



Full length article

Five-band multispectral image synthesis from UAV RGB imagery using a spectral-informed GAN for agroforestry monitoring

Dalila Anwar Dini Rasyidah^a, Riyanarto Sarno^{a,*}, Kelly Rossa Sungkono^a, Agus Tri Haryono^a, Rizqy Ahsana Putri^a, Sholiq^b, Dhimas Wiratmoko^c, Siti Halimah Larekeng^d, Arif Abdullah Sagan^e

^a Department of Informatics, Institut Teknologi Sepuluh Nopember (ITS), Kampus ITS Sukolilo, Surabaya, 60111, East Java, Indonesia

^b Department of Information Systems, Institut Teknologi Sepuluh Nopember (ITS), Kampus ITS Sukolilo, Surabaya, 60111, East Java, Indonesia

^c Indonesian Oil Palm Research Institute (IOPRI), Medan, Indonesia

^d Faculty of Forestry, Universitas Hasanuddin, Makassar, Indonesia

^e Faculdade de Ciências Políticas, Universidade de Dili, Dili, Timor-Leste

ARTICLE INFO

Keywords:

Cocoa agroforestry
Multispectral synthesis
Smallholder agriculture
Sustainability indicators
UAV remote sensing
Vegetation indices
West Africa

ABSTRACT

Vegetation indices (VI), including the Normalized Difference Vegetation Index (NDVI) and Normalized Difference Red Edge Index (NDRE), are established proxies for chlorophyll content and plant stress in tropical agroforestry systems. Computing these indices requires Red Edge and Near-Infrared (NIR) bands that standard Red-Green-Blue (RGB) cameras cannot capture. Smallholder cocoa farmers in Côte d'Ivoire, who manage around 40% of global cocoa production, operate RGB-equipped Unmanned Aerial Vehicles (UAVs) without multispectral sensors, which restricts reliable vegetation health monitoring. This study proposes SMH-GAN (Serial dual attention, Multi-head decoder, Hybrid spectral-informed Generative Adversarial Network), a modification of Pix2Next, to synthesize five-band multispectral imagery from RGB UAV data to enable VI computation without dedicated sensors. Four modifications address Pix2Next limitations. Serial Dual Attention (SDA) resolves absent band-specific feature separation by applying spectral then spatial attention sequentially. A Spectral-Informed Bottleneck (SIB) enforces vegetation reflectance constraints at the latent level without target band inputs. A Split-Head Decoder reduces gradient conflicts from joint optimization of heterogeneous bands. A Spectral Consistency Loss penalizes inter-band ratio violations to preserve VI reliability during training. SMH-GAN was evaluated on a public UAV dataset of ten cocoa agroforestry plots in Divo, Côte d'Ivoire, against Pix2Pix, TaijiGNN, and Pix2Next. The proposed method achieved mean Structural Similarity Index Measure (SSIM) of 0.876, a 4.8% improvement over Pix2Next. NDVI achieved $r = 0.91$ with ground truth, and six indices exceeded $r = 0.87$. These results indicate that the synthesized bands preserved the spectral relationships required for vegetation index computation on the evaluated dataset.

1. Introduction

Vegetation health monitoring underpins sustainable land management in tropical agroforestry systems, where canopy condition, chlorophyll content, and plant stress dictate productivity and ecosystem services (Okoyere et al., 2024; Salehin et al., 2025; Sishodia et al., 2020; Sahoo et al., 2025). Spectral vegetation indices, including the Normalized Difference Vegetation Index (NDVI), Green Normalized Difference Vegetation Index (GNDVI), and Normalized Difference Red Edge Index (NDRE), quantify biophysical states across diverse forest and agricultural landscapes (Rashid and Kashani, 2025; Yuan et al., 2022; Boiarskii and Hasegawa, 2019). UAVs equipped with multispectral sensors provide high-spatial-resolution data that identifies stress,

disease, and canopy degradation invisible to standard visible-range imagery (Salehin et al., 2025; Yang et al., 2017). The practical utility of these indicators hinges on whether smallholder farming communities in resource-constrained regions can access the sensor technology required to produce them.

Cocoa agroforestry systems in Côte d'Ivoire represent a context where this access gap carries direct economic and environmental consequences. The country accounts for around 40% of global cocoa production, and most output is managed by smallholder farmers (Kouassi et al., 2023). Smallholder operations in this region face persistent barriers to accessing advanced sensing technology, as limited financial and institutional capacity constrains technology adoption (Antwi-Agyei

* Corresponding author.

E-mail address: riyanarto@its.ac.id (R. Sarno).

<https://doi.org/10.1016/j.srs.2026.100493>

Received 28 April 2026; Received in revised form 2 July 2026; Accepted 17 August 2026

Available online 22 August 2026

2666-0172/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

et al., 2021). The region also exhibits high rates of deforestation driven by agricultural expansion (Lammoglia et al., 2024). Vegetation condition monitoring therefore serves crop productivity and canopy cover tracking as complementary environmental sustainability indicators. Early detection of diseases such as *Phytophthora* requires spectral information in the NIR and Red Edge regions, where pathogen-induced stress manifests before visible symptoms appear (Ndoumbe-Nkeng et al., 2004). RGB cameras cannot capture NIR and Red Edge bands, and studies have associated reliance on visible-range data with delayed disease detection and increased crop losses (Zhou et al., 2023). The gap is therefore spectral, as consumer-grade RGB platforms available to smallholder operations lack the band coverage that vegetation health assessment requires.

Farmers in resource-constrained regions lack access to multispectral UAV, and deep learning-based spectral synthesis from RGB imagery has emerged as a technical pathway to bridge this spectral gap (Aslahishahri et al., 2021; Zhang et al., 2022). Deep learning approaches have advanced this task through convolutional networks and Generative Adversarial Networks (GANs), with applications in RGB-to-NIR translation (Aslahishahri et al., 2021; De Souza et al., 2023) and aerial multispectral reconstruction (Zhang et al., 2022; Rahmawati et al., 2022). Remote sensing-based monitoring in regions with constrained sensor access has confirmed that band availability limits the reliability of vegetation indices derived from optical data (Rashid and Kashani, 2025). The viability of synthesis-based approaches therefore depends on whether generated bands preserve the inter-band spectral relationships that vegetation index computation requires.

Spectral synthesis methods have progressed from sparse coding approaches that were limited by an inability to model nonlinear spectral relationships (Ahmed et al., 2025), through convolutional neural networks that improved spectral transformation learning from paired training data (Ronneberger et al., 2015; Zeng et al., 2021; Fajar et al., 2023), to Generative Adversarial Networks (GANs) that demonstrated improved perceptual realism in RGB-to-NIR translation (Aslahishahri et al., 2021; De Souza et al., 2023) and aerial multispectral reconstruction (Zhang et al., 2022; Rahmawati et al., 2022). Vegetation indices combine bands as normalized ratios, so NDVI depends on the NIR–Red contrast and NDRE on the NIR–Red Edge contrast (Boiarskii and Hasegawa, 2019). The adversarial objective optimizes perceptual realism (De Souza et al., 2023) and does not constrain the relative magnitude between bands, so a reconstruction can look plausible yet shift the ratio. RGB-to-spectral reconstruction is an ill-posed problem, and a reconstruction can match the visible range yet show discrepancies in the Near-Infrared region from 700 to 1000 nm (Ahmed et al., 2025), so the vegetation index inherits the ratio error. A prior RGB-to-multispectral reconstruction study on UAV imagery quantified this band-wise gap, with the non-visible Red Edge band reaching reconstruction error near twice that of the visible bands and a derived NDVI agreement of 69.04% by an intersection-over-union measure (Zhang et al., 2022). Pix2Next advanced this area through pre-trained Vision Foundation Model (VFM) representations and cross-attention mechanisms and achieved competitive RGB-to-NIR results on the RANUS benchmark (Jin et al., 2025). Evaluated against Pix2Pix, Pix2PixHD, CycleGAN, and BBDM, Pix2Next recorded SSIM of 0.803 and FID of 28.01, outperforming all competing methods across six evaluation metrics (Jin et al., 2025). This superiority over existing image-to-image translation methods establishes Pix2Next as the most suitable architectural foundation for spectral synthesis tasks, yet its design for single-band NIR translation limits its direct application to five-band agricultural synthesis. Five-band multispectral synthesis from UAV RGB imagery with downstream vegetation index validation remains under-explored in the existing literature (Zhang et al., 2022; Rahmawati et al., 2022; Aslahishahri et al., 2021; Ahmed et al., 2025).

Three aspects of Pix2Next constrain its application to multi-band agricultural synthesis. The uniform cross-attention design treats all spectral bands without separating spectral from spatial processing,

which is suboptimal when output bands exhibit different reconstruction characteristics; sequential channel-spatial attention has demonstrated consistent gains in feature selectivity across image processing tasks (Jin et al., 2022; Farokhah et al., 2023; Woo et al., 2018), and learned spectral cues have improved band relationship preservation in remote sensing fusion contexts (Xing et al., 2022; Guo et al., 2022; Haryono et al., 2023). The bottleneck layer lacks constraints based on characteristic vegetation reflectance patterns governed by pigment absorption and leaf cellular structure (Gitelson et al., 2003); domain-informed constraints at intermediate representations have improved spectral fidelity in hyperspectral fusion (Liu et al., 2022; Xiao et al., 2021; Q. Zhu et al., 2024; Zheng et al., 2021), though such approaches require target spectral reference inputs absent in RGB-to-multispectral synthesis. The unified decoder does not account for divergent optimization requirements across five spectral bands; standard reconstruction losses optimize per-band fidelity without penalizing inter-band correlation violations (Liu et al., 2022; Gao et al., 2022), and vegetation indices sensitive to inter-band ratios (Boiarskii and Hasegawa, 2019; Xing et al., 2022) are not preserved by unified output layers (Zeng et al., 2021; Aslahishahri et al., 2021; Jin et al., 2025) when tasks with conflicting optimization landscapes are learned jointly (Standley et al., 2020). These three limitations together constrain the reliability of synthesized vegetation indicators for agricultural monitoring applications, and none has been addressed in the context of five-band UAV synthesis from RGB imagery.

Serial dual attention, Multi-head decoder, Hybrid spectral-informed GAN (SMH-GAN) is proposed as a modification of Pix2Next (Jin et al., 2025) to address these three limitations for five-band agricultural synthesis. Serial Dual Attention (SDA) replaces uniform cross-attention with sequential spectral-then-spatial processing to separate channel-wise band selection from spatial localization, accommodating the heterogeneous reconstruction characteristics of individual bands (Woo et al., 2018; Jin et al., 2022). Spectral-Informed Bottleneck (SIB) introduces domain-informed constraints at the latent representation level through dual-path processing with adaptive gating, enforcing vegetation reflectance relationships at the bottleneck without requiring target spectral reference inputs (Gitelson et al., 2003). Split-Head Decoder separates visible band reconstruction from non-visible band inference through asymmetric dedicated branches, reducing gradient conflicts that arise when tasks with divergent optimization landscapes share decoder parameters (Standley et al., 2020). Spectral consistency loss constrains the inter-band relationships of NDVