

## Article

# From RGB to Synthetic NIR: Image-to-Image Translation for Pineapple Crop Monitoring Using Pix2PixHD

Darío Doria Usta <sup>1,\*</sup>, Ricardo Hundelshaussen <sup>1,\*</sup>, Carlos Martínez López <sup>1</sup>, Delio Salgado Chamorro <sup>1</sup>, César López Martínez <sup>1</sup>, João Felipe Coimbra Leite Costa <sup>2</sup> and Marcel Arcari Bassani <sup>2</sup>

<sup>1</sup> Escuela de Ingenierías y Arquitectura, Facultad de Ingeniería Industrial, Universidad Pontificia Bolivariana, Carrera 6 No. 97 A-99, Montería 230029, Colombia; carlos.martinezlo@upb.edu.co (C.M.L.); delio.salgado@upb.edu.co (D.S.C.); cesar.lopezma@upb.edu.co (C.L.M.)

<sup>2</sup> Departamento de Engenharia de Minas, Escola de Engenharia, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves, 9500-Bloco IV-Prédio 75-Sala 104, Bairro Agronomia, Porto Alegre 91501-970, Brazil; jfelipe@ufrgs.br (J.F.C.L.C.); marcel.bassani@ufrgs.br (M.A.B.)

\* Correspondence: dario.doria@upb.edu.co (D.D.U.); ricardo.hundelshaussen@upb.edu.co (R.H.)

## Abstract

Near-infrared (NIR) imaging plays a crucial role in precision agriculture; however, the high cost of multispectral sensors limits its widespread adoption. In this study, we generate synthetic NIR images (2592 × 1944 pixels) of pineapple crops from standard RGB drone imagery using the Pix2PixHD framework. The model was trained for 580 epochs, saving the first model after epoch 1 and then every 10 epochs thereafter. While models trained beyond epoch 460 achieved marginally higher metrics, they introduced visible artifacts. Model 410 was identified as the most effective, offering consistent quantitative performance while producing artifact-free results. Evaluation of Model 410 across 229 test images showed a mean SSIM of 0.6873, PSNR of 29.92, RMSE of 8.146, and PCC of 0.6565, indicating moderate to high structural similarity and reliable spectral accuracy of the synthetic NIR data. The proposed approach demonstrates that reliable NIR information can be obtained without expensive multispectral equipment, reducing costs and enhancing accessibility for farmers. By enabling advanced tasks such as vegetation segmentation and crop health monitoring, this work highlights the potential of deep learning-based image translation to support sustainable and data-driven agricultural practices. Future directions include extending the method to other crops, environmental conditions and real-time drone monitoring.



Academic Editor: Hongying Meng

Received: 22 October 2025

Revised: 19 November 2025

Accepted: 28 November 2025

Published: 5 December 2025

**Citation:** Doria Usta, D.; Hundelshaussen, R.; Martínez López, C.; Salgado Chamorro, D.; López Martínez, C.; Coimbra Leite Costa, J.F.; Arcari Bassani, M. From RGB to Synthetic NIR: Image-to-Image Translation for Pineapple Crop Monitoring Using Pix2PixHD.

*Technologies* **2025**, *13*, 569.

<https://doi.org/10.3390/technologies13120569>

**Copyright:** © 2025 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 (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** precision agriculture; Pix2PixHD; image-to-image translation; synthetic NIR generation; GANs; pineapple crop; multispectral images

## 1. Introduction

The increasing global population and the pressing need for sustainable food production have propelled the field of precision agriculture to the forefront of modern farming [1,2]. Unmanned Aerial Vehicles (UAVs), commonly known as drones, have become a pillar of this revolution by enabling the rapid, non-destructive, and cost-effective acquisition of extensive data from agricultural fields [2]. This remote sensing data is invaluable for tasks such as monitoring crop health, estimating yields, optimizing irrigation, and detecting pests and diseases early [1–4].

Central to these applications is the ability to analyze information beyond the visible spectrum. Near-Infrared (NIR) bands are particularly critical, as they provide rich spectral information about a plant's biological processes that are not perceptible to the human

eye [5]. For instance, healthy vegetation exhibits high reflectance in the NIR range, a property that forms the basis for essential vegetation indices like the Normalized Difference Vegetation Index (NDVI) [1,6–8]. However, the widespread adoption of NIR imaging is hindered by significant barriers. Dedicated multispectral cameras required to capture these bands are often excessively expensive—ranging from thousands to tens of thousands of dollars—and typically offer lower spatial resolution compared to standard RGB (Red, Green, and Blue) cameras. This financial and technical gap prevents many small and medium-scale producers from accessing powerful technology that could enhance their operations [4,6].

To overcome this limitation, a growing body of research is focused on synthesizing NIR data from readily available and low-cost RGB images [9]. This task, known as image-to-image (I2I) translation [2,10–13], aims to create a virtual NIR channel from a visible-light image, providing a practical and affordable alternative for agricultural monitoring. The generation of spectral images from RGB images for crop monitoring has evolved significantly over the years. Initial approaches were based on physical models or linear and polynomial regressions to estimate the NIR band from observed RGB pixel values [14]. However, these methods proved to be highly sensitive to variable environmental conditions—such as lighting, atmospheric humidity, or solar angle—and lack the ability to model the complex nonlinear relationships inherent in the plant structures of different crops and phenological stages, which severely limits their generalization [2,15]. Machine learning and, specifically, deep learning models have shown remarkable promise in this domain due to their ability to learn the complex, nonlinear correlations between RGB and NIR reflectance that physical models cannot easily define [6,15,16].

Nevertheless, most existing studies present limitations that restrict their applicability to real agricultural environments. Many are based on paired RGB–NIR datasets [6,15] that are difficult to acquire or produce low-resolution results that are unsuitable for precision tasks (typically  $\leq 1024 \times 1024$  pixels). Furthermore, validation is often limited to crops with simple canopies, while tropical species such as pineapple remain largely unexplored. These limitations highlight the need for crop-specific approaches that can be generalized to field conditions [7–9].

Recent research in I2I translation has been largely shaped by Generative Adversarial Networks (GANs) [17], a class of deep learning models in which a generator and a discriminator are trained in competition. Conditional GANs (cGANs) [18] demonstrated strong performance on tasks requiring paired datasets, often employing U-Net architectures [19] to preserve fine spatial and textural details. Extending this paradigm, Cycle-consistent GANs (CycleGANs) [20] enable unsupervised I2I translation by removing the need for paired data, using cycle consistency to maintain semantic coherence across domains. A comprehensive overview of these and related approaches is provided in [21], which discusses key methodologies, applications, challenges, and future research directions.

Within this landscape, Pix2PixHD [22] has emerged as a variant of cGANs capable of generating images at megapixel resolution through a coarse-to-fine generator and multi-scale discriminators, which help preserve both global structure and local spectral consistency. It has achieved excellent results in translating visible images to NIR [23,24] and in agricultural applications such as plant phenotyping [25,26] and disease detection [27]. However, to date, no study has applied Pix2PixHD to synthesize NIR imagery of pineapple crops, which exhibit complex canopies and unique spectral responses that challenge standard RGB-to-NIR translation methods.

GAN-based models such as Pix2Pix [10], CycleGAN [20], and Pix2PixHD [22] have demonstrated strong capabilities in image translation and have had a profound impact across multiple domains, including precision agriculture [1,4,6,9,28,29], robotics [30], med-

ical imaging [31], human face synthesis [32], and fire detection [33]. These models rely on an adversarial process in which an encoder–decoder generator progressively learns to produce realistic images in the target domain while a discriminator evaluates their authenticity. More advanced methods, such as Pix2Next [34], have been developed to enhance image quality by integrating Vision Foundation Models and cross-attention mechanisms, demonstrating a new state-of-the-art in the field.

In this study, we addressed the challenge of generating synthetic NIR images with a resolution of  $2592 \times 1944$  pixels of pineapple crops from RGB data by employing the Pix2PixHD method, specifically designed for high-quality imagery synthesis. Our approach demonstrates that reliable NIR information can be derived from standard RGB drone imagery, thereby eliminating the need for costly multispectral equipment. By producing spectrally accurate and high-quality synthetic NIR data, this work offers a practical and cost-effective pathway for precision agriculture applications, enabling advanced tasks such as vegetation segmentation and crop health monitoring. These findings highlight the potential of deep learning–based image translation to make sophisticated agricultural analysis more accessible to farmers and practitioners worldwide.

To synthesize the limitations of prior work, existing NIR-synthesis studies face three persistent constraints: (i) the scarcity and difficulty of acquiring paired RGB–NIR datasets, which limits generalization; (ii) the predominance of low-resolution outputs that hinder applications requiring fine spatial detail; and (iii) the lack of evaluations on complex tropical crops, resulting in limited evidence of generalization to real-world agricultural structures. These constraints reveal a clear research gap: no previous work provides paired RGB–NIR synthesis for tropical crops with complex canopies—such as pineapple—using a GAN-based framework capable of handling large-scale imagery.

To address this gap, this study offers three main contributions: (i) the first application of Pix2PixHD to generate synthetic NIR images specifically for pineapple crops, a use case characterized by complex canopies; (ii) the development of a new paired RGB–NIR aerial dataset with a resolution of  $2592 \times 1944$  pixels, which is larger than those typically found in previous research on NIR synthesis; and (iii) a systematic evaluation of 580 training checkpoints to examine the relationship between quantitative performance and perceptual quality. These contributions improve the methodological basis for GAN-based spectral synthesis in the context of pineapple crop monitoring.

Specifically, the objectives of this work are: (i) to train and evaluate a Pix2PixHD model to generate synthetic NIR images from RGB aerial data of pineapple crops; (ii) to assess the quality of the synthetic NIR images by comparing them with the corresponding real test images, using both quantitative metrics (SSIM, PSNR, RMSE, and PPC) and visual assessment to balance numerical performance with artifact-free appearance; and (iii) to analyze the potential of this approach to reduce equipment costs while maintaining spectral accuracy in tropical agriculture.

## 2. Materials and Methods

### 2.1. Dataset Collection

The multispectral dataset was acquired using a DJI Mavic 3M UAV. The equipment manufacturer is DJI, and it was sourced from Montería, Colombia. This is equipped with an integrated sensor system that captures images simultaneously in four spectral bands (green, red, red edge, and near-infrared) and one RGB image composed of three color channels. The DJI Mavic 3M covers green, red, red edge, and NIR wavelengths centering at 560 nm, 650 nm, 730 nm and 860, respectively. For the visual recording of the pineapple crop, a systematic image acquisition was carried out during the 2024–2025 cycle. The captures were programmed to occur at a quarterly frequency to document the crop at different stages

of its growth. All photographs were taken at an altitude of 20 m above the plantation and from an angle perpendicular to the ground surface.

In this study, pairs of RGB and NIR images were used to train and evaluate the Pix2PixHD model. The RGB images, composed of three color channels, served as the input data, while the corresponding NIR band—centered at a wavelength of 860 nm—was used as the target output to generate synthetic near-infrared representations. This pairing enables the model to learn spectral transformations from the visible domain to the NIR domain. The dataset consists of 1521 RGB–NIR paired images, each with a resolution of  $2592 \times 1944$  pixels, with 70% of the samples randomly assigned to training and 30% to testing. The RGB–NIR image pairs collected during 2024 and 2025 were jointly used for training and evaluation to encompass a wider diversity of phenological and illumination conditions and to ensure a more representative dataset size. To complement this primary split and to evaluate the model's temporal robustness, an additional cross-year validation was performed. In this analysis, the model was trained exclusively on the 1014 paired images collected in 2024 and subsequently evaluated on the 507 images acquired in 2025, enabling an independent assessment of inter-annual variation effects. The results of this validation are presented in Section 3.4.

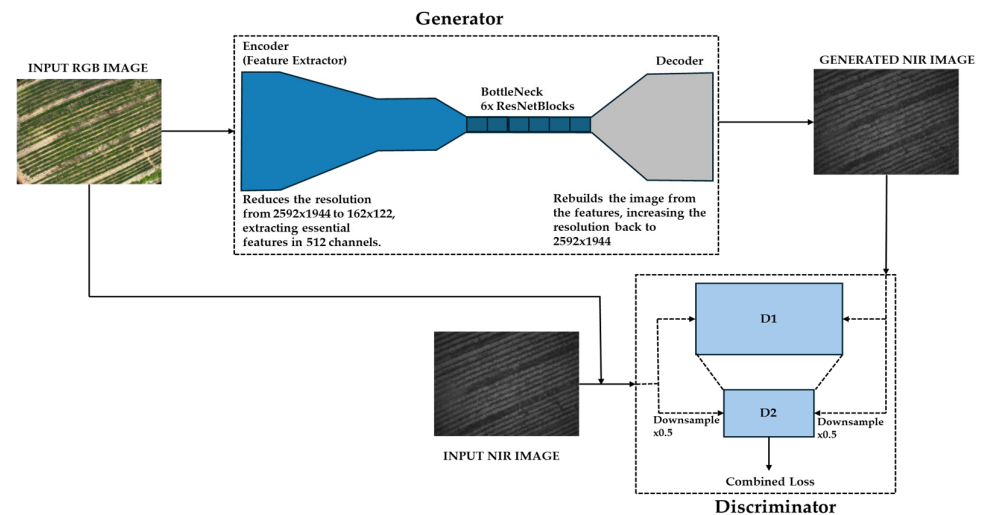
## 2.2. Methods

The Pix2Pix method [10] is a cGAN-based framework used for I2I translation tasks. It consists of a generator and a discriminator: the generator aims to translate images in RGB format to images in the NIR spectrum. In contrast, the discriminator seeks to differentiate between authentic and model-generated images. This approach is trained in a supervised manner, using a dataset composed of pairs of corresponding images ( $x_i, y_i$ ), where  $x_i$  is an RGB image and  $y_i$  its respective NIR image. Pix2pix employs a U-Net architecture [18], and as a discriminator, a fully convolutional network known as PatchGAN [10].

Since the images used in this study have a resolution of  $2592 \times 1944$  pixels, we chose to employ Pix2pixHD, an extension of the original framework proposed by [21], explicitly designed for high-quality images synthesis at large resolutions. This model introduces a coarse-to-fine generator, a set of multi-scale discriminators, and a more robust adversarial loss function. The generator is decomposed into two sub-networks: G1, which acts as a global generator, and G2, which refines the results locally, forming a generator (G1, G2). The discriminator, on the other hand, adapts to challenges associated with large-resolution images by using multiple discriminators in parallel with the same architecture, but operating at different spatial scales, allowing a more efficient and stable evaluation without overfitting or excessive memory usage.

## 2.3. Network Architecture

The architecture used in this work corresponds to an adaptation of the Pix2pixHD algorithm, in which only the G1 subnetwork of the generator is used. This decision addresses computational limitations that prevent the implementation of the local enhancement provided by the G2 subnetwork. Also, the discriminator was configured with two multi-scale PatchGAN networks operating in parallel. However, these modifications were made without compromising training efficiency or stability. The architecture was configured to process images at a resolution of  $2592 \times 1944$  pixels efficiently in terms of memory usage, while ensuring reliability in the generated images. Algorithm A1 (Appendix A) outlines the training procedure, which maintains the conventional scheme of cGANs, consisting of alternating updates of the generator and discriminator parameters (Figure 1).



**Figure 1.** Architecture of the Pix2pixHD implemented for RGB to NIR image translation.

### 2.3.1. Feature Extractor

The feature extractor comprises a series of convolutional operations that perform a transformation on the pixels of the input image into a semantic representation. The process starts with an RGB image of dimensions  $2592 \times 1944$  and three channels, which is processed by an initial convolutional layer that converts these channels into 32 feature maps. From there, the information goes through four sequential convolutional blocks, where each one reduces the spatial resolution by half and doubles the number of channels, as shown in Table 1. In the final stage of such extraction development, the output point is attributed as a compact feature map, which has a limited resolution ( $122 \times 162$ ) and a massive depth (512 channels). So far, this tensor is a compact encoding of the original visualization, as it integrates the semantic information implemented by the ResnetBlocks in the bottleneck.

**Table 1.** Encoder Downsampling Blocks.

| Step | Input Dimension    | Input Channels | Output Dimension  | Output Channels |
|------|--------------------|----------------|-------------------|-----------------|
| 1    | $2592 \times 1944$ | 32             | $1296 \times 972$ | 64              |
| 2    | $1296 \times 972$  | 64             | $648 \times 486$  | 128             |
| 3    | $648 \times 486$   | 128            | $324 \times 243$  | 256             |
| 4    | $324 \times 243$   | 256            | $162 \times 122$  | 512             |

### 2.3.2. Generator

The first half of the generator network corresponds to the feature extractor, whose output enters the processing stage through the bottleneck; this is the central and deepest part of the network. Here, the compact feature map passes through a series of six ResnetBlocks, where each applies complex transformations to the data without changing its resolution. The key to these blocks lies in their skip connections, which allow the network to go deeper and learn complex patterns without losing information from previous layers. This is the stage at which the network understands the image content and defines how it will be translated. After finishing the processing at the bottleneck, the decoder starts to fulfill its function of performing the inverse task, i.e., reconstructing the final image, this is performed through four upsampling blocks that work as an inverse convolution where each block doubles the spatial resolution and, at the same time, halves the number of channels, inversely replicating the encoder process, as shown in Table 2. Finally, a last convolutional layer converts the feature maps into the channels of the output NIR image, and a Tanh activation function normalizes the pixel values, thus completing the generated image at its original resolution.

