harmonized Satellite-UAS Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2026**, *19*, 8740–8755. [CrossRef]
65. Toosi, A.; Samadzadegan, F.; Javan, F.D. $S^3$-ESRGAN: Enhanced Super-Resolution Generative Adversarial Network for Remote Sensing Imagery Spatial Resolution Improvement—An Application Using Sentinel-2 and UAV Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2026**, *19*, 2149–2172. [CrossRef]
66. Ren, A.; Zhou, X.; Li, G.; Li, J.; Ma, Y.; Qiao, L.; Yan, J.; Shen, J. Integrated UAV and Satellite Remote Sensing: A Survey on Integration Strategies and Data Fusion Methods. In *Proceedings of the IGARSS 2025—2025 IEEE International Geoscience and Remote Sensing Symposium*; IEEE: Piscataway, NJ, USA, 2025; pp. 8835–8840.
67. Enriqueta, G.-A. *Modificaciones al Sistema de Clasificación Climática de Köppen*; Universidad Nacional Autónoma de México Press: Mexico City, Mexico, 2004; ISBN 970-32-1010-4.
68. Murguia-Cozar, A.; López-Canteñs, G.d.J.; Macedo-Cruz, A.; Velazquez-López, N.; Ontiveros-Capurata, R.E. Fusion of UAV–Sentinel-2 Images Using Convolutional Neural Networks for Remote-Sensing Analysis. *Int. J. Image Data Fusion* **2026**, *17*, 2701715. [CrossRef]

69. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.; Woo, W. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In *Proceedings of the NIPS'15: Proceedings of the 29th International Conference on Neural Information Processing Systems*; Cortes, C., Lee, D.D., Sugiyama, M., Garnett, R., Eds.; MIT Press: Montreal, QC, Canada, 2015; Volume 1.
70. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In *Proceedings of the NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*; von Luxburg, U., Guyon, I., Bengio, S., Wallach, H., Fergus, R., Eds.; Curran Associates Inc.: Long Beach, CA, USA, 2017; pp. 6000–6010.
71. Perez, E.; Strub, F.; de Vries, H.; Dumoulin, V.; Courville, A. FiLM: Visual Reasoning with a General Conditioning Layer. In *Proceedings of the AAAI'18/IAAI'18/EAAI'18: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*; Mcllraith, S.A., Weinberger, K.Q., Eds.; AAAI Press: New Orleans, LA, USA, 2018; pp. 3942–3951.
72. Wald, L. *Data Fusion. Definitions and Architectures-Fusion of Images of Different Spatial Resolutions*; Presses de l'Ecole, Ecole des Mines de Paris: Paris, France, 2002.
73. Lin, P.-L.; Huang, P.-Y. Fusion Methods Based on Dynamic-Segmented Morphological Wavelet or Cut and Paste for Multifocus Images. *Signal Process.* **2008**, *88*, 1511–1527. [CrossRef]
74. Liu, Z.; Forsyth, D.S.; Laganière, R. A Feature-Based Metric for the Quantitative Evaluation of Pixel-Level Image Fusion. *Comput. Vis. Image Underst.* **2008**, *109*, 56–68. [CrossRef]
75. Aybar, C.; Contreras, J.; Donike, S.; Portalés-Julià, E.; Mateo-García, G.; Gómez-Chova, L. A Radiometrically and Spatially Consistent Super-Resolution Framework for Sentinel-2. *Remote Sens. Environ.* **2026**, *334*, 115222. [CrossRef]
76. Chang, Y.; Luo, B. Bidirectional Convolutional LSTM Neural Network for Remote Sensing Image Super-Resolution. *Remote Sens.* **2019**, *11*, 2333. [CrossRef]
77. Michel, J.; Inglada, J. Temporal Attention Multi-Resolution Fusion of Satellite Image Time-Series, Applied to Landsat-8/9 and Sentinel-2: All Bands, Any Time, at Best Spatial Resolution. *Remote Sens. Environ.* **2026**, *334*, 115159. [CrossRef]
78. Xiong, Y.; Guo, S.; Chen, J.; Deng, X.; Sun, L.; Zheng, X.; Xu, W. Improved SRGAN for Remote Sensing Image Super-Resolution across Locations and Sensors. *Remote Sens.* **2020**, *12*, 1263. [CrossRef]
79. Xiao, Y.; Su, X.; Yuan, Q.; Liu, D.; Shen, H.; Zhang, L. Satellite Video Super-Resolution via Multiscale Deformable Convolution Alignment and Temporal Grouping Projection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5606015. [CrossRef]
80. Wang, X.; Yi, J.; Guo, J.; Song, Y.; Lyu, J.; Xu, J.; Yan, W.; Zhao, J.; Cai, Q.; Min, H. A Review of Image Super-Resolution Approaches Based on Deep Learning and Applications in Remote Sensing. *Remote Sens.* **2022**, *14*, 5423. [CrossRef]
81. Chen, H.; Li, W.; Gu, J.; Ren, J.; Sun, H.; Zou, X.; Zhang, Z.; Yan, Y.; Zhu, L. Low-Res Leads the Way: Improving Generalization for Super-Resolution by Self-Supervised Learning. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024; pp. 25857–25867.


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.