**Table 2.** Decoder Upsampling Blocks.

| Block | Input Dimension   | Input Channels | Output Dimension   | Output Channels |
|-------|-------------------|----------------|--------------------|-----------------|
| 1     | $162 \times 122$  | 512            | $324 \times 243$   | 256             |
| 2     | $324 \times 243$  | 256            | $648 \times 486$   | 128             |
| 3     | $648 \times 486$  | 128            | $1296 \times 972$  | 64              |
| 4     | $1296 \times 972$ | 64             | $2592 \times 1944$ | 32              |

### 2.3.3. Discriminator

Discriminator is the second essential component of the architecture used in this work. Its function is to act as an “art critic,” judging how realistic the NIR image generated by the model is by comparing it with a real NIR image. To achieve this, it does not focus on the NIR image in isolation. Still, it analyzes the complete pair, i.e., the RGB image input together with the NIR image (whether real or generated). This component is based on two key concepts: multi-scale and PatchGAN. The multi-scale approach uses several discriminators. In this work, two discriminators were used that work in parallel at different scales. The first discriminator operates at full resolution with the image and is responsible for evaluating fine details, edge sharpness, and texture consistency. Conversely, the second discriminator operates on a version of the image reduced by half, specializing in overall coherence, general composition, and whether objects have the correct structure on a large scale. This dual approach forces the generator to produce images that are not only detailed but also structurally correct. In the PatchGAN approach, the discriminator is not limited to giving a single verdict on whether the image is real or fake. Instead, it functions as a fully convolutional network that generates a feature map, where each pixel represents the evaluation of a small overlapping patch of the input image. That value represents the discriminator’s decision on the authenticity of that fragment. In this way, instead of judging the image as a whole, the discriminator analyzes it in parts, forcing the generator to attend and improve the realism in all areas of the image simultaneously.

### 2.3.4. Output Configuration and NIR Channel Representation

In this work, the generator’s output was configured with three channels, following the original Pix2PixHD architecture. Although the target domain is near-infrared (NIR) images, maintaining the three-channel configuration ensures compatibility with the model’s pretrained layers and facilitates stable convergence during adversarial training. However, each output channel does not represent a distinct wavelength or a separate NIR band. Instead, the three channels jointly encode the learned spatial and textural representation of the synthetic NIR domain, effectively reproducing the visual and radiometric characteristics of real NIR images acquired at a wavelength centered at 860 nm.

Therefore, only one NIR band—corresponding to the near-infrared spectral range of the original DJI Mavic 3M sensor—was modeled, and a continuous NIR spectrum was not generated. As a result, spectral profiles are not reported, since the model’s output is a synthetic image-level reconstruction rather than a pixel-level spectral estimate. This design choice allows us to leverage Pix2PixHD’s high-capacity feature mapping while preserving structural and radiometric consistency with the real NIR domain.

## 2.4. Loss Function

Model performance was optimized by incorporating additional components into the standard loss function used in GANs, thereby improving the quality of the generated images. We used L1-Loss, VGG Loss, and a Feature Matching Loss (see Appendix B).

### 2.5. Training Checkpoints and Model Notation

During training, the model was optimized for a total of 580 epochs. To monitor the evolution of the learning process, the first checkpoint was saved after epoch 1 and subsequent checkpoints were stored every 10 epochs. Throughout the manuscript, the notation model followed by a number (e.g., model 410, model 460, model 500) refers to the complete state of the Pix2PixHD network—both generator and discriminator—saved at the corresponding epoch. This convention enables a progressive evaluation of the model's behavior during training and facilitates the identification of convergence trends, stability patterns, and the selection of the most reliable configuration for subsequent analysis.

## 3. Results

### 3.1. Quantitative Performance Analysis

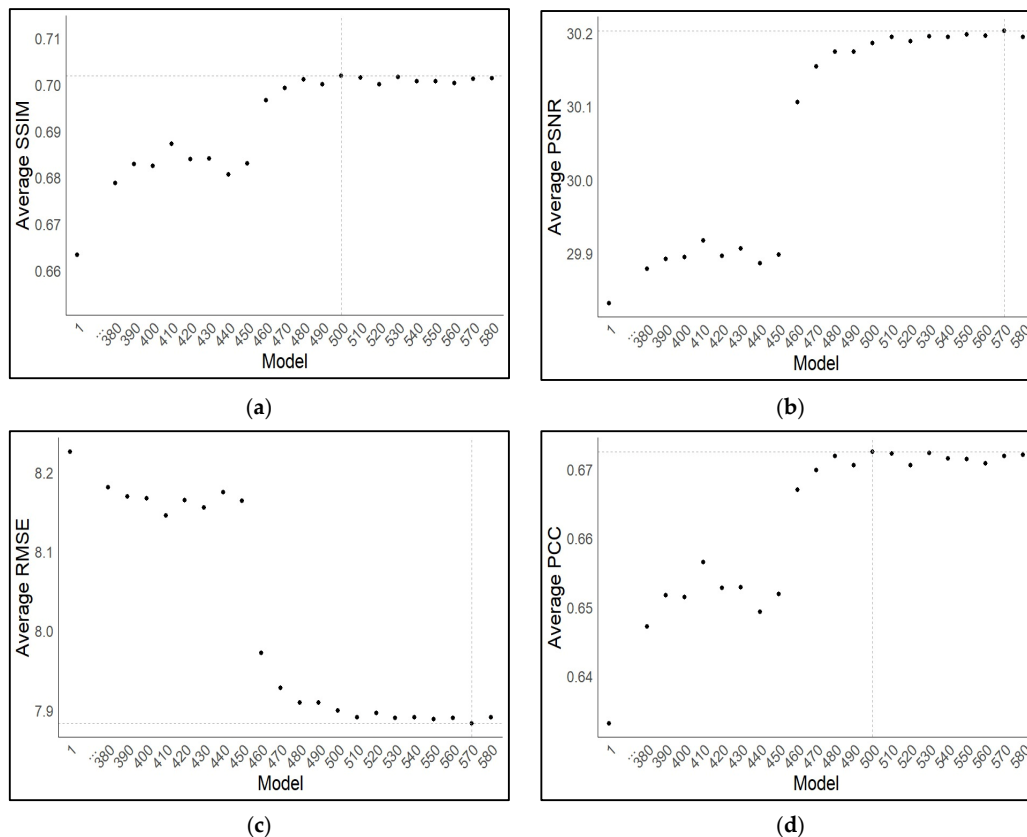
To assess the performance of the proposed models in estimating NIR images from RGB inputs, we report four key evaluation metrics that collectively capture structural similarity, signal quality, reconstruction error, and correlation with ground truth data. Specifically, we employ SSIM, PSNR, RMSE, and PCC metrics [24,35]. These metrics were computed as averages over an independent test dataset composed of 229 RGB–NIR image pairs that were not used during training, ensuring an unbiased validation of model performance.

The progression of SSIM values across the different models demonstrates a consistent improvement in structural similarity between the predicted and ground truth NIR images. As illustrated in Figure 2a, SSIM starts at 0.6633 for the initial model and gradually increases with training, surpassing 0.70 around model 480. From model 500 onwards, SSIM values remain stable at approximately 0.701, with the highest value (0.7019) recorded for model 500. This plateau suggests that the model has reached convergence in terms of structural fidelity. The consistent performance observed beyond model 500 indicates that additional training offers limited gains in perceptual similarity, making this stage a suitable point for model selection based on SSIM.

The average PSNR values exhibit a clear upward trend as training progresses, reflecting an overall improvement in image quality and a reduction in reconstruction noise. As illustrated in Figure 2b, the initial model reaches a PSNR of 29.832 dB, which steadily increases across subsequent models. A noticeable jump occurs around model 460, where PSNR surpasses 30 dB, and continues to rise until it peaks at 30.203 dB in model 570. These results suggest that model 570 offers the best compromise in terms of signal-to-noise ratio, with minimal gains observed in later models.

The RMSE values exhibit a steady decline throughout the training process, indicating a continuous reduction in the prediction error of the estimated NIR images. As shown in Figure 2c, the initial model has the highest RMSE at 8.2264, followed by a gradual decrease across subsequent models. A marked drop occurs around model 460, where RMSE falls below 8.0, and continues to improve until it reaches its lowest point at model 570 with an RMSE of 7.8832. Beyond this point, RMSE values remain nearly constant, suggesting that the model has reached error convergence.

The PCC results, shown in Figure 2d, demonstrate a steady improvement throughout the training process. Starting at 0.6331 for model 1, the values gradually increase, reaching 0.667 by model 460. From that point onward, a sharper rise is observed, followed by a stabilization above 0.67. The highest value is observed at model 500 (0.6726), closely followed by models 530 (0.6724) and 580 (0.6721). This plateau suggests that the model has reached its peak capacity for capturing linear correlations between the estimated NIR images and the ground truth. Beyond this stage, additional training does not yield significant gains. These results highlight a consistent enhancement in correlation performance, with models around iteration 500 showing the most reliable outcomes in terms of PCC.



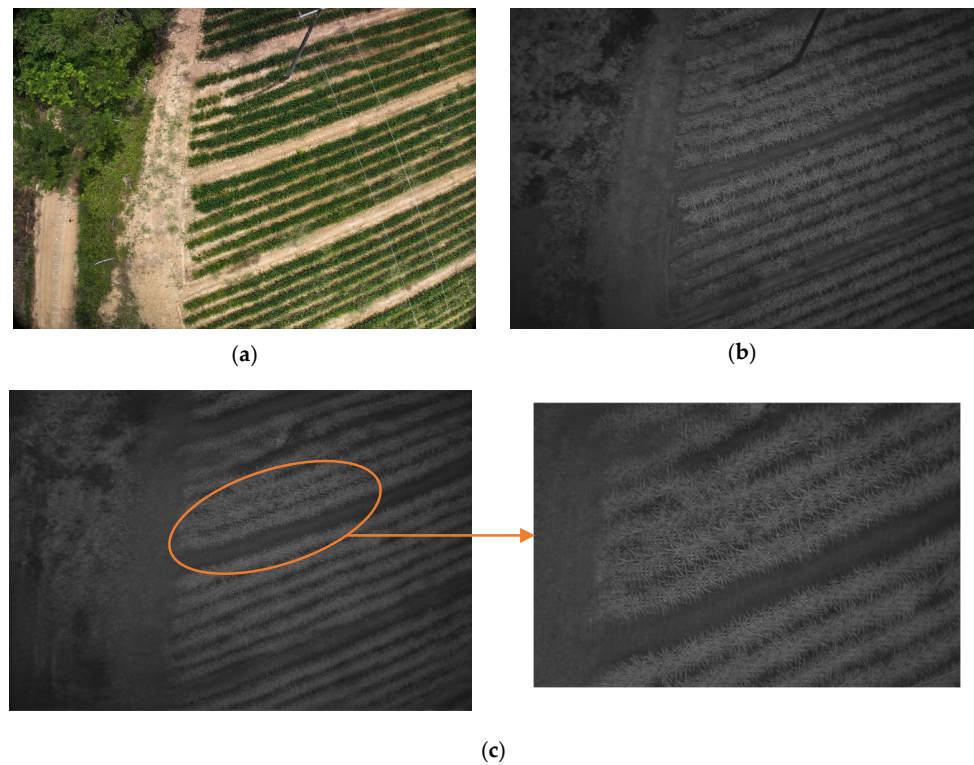
**Figure 2.** Quantitative Evaluation Metrics (average): (a) SSIM, (b) PSNR, (c) RMSE, (d) PCC. The dot and dashed lines represent the model with the best performance across the evaluation metrics and models indexed from 2 to 370 are not shown in detail in the figure, as they were excluded due to low performance.

### 3.2. Qualitative Analysis

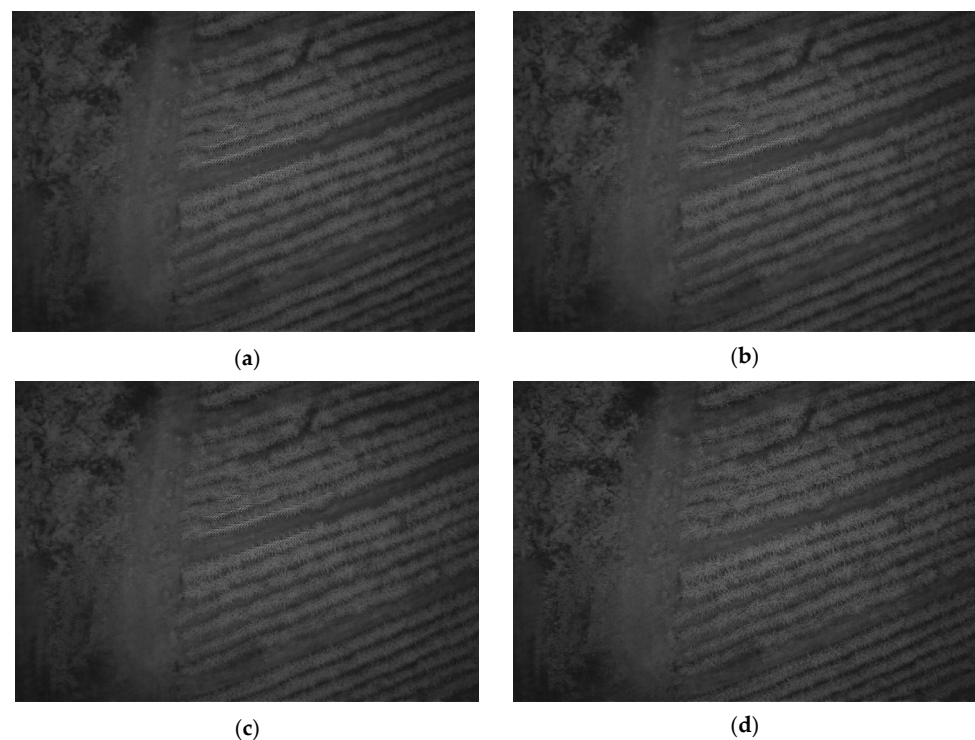
To evaluate the performance of the models in estimating NIR images from RGB inputs, a qualitative analysis is presented. Figure 3a,b display a visual reference consisting of an RGB image and its corresponding NIR ground truth, both selected from the test dataset. The NIR images correspond to a spectral band centered around 860 nm, consistent with the near-infrared range captured by the sensor used in this study.

Model 1 serves as the initial baseline in our training pipeline for estimating NIR images from RGB inputs. Quantitatively, this model exhibits suboptimal performance metrics, with an SSIM of 0.6663, PSNR of 29.832, RMSE of 8.2264, and a PCC of 0.6331. These values reflect significant discrepancies between the predicted and ground truth NIR images. Qualitatively, the NIR image estimated by Model 1, shown in Figure 3c, contains noticeable artifacts [36] that affect the overall image quality.

The NIR image estimated by Model 500, presented in Figure 4a, exhibits some noticeable artifacts despite its improved quantitative performance. These artifacts indicate that, while the model better captures the overall spectral characteristics and structural details compared to other versions, it has not fully eliminated the visual distortions in the predicted images.



**Figure 3.** (a) RGB real test image; (b) NIR real test image; (c) NIR image estimate using model 1.



**Figure 4.** NIR image estimated using: (a) model 500; (b) model 570; (c) model 460; (d) model 410.

The NIR image estimated by Model 570, shown in Figure 4b, demonstrates good quantitative performance, particularly with a high PSNR of 30.203. However, despite these promising metrics, the image still contains noticeable artifacts. This suggests that although Model 570 improves signal quality, it does not entirely overcome the visual distortions present in the estimated NIR images.

Starting from model 460, a noticeable positive shift in quantitative performance is observed across all metrics, as detailed in the previous analysis. This improvement continues through models 500 and 570, with the latter achieving peak PSNR and minimal RMSE. Despite these advances, the NIR image estimated by Model 460, shown in Figure 4c, still exhibits artifacts, although these are less perceptible compared to those produced by models 500 and 570.

The NIR image estimated by Model 410, shown in Figure 4d, does not exhibit noticeable artifacts, distinguishing it from models 460, 500, and 570. Quantitatively, Model 410 achieves an SSIM of 0.6873, PSNR of 29.917, RMSE of 8.1458, and PCC of 0.6565. While these metrics reflect a positive trend compared to earlier models, the differences with models 460, 500, and 570 are relatively small, indicating incremental improvements in structural similarity, signal quality, and reconstruction accuracy. This suggests that although Model 410 performs well without visible artifacts, subsequent models provide modest enhancements in quantitative metrics at the cost of introducing some visual distortions.

### 3.3. Analysis of Model 410 Performance

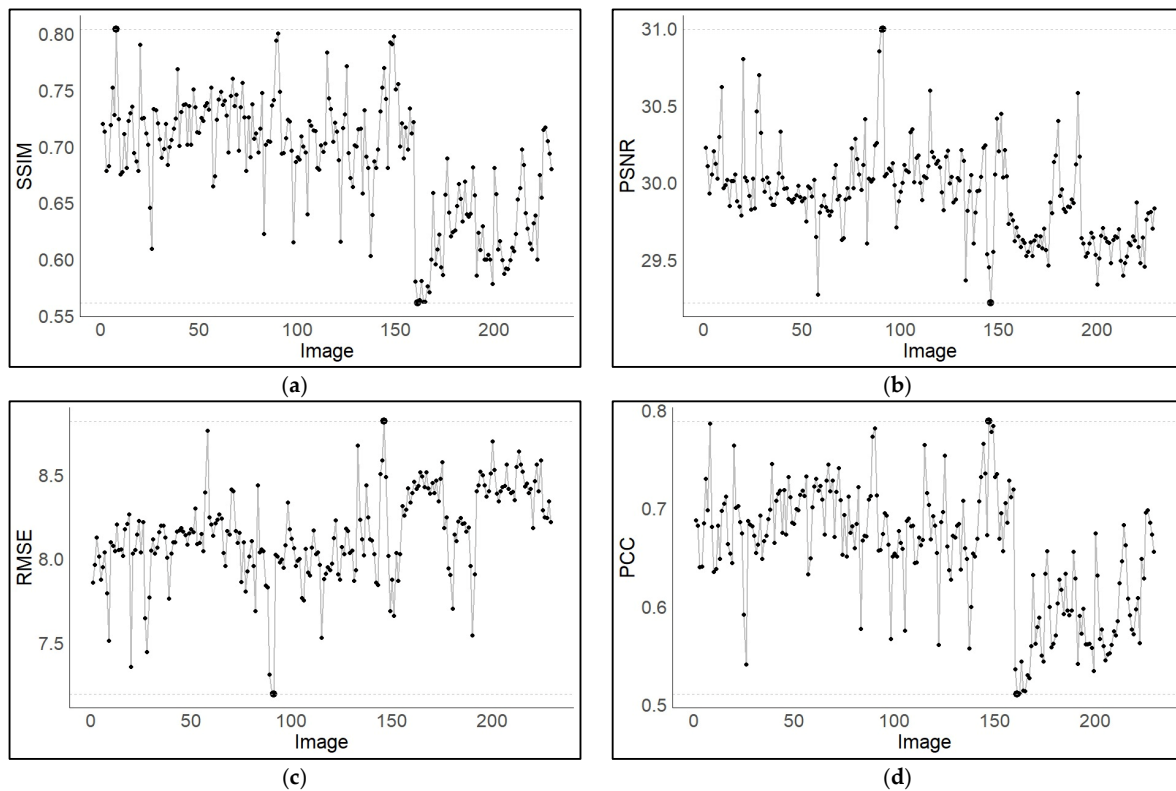
Table 3 presents the descriptive statistics for Model 410 across all evaluation metrics. The quantitative evaluation of Model 410 across 229 images demonstrates consistent performance in terms of structural similarity, peak signal-to-noise ratio, and error metrics. The SSIM values show a mean of 0.6873 (SD = 0.0543), indicating moderate to high structural similarity between the estimated and ground truth NIR images. The interquartile range (Q1 = 0.6530, Q3 = 0.7248) suggests that most images achieve SSIM values close to 0.70, with a maximum of 0.8037 reflecting strong structural fidelity in the best cases.

**Table 3.** Descriptive Statistics for Model 410 (n = 229 images).

| Metric | Mean   | Std. Dev. | Min    | Q1     | Median | Q3     | Max    |
|--------|--------|-----------|--------|--------|--------|--------|--------|
| SSIM   | 0.6873 | 0.0543    | 0.5615 | 0.6530 | 0.6975 | 0.7248 | 0.8037 |
| PSNR   | 29.92  | 0.2907    | 29.22  | 29.70  | 29.91  | 30.05  | 30.99  |
| RMSE   | 8.146  | 0.2698    | 7.193  | 8.016  | 8.145  | 8.349  | 8.820  |
| PCC    | 0.6565 | 0.0625    | 0.5107 | 0.6172 | 0.6689 | 0.6989 | 0.7886 |

Additionally, the trend observed in Figure 5a illustrates the variation in SSIM values across the 229 evaluated images. While the overall distribution remains concentrated around 0.70, the plot reveals localized fluctuations, with certain sequences presenting higher structural similarity peaks (close to 0.80) and others exhibiting noticeable drops near 0.55. These variations suggest that although the model generally maintains structural fidelity, its performance is sensitive to specific image characteristics or regions within the dataset. Notably, the consistency in mid-range values aligns with the interquartile range reported in Table 3, confirming that most images retain a moderate to high degree of structural similarity despite these isolated deviations.

As shown in Table 3, the PSNR values for Model 410 display a mean of 29.92 dB (SD = 0.2907), with a range from 29.22 to 30.99 dB, indicating consistent reconstruction quality across the dataset. Figure 5b illustrates the PSNR evolution for all 229 images, highlighting that most images cluster closely around the mean, while occasional peaks and dips represent slightly higher or lower reconstruction fidelity. Overall, the plot confirms the model's stability, demonstrating that the predicted images maintain low distortion relative to the original NIR references across the entire set.



**Figure 5.** Performance metrics for Model 410 across 229 evaluated images: (a) SSIM, (b) PSNR, (c) RMSE, and (d) PCC.

The trend observed in Figure 5c illustrates the variation in RMSE values across the 229 evaluated images. While the overall distribution is concentrated around the mean of 8.146, the plot reveals moderate fluctuations in error values across the images. The RMSE values remain relatively consistent, with a minimum of 7.193 and a maximum of 8.820, suggesting that the model's performance is fairly stable throughout the dataset. These variations, while present, are not extreme, indicating that the reconstruction errors are generally uniform across most images.

The limited range of RMSE values suggests that the model exhibits consistent performance across different image types, with only slight variations in error magnitude. The plot further reveals that while certain images are reconstructed with slightly lower RMSE, others show a modest increase in error, reflecting the model's sensitivity to specific image characteristics. The overall stability of the RMSE values aligns with the interquartile range reported in Table 3, confirming that most images retain relatively similar reconstruction accuracy, with only minor deviations from the mean. The model appears to maintain a reliable level of performance across the dataset, with variations largely falling within an acceptable range.

Finally, Table 3 shows that the PCC averages 0.6565 (SD = 0.0625), demonstrating a moderate positive correlation between predicted and actual NIR intensities. The minimum value of 0.5107 indicates that there are some cases with relatively weaker correlations, suggesting that certain images pose more challenges for the model. However, the upper quartile (Q3 = 0.6989) and maximum (0.7886) highlight that a substantial portion of the dataset exhibits a strong positive correlation between the predicted and actual values.

This trend is further corroborated by Figure 5d, which visually represents the fluctuations in PCC across the 229 images. The plot shows that while there are some instances with lower PCC values around 0.55, the majority of images exhibit PCC values closer to the upper quartile, with several reaching values above 0.75. These fluctuations are relatively

modest, indicating a generally stable performance across the dataset, where most images maintain a moderate to strong correlation between the predicted and actual intensities.

### 3.4. Cross-Year Validation

To assess the temporal robustness of the model and its ability to generalize across acquisition periods, a cross-year validation was conducted. In this analysis, the model was trained on the 1014 paired images collected in 2024 and evaluated on the 507 images acquired in 2025, allowing the effect of inter-annual variation to be independently examined.

The summary statistics obtained from the cross-year validation are presented in Table 4. Overall, the model exhibited moderate performance when evaluated on the 507 images acquired in 2025, with an average SSIM of 0.5637 and a PSNR of 23.621 dB. The RMSE showed a mean value of 16.938, while the PCC averaged 0.5258, indicating limited but consistent pixel-level agreement with the ground-truth NIR images. The distributional metrics (Q1, median, and Q3) further reveal stable central tendencies across all measures, although a wider spread was observed in RMSE and PSNR, reflecting increased variation in reconstruction accuracy under the 2025 acquisition conditions.

**Table 4.** Descriptive Statistics Cross-Year Validation.

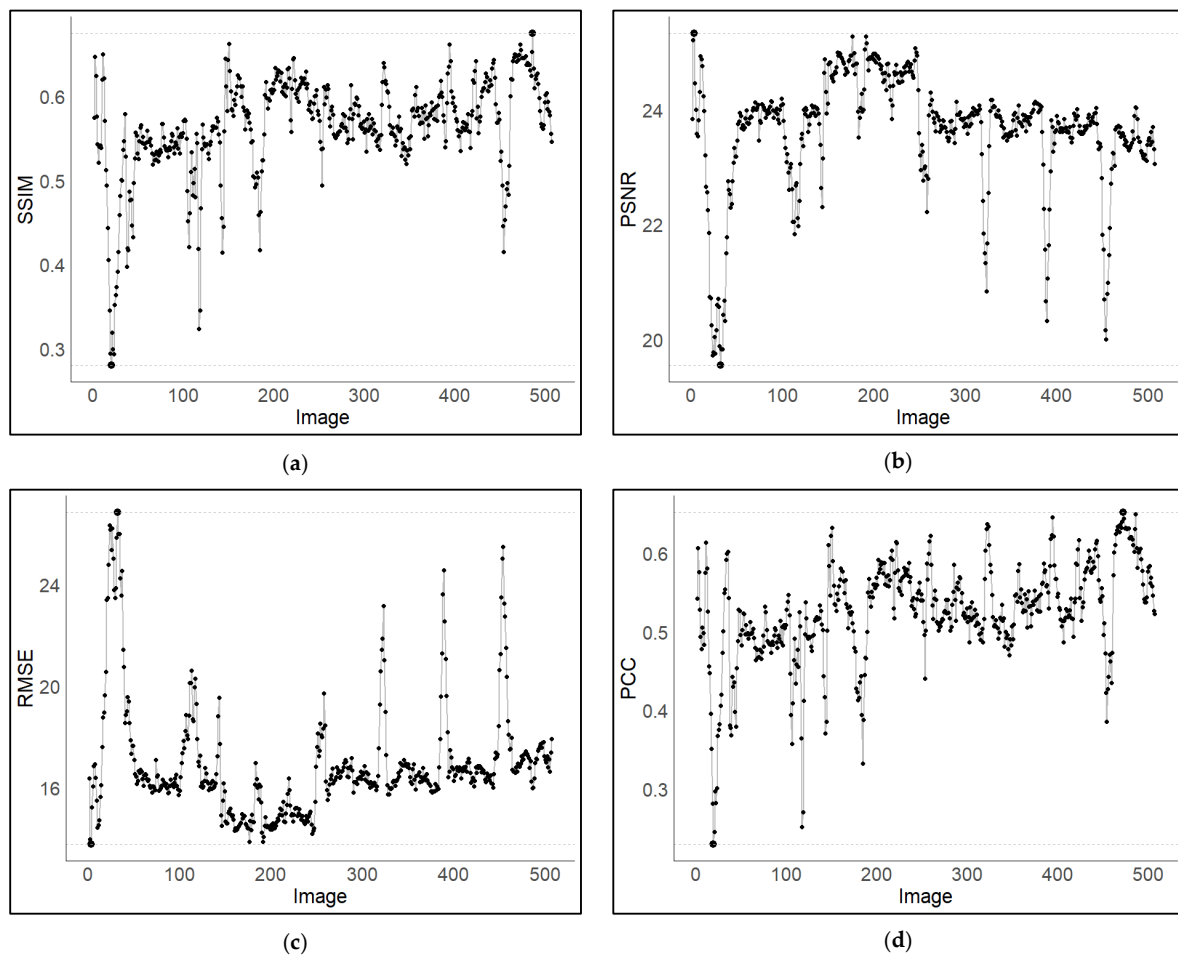
| Metric | Mean   | Std. Dev. | Min    | Q1     | Median | Q3     | Max    |
|--------|--------|-----------|--------|--------|--------|--------|--------|
| SSIM   | 0.5637 | 0.0603    | 0.2806 | 0.5430 | 0.5691 | 0.6034 | 0.6751 |
| PSNR   | 23.621 | 1.0471    | 19.559 | 23.447 | 23.811 | 24.076 | 25.328 |
| RMSE   | 16.938 | 2.2798    | 13.807 | 15.948 | 16.442 | 17.151 | 26.825 |
| PCC    | 0.5258 | 0.0655    | 0.2300 | 0.4971 | 0.5283 | 0.5676 | 0.6521 |

These results indicate that the model retains part of its predictive capability when transferred to a different acquisition year, but with noticeable degradation relative to the within-year evaluation. The lower SSIM and PCC values suggest reduced structural and correlational fidelity, while the drop in PSNR and the higher RMSE reflect increased reconstruction error driven by inter-annual variability in scene characteristics. Although performance remains coherent across the dataset—as evidenced by the relatively narrow interquartile ranges—the overall decrease in all four metrics highlights the sensitivity of the RGB-to-NIR mapping to temporal domain shift, emphasizing the importance of incorporating temporal diversity or adaptation strategies in future model extensions.

The per-image evolution of the four evaluation metrics provides additional insight into the temporal behavior of the model under cross-year conditions. As shown in Figure 6a,d, the SSIM and PCC curves display a consistent upward trend during the initial portion of the test sequence, followed by a stabilization phase in which values oscillate around their respective means. Although occasional drops appear, these isolated fluctuations do not disrupt the overall pattern of moderate structural correspondence between the synthetic and reference NIR images. Similarly, the PSNR distribution exhibits localized decreases associated with specific image groups, while the majority of the predictions cluster near the global mean reported in Table 4, confirming that the degradation is not uniform across the dataset but concentrated in subsets of more challenging scenes.

The RMSE profile further reinforces this behavior, with a dominant region of stable error levels interspersed with pronounced peaks. These high-error instances correspond to abrupt deviations in PSNR and coincide with low PCC and SSIM values, suggesting the presence of acquisition conditions or scene characteristics in 2025 that differ substantially from the 2024 training data. Despite these peaks, the central tendency of RMSE remains relatively compact, aligned with the interquartile ranges previously reported. Taken together, the graphical patterns support the statistical findings by illustrating that

performance degradation under cross-year evaluation is driven not by a systematic collapse of the model, but rather by specific image subsets where inter-annual domain shift is more pronounced.



**Figure 6.** Performance metrics for cross-year validation: (a) SSIM, (b) PSNR, (c) RMSE, and (d) PPC.

#### 4. Discussion

Based on the joint analysis of SSIM, PSNR, RMSE, and PCC metrics across all models (Figure 2), a clear convergence trend is observed, with performance improvements stabilizing after model 460. SSIM steadily increases throughout training, with values surpassing 0.70 from model 480 onward and peaking at model 500 (0.7019). Similarly, PSNR improves from 29.83 (model 1) to over 30.20, reaching its maximum at model 570 (30.203). In terms of RMSE, which measures reconstruction error, values progressively decreased from 8.2264 to a minimum of 7.8832 at model 570, indicating a reduction in pixel-level deviation. PCC, reflecting linear correlation with the ground truth, also stabilizes above 0.67 after model 460, reaching a peak of 0.6726 at model 500.

When evaluating these trends collectively, model 500 emerges as the most balanced candidate. It achieves a high SSIM (0.7019), a strong PSNR (30.1857), low RMSE (7.8989), and the highest PCC (0.6726). While a few later models (e.g., 570 or 580) show very similar values, the marginal gains are minimal and do not justify further training in terms of performance improvement. Thus, model 500 is recommended as the most balanced model, providing the best trade-off between structural similarity, signal quality, reconstruction accuracy, and correlation fidelity.

Qualitative analysis highlights a complex relationship between quantitative performance metrics and the perceptual quality of the estimated NIR images. While Model 1

establishes a baseline with relatively poor results, subsequent models such as 460, 500, and 570 exhibit steady improvements in SSIM, PSNR, RMSE, and PCC. However, these gains come at the expense of persistent visual artifacts that compromise the overall fidelity of the reconstructed images. Notably, Model 410 achieves a balanced outcome by delivering artifact-free results, despite only modest improvements in quantitative scores compared to later models. This contrast suggests that higher numerical performance does not necessarily guarantee superior perceptual quality, underscoring the importance of considering both objective metrics and visual assessment when evaluating model performance for NIR image estimation.

The evaluation of Model 410 underscores its stability and reliability across the dataset, with consistent performance observed in SSIM, PSNR, RMSE, and PCC metrics. Although structural similarity values indicate moderate to high fidelity (mean SSIM  $\approx 0.69$ ), localized fluctuations reveal that performance can vary depending on image-specific characteristics. Similarly, PSNR results demonstrate stable reconstruction quality, with minimal variance around 30 dB, while RMSE values remain within a narrow range, suggesting uniform reconstruction accuracy across most images. The moderate positive correlations captured by PCC further highlight that, despite some cases of weaker alignment between predicted and ground-truth intensities, the majority of images exhibit strong spectral consistency. Collectively, these findings suggest that Model 410 achieves a balanced trade-off between structural fidelity and reconstruction stability, offering robust performance without extreme deviations, although sensitivity to certain image features remains a limiting factor.

In the specific case of pineapple, several factors make NIR synthesis particularly difficult compared to other crops. Pineapple plants form a rosette-shaped structure with long, rigid, and highly overlapping leaves that introduce strong geometric occlusions and self-shading effects. These intertwined leaves generate abrupt transitions in reflectance and heterogeneous textural patterns, complicating the modeling of RGB and NIR relationships. In addition, pineapple exhibits pronounced spectral variability in different parts of the plant—young leaves, mature leaves, and senescent tissues often coexist within the same frame—resulting in greater intraclass variation than is typically observed in more homogeneous crops. The crop also grows at different heights throughout the plantation, producing irregular canopy surfaces that further affect light distribution and the resulting NIR signal. These spatial and spectral complexities differ such as rice [37] or wheat [38], whose canopy is typically slightly more homogeneous and has consistent reflectance patterns, which facilitate NIR synthesis. Consequently, NIR synthesis for pineapple inherently demands greater robustness against structural discontinuities, localized shadows, and spectrum-altering leaf orientations, making it a substantially more difficult target for RGB-based spectral translation than many temperate or uniformly fielded crops.

Cross-year evaluation further highlights the sensitivity of RGB-to-NIR translation to temporal domain shift in pineapple crops. When the model trained with 2024 images was tested on the 2025 subset, all metrics showed a consistent decline (SSIM = 0.5637, PSNR = 23.621 dB, RMSE = 16.938, PCC = 0.5258). Overall, the inter-year analysis shows that, although the model maintains partial predictive power, future extensions should incorporate temporal diversification or adaptation strategies to improve robustness under multi-temporal operating conditions.

From an application standpoint, the proposed RGB-to-NIR translation framework could reduce reliance on specialized multispectral sensors in small- and medium-sized farms. Synthetic NIR maps generated from low-cost RGB images can facilitate the calculation of vegetation indices and yield estimation. Consequently, the synthetic generation of NIR from RGB images has the potential to democratize access to spectral data, especially in

developing regions, by turning any low-cost RGB sensor into a useful tool for data-driven agronomic analysis.

## 5. Conclusions

This study demonstrates the feasibility of generating synthetic NIR images at a resolution of  $2592 \times 1944$  pixels of pineapple crops from standard RGB drone imagery using the Pix2PixHD framework. Among the trained models, model 410 was selected as the most suitable candidate due to its ability to deliver artifact-free outputs while maintaining consistent quantitative performance across all evaluation metrics. Although subsequent models (460–580) achieved slightly higher SSIM, PSNR, and RMSE values, these improvements were accompanied by visible artifacts that compromised perceptual quality. The choice of Model 410 therefore highlights the importance of balancing numerical performance with visual fidelity when applying deep learning models to agricultural imaging. Importantly, our results suggest that reliable and spectrally accurate NIR data can be generated without costly multispectral equipment, laying the groundwork for future advances in digital crop monitoring, sustainability, and smart agriculture technologies by reducing operating costs and expanding accessibility to precision agriculture. These synthetic NIR images enable advanced applications such as vegetation segmentation and crop health monitoring, underscoring the potential of deep learning-based image translation to support sustainable and data-driven farming practices. In summary, this study bridges the gap between affordable RGB imaging and high-value NIR information, laying the foundation for the next generation of low-cost, intelligent sensing systems in precision agriculture. Future research should focus on further reducing residual distortions, extending the model's generalization to different crops and environments, incorporating cross-validation by year, and integrating the approach into real-time drone-based monitoring systems to maximize its practical utility. In addition, the cross-year validation performed in this study provides evidence of the model's partial temporal robustness while also exposing its sensitivity to inter-annual variation. The reduction in SSIM, PSNR, PCC and the increase in RMSE when transferring from 2024 to 2025 images underscores the need to incorporate multi-season samples or domain-adaptation mechanisms in future work. These findings highlight that, while the proposed Pix2PixHD-based framework can generate reliable NIR imagery from RGB data within the same acquisition period, improving temporal generalization is a key direction for enhancing operational deployment in real agricultural environments.

**Author Contributions:** Conceptualization, D.D.U. and R.H.; methodology, D.D.U. and C.M.L.; software, D.D.U.; validation, D.D.U., R.H. and D.S.C.; formal analysis, D.D.U., R.H. and D.S.C.; investigation, D.D.U. and R.H.; resources, C.L.M.; data curation, C.M.L.; writing—original draft preparation, R.H.; writing—review and editing, D.D.U., R.H., J.F.C.L.C. and M.A.B.; visualization, C.L.M. and C.M.L.; supervision, D.D.U., R.H., J.F.C.L.C. and M.A.B.; funding acquisition, R.H., J.F.C.L.C. and M.A.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work has been funded by the Universidad Pontificia Bolivariana—Seccional Montería—(278-01/25-G002) and the Ministerio de Ciencias, Tecnología e Innovación—SNCTI—(Convocatoria 934-2023).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data are available upon request. Our code is available at DoriaU22/RGB2NIR-Pineapple-Pix2PixHD.

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

## Appendix A

---

### Algorithm A1. Training Pix2pixHD for RGB to NIR Image Translation

---

#### Step by step

---

Require: Dataset of image pairs (x, y) where x is RGB and y is NIR  
 Require: Generator  $G_1$  with its parameters  $\theta_G$   
 Require: Multi-scale discriminator  $D = \{D_1, D_2\}$  with its parameters  $\theta_D$   
 Require: Opt\_G and Opt\_D optimizers (Adam) and their learning rate  $\eta_G, \eta_D$   
 Require: Loss functions  $L_{GAN}, L_{L1}, L_{VGG}, L_{Feat}$   
 Require: Hyperparameters  $\lambda_{l1}, \lambda_{feat}, \lambda_{vgg}$   
 Require: Number of epochs N, batch size B

1. for epoch = 1 to N do
2.     for each mini-batch (x, y) of size B do
3.         //Discriminator Update Phase: The Generator only produces; it does not learn in this phase.
4.         Generate false image:  $\hat{y} = G(x)$
5.         Freeze Generator: Disable gradient calculation for  $\theta_G$
6.         Calculate loss in real pairs:  $L_{D\_real} = D(x, y)$  // D must classify it as “real”.
7.         Calculate loss in fake pairs:  $L_{D\_fake} = D(x, \hat{y}.detach())$
8.         Total Discriminator Loss:  $L_{D\_total} = (L_{D\_real} + L_{D\_fake}) * 0.5$
9.         //Update weights of D:
10.         Opt\_D.zero\_grad() //Clear previous gradients of D
11.          $L_{D\_total}.backward()$  //Calculate gradients for D
12.         Opt\_D.step() //Apply update a  $\theta_D$
13.         //Generator Update Phase: The Discriminator only judges, he does not learn in this phase.
14.         Freeze Discriminator: Disable gradient calculation for  $\theta_D$
15.         Loss GAN:  $L_{GAN} = D(x, \hat{y})$  // G wants D to classify  $\hat{y}$  as “real”.
16.          $L1 = L_{L1}(\hat{y}, y)$  //Pixel-to-pixel similarity
17.          $LVGG = L_{VGG}(\hat{y}, y)$  //Perceptual similarity
18.          $L_{Feat} = L_{Feat}(D(x, \hat{y}), D(x, y))$  //Similarity of internal characteristics of D
19.         Total Generator Loss:  $L_{G\_total} = L_{GAN} + \lambda_{l1}*L1 + \lambda_{vgg}*LVGG + \lambda_{feat}*L_{Feat}$
20.         //Update weights of G:
21.         Opt\_G.zero\_grad() //Clear previous gradients of G
22.          $L_{G\_total}.backward()$  //Calculate gradients for G
23.         Opt\_G.step() //Apply update a  $\theta_G$
24.     end for
25. end for

---

## Appendix B

### Appendix B.1. GAN Loss

We use LSGAN, a more stable variant of the traditional GAN loss function, which replaces the cross-entropy loss function with the mean squared error. This variant defines two equations: one for the discriminator and one for the generator, as shown below:

$$\mathcal{L}_{GAN}(D) = \frac{1}{2} \mathbb{E}_{(x,y)} \left[ (D(x, y) - 1)^2 \right] + \frac{1}{2} \mathbb{E}_x \left[ (D(x, G(x)))^2 \right] \quad (A1)$$

$$\mathcal{L}_{GAN}(G) = \frac{1}{2} \mathbb{E}_x \left[ (D(x, G(x)) - 1)^2 \right] \quad (A2)$$

where  $\mathbb{E}$  denotes the expected value,  $D(x, y)$  is the output for a real pair, and  $D(x, G(x))$  is the discriminator output for a generated pair.

#### Appendix B.2. L1-Loss

It is a direct reconstruction loss that forces the generator to create images that are structurally similar to the real image, pixel by pixel. This helps reduce blurring and ensures that the overall content of the image generated is correct. The following equation can define this process:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{(x,y)} \left[ \|y - G(x)\|_1 \right] \quad (A3)$$

where  $\|y - G(x)\|_1$ , or L1 norm is the sum of the absolute differences between each pixel of the real image and the generated image.

#### Appendix B.3. VGG Loss

It measures the similarity between images on a perceptual level, using a pre-trained VGG19 network [3], in order to generate images with textures and content that are visually similar to the real image, resulting in a more realistic image to the human eye. The VGG Loss is formulated as follows:

$$\mathcal{L}_{VGG}(G) = \sum_{j=1}^M \frac{1}{N_j} \|\phi_j(y) - \phi_j(G(x))\|_1 \quad (A4)$$

where  $\phi$  is the VGG19 network, only used to extract features,  $\phi_j$  is the feature map of the  $j$ -th layer,  $M$  refers to the number of VGG layers used in the comparison and  $N$  to the number of elements in the feature map of the  $j$ -th layer.

#### Appendix B.4. Feature Matching Loss

$$\mathcal{L}_{Feat}(G, D) = \sum_{k=1}^K \sum_{t=1}^T \frac{1}{N_i} \|D_K^{(i)}(x, y) - D_K^{(i)}(x, G(x))\|_1 \quad (A5)$$

where  $\mathcal{L}_{Feat}(G, D)$  is the final loss function that depends on both the generator ( $G$ ) and the discriminator ( $D$ ). It is computed as the sum over  $K$  discriminators in the multi-scale architecture and over  $T$  intermediate layers of the discriminator from which the feature maps are extracted.  $D_K^{(i)}(x, y)$  Represents the output feature map of the  $i$ -th layer of the  $k$ -th discriminator,  $x$  is the input RGB image,  $y$  is the real NIR image,  $G(x)$  is the generated NIR image,  $\|\cdot\|_1$  denotes the L1 norm to calculate the absolute element-wise distance between the feature maps. Finally,  $1/N_i$  is the normalization factor where  $N_i$  is the total number of elements in the feature map of the  $i$ -th layer.

## References

- Davidson, C.; Jaganathan, V.; Sivakumar, A.N.; Czarnecki, J.M.P.; Chowdhary, G. NDVI/NDRE prediction from standard RGB aerial imagery using deep learning. *Comput. Electron. Agric.* **2022**, *203*, 107396. [\[CrossRef\]](#)
- Aslahishahri, M.; Stanley, K.G.; Duddu, H.; Shirtliffe, S.; Vail, S.; Bett, K.; Pozniak, C.; Stavness, I. From RGB to NIR: Predicting of near infrared reflectance from visible spectrum aerial images of crops. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2021; pp. 1312–1322. [\[CrossRef\]](#)
- Shanmugapriya, P.; Rathika, S.; Ramesh, T.; Janaki, P. Applications of Remote Sensing in Agriculture—A Review. *Int. J. Curr. Microbiol. Appl. Sci.* **2019**, *8*, 2270–2283. [\[CrossRef\]](#)
- de Lima, D.C.; Saqui, D.; Ataky, S.; Jorge, L.A.d.C.; Ferreira, E.J.; Saito, J.H. *Estimating Agriculture NIR Images from Aerial RGB Data*; Lecture Notes in Computer Science; Springer Verlag: Cham, Switzerland, 2019; pp. 562–574. [\[CrossRef\]](#)

5. Sun, T.; Jung, C.; Fu, Q.; Han, Q. Nir to RGB domain translation using asymmetric cycle generative adversarial networks. *IEEE Access* **2019**, *7*, 112459–112469. [\[CrossRef\]](#)
6. de Lima, D.C.; Saqui, D.; Mpinda, S.A.T.; Saito, J.H. Pix2Pix Network to Estimate Agricultural Near Infrared Images from RGB Data. *Can. J. Remote Sens.* **2022**, *48*, 299–315. [\[CrossRef\]](#)
7. Moscovini, L.; Ortenzi, L.; Pallottino, F.; Figorilli, S.; Violino, S.; Pane, C.; Capparella, V.; Vasta, S.; Costa, C. An open-source machine-learning application for predicting pixel-to-pixel NDVI regression from RGB calibrated images. *Comput. Electron. Agric.* **2024**, *216*, 108536. [\[CrossRef\]](#)
8. Wang, J.; Chen, C.; Wang, J.; Yao, Z.; Wang, Y.; Zhao, Y.; Sun, Y.; Wu, F.; Han, D.; Yang, G.; et al. NDVI Estimation Throughout the Whole Growth Period of Multi-Crops Using RGB Images and Deep Learning. *Agronomy* **2025**, *15*, 63. [\[CrossRef\]](#)
9. Gkillas, A.; Kosmopoulos, D.; Berberidis, K. Cost-efficient coupled learning methods for recovering near-infrared information from RGB signals: Application in precision agriculture. *Comput. Electron. Agric.* **2023**, *209*, 107833. [\[CrossRef\]](#)
10. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2017; pp. 5967–5976. [\[CrossRef\]](#)
11. Zhao, X.; Yu, H.; Bian, H. Image to Image Translation Based on Differential Image Pix2Pix Model. *Comput. Mater. Contin.* **2023**, *77*, 181–198. [\[CrossRef\]](#)
12. Shen, Z.; Huang, M.; Shi, J.; Xue, X.; Huang, T.S. Towards instance-level image-to-image translation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; IEEE Computer Society: Los Alamitos, CA, USA, 2019; pp. 3678–3687. [\[CrossRef\]](#)
13. Huang, J.; Liao, J.; Kwong, S. Unsupervised Image-to-Image Translation via Pre-Trained StyleGAN2 Network. *IEEE Trans. Multimed.* **2021**, *24*, 1435–1448. [\[CrossRef\]](#)
14. Fsiian, A.N.; Thomas, J.B.; Hardeberg, J.Y.; Gouton, P. Spectral Reconstruction from RGB Imagery: A Potential Option for Infinite Spectral Data? *Sensors* **2024**, *24*, 3666. [\[CrossRef\]](#)
15. Illarionova, S.; Shadrin, D.; Trekin, A.; Ignatiev, V.; Oseledets, I. Generation of the nir spectral band for satellite images with convolutional neural networks. *Sensors* **2021**, *21*, 5646. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Picon, A.; Bereciartua-Perez, A.; Eguskiza, I.; Romero-Rodriguez, J.; Jimenez-Ruiz, C.J.; Eggers, T.; Klukas, C.; Navarra-Mestre, R. Deep convolutional neural network for damaged vegetation segmentation from RGB images based on virtual NIR-channel estimation. *Artif. Intell. Agric.* **2022**, *6*, 199–210. [\[CrossRef\]](#)
17. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D. Generative Adversarial Nets. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680. Available online: <https://dl.acm.org/doi/10.5555/2969033.2969125> (accessed on 7 September 2025).
18. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:1411.1784. [\[CrossRef\]](#)
19. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 234–241. [\[CrossRef\]](#)
20. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2017; pp. 2242–2251. [\[CrossRef\]](#)
21. Chakraborty, T.; K S, U.R.; Naik, S.M.; Panja, M.; Manvitha, B. Ten years of generative adversarial nets (GANs): A survey of the state-of-the-art. *Mach. Learn. Sci. Technol.* **2024**, *5*, 011001. [\[CrossRef\]](#)
22. Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; IEEE Computer Society: Los Alamitos, CA, USA, 2018; pp. 8798–8807. [\[CrossRef\]](#)
23. Sa, I.; Lim, J.; Ahn, H.; Macdonald, B. deepNIR: Datasets for Generating Synthetic NIR Images and Improved Fruit Detection System Using Deep Learning Techniques. *Sensors* **2022**, *22*, 4721. [\[CrossRef\]](#)
24. Ravaglia, L.; Longo, R.; Wang, K.; Van Hamme, D.; Moeyersoms, J.; Stoffelen, B.; De Schepper, T. RGB-to-Infrared Translation Using Ensemble Learning Applied to Driving Scenarios. *J. Imaging* **2025**, *11*, 206. [\[CrossRef\]](#)
25. Lou, X.; Fu, Z.; Lin, E.; Liu, H.; He, Y.; Huang, H.; Liu, F.; Weng, Y.; Liang, H. Phenotypic measurements of broadleaf tree seedlings based on improved UNet and Pix2PixHD. *Ind. Crops Prod.* **2024**, *222*, 119880. [\[CrossRef\]](#)
26. Thesma, V.; Velni, J.M. Plant Root Phenotyping Using Deep Conditional GANs and Binary Semantic Segmentation. *Sensors* **2023**, *23*, 309. [\[CrossRef\]](#)
27. Farooque, A.A.; Afzaal, H.; Benlamri, R.; Al-Naemi, S.; MacDonald, E.; Abbas, F.; MacLeod, K.; Ali, H. Red-green-blue to normalized difference vegetation index translation: A robust and inexpensive approach for vegetation monitoring using machine vision and generative adversarial networks. *Precis. Agric.* **2023**, *24*, 1097–1115. [\[CrossRef\]](#)

28. Shukla, A.; Upadhyay, A.; Sharma, M.; Chinnusamy, V.; Kumar, S. High-Resolution NIR Prediction from RGB Images: Application to Plant Phenotyping. In Proceedings of the International Conference on Image Processing, ICIP, Bordeaux, France, 16–19 October 2022; IEEE Computer Society: Los Alamitos, CA, USA, 2022; pp. 4058–4062. [\[CrossRef\]](#)
29. Krestenitis, M.; Ioannidis, K.; Vrochidis, S.; Kompatsiaris, I. Visual to near-infrared image translation for precision agriculture operations using GANs and aerial images. *Comput. Electron. Agric.* **2025**, *237*, 110720. [\[CrossRef\]](#)
30. Divyanth, L.; Rathore, D.; Senthilkumar, P.; Patidar, P.; Zhang, X.; Karkee, M.; Machavaram, R.; Soni, P. Estimating depth from RGB images using deep-learning for robotic applications in apple orchards. *Smart Agric. Technol.* **2023**, *6*, 100345. [\[CrossRef\]](#)
31. Haiderbhai, M.; Ledesma, S.; Lee, S.C.; Seibold, M.; Fürnstahl, P.; Navab, N.; Fallavollita, P. pix2xray: Converting RGB images into X-rays using generative adversarial networks. *Int. J. Comput. Assist. Radiol. Surg.* **2020**, *15*, 973–980. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Ulusoy, U.; Yilmaz, K.; Özşahin, G. Generative Adversarial Network for Generating Synthetic Infrared Image from Visible Image. *Gazi Üniversitesi Fen Bilim. Derg. Part C Tasarım Ve Teknol.* **2022**, *10*, 286–299. [\[CrossRef\]](#)
33. Akagic, A.; Buza, E.; Horvat, M.; Akagi, A. Mapping RGB-to-NIR with Pix2Pix Image-to-Image Translation for Fire Detection Applications. In Proceedings of the 34th Central European Conference on Information and Intelligent Systems (CECIIS 2023), Dubrovnik, Croatia, 20–22 September 2023. Available online: <https://www.researchgate.net/publication/374145237> (accessed on 11 September 2025).
34. Jin, Y.; Park, I.; Song, H.; Ju, H.; Nalcakan, Y.; Kim, S. Pix2Next: Leveraging Vision Foundation Models for RGB to NIR Image Translation. *Technologies* **2025**, *13*, 154. [\[CrossRef\]](#)
35. Yuan, X.; Tian, J.; Reinartz, P. Generating artificial near infrared spectral band from rgb image using conditional generative adversarial network. *ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci.* **2020**, *V-3-2020*, 279–285. [\[CrossRef\]](#)
36. Wu, X.; Chao, D.; Yang, Y. High-Resolution Image Translation Model Based on Grayscale Redefinition. *arXiv* **2024**, arXiv:2403.17639. [\[CrossRef\]](#)
37. Li, J.; Cao, Q.; Wang, S.; Li, J.; Zhao, D.; Feng, S.; Cao, Y.; Xu, T. Improved Multi-Stage Rice Above-Ground Biomass Estimation Using Wavelet-Texture-Fused Vegetation Indices from UAV Remote Sensing. *Plants* **2025**, *14*, 2903. [\[CrossRef\]](#)
38. Zhang, D.; Hou, L.; Lv, L.; Qi, H.; Sun, H.; Zhang, X.; Li, S.; Min, J.; Liu, Y.; Tang, Y.; et al. Precision Agriculture: Temporal and Spatial Modeling of Wheat Canopy Spectral Characteristics. *Agriculture* **2025**, *15*, 326. [\[CrossRef\]](#)

**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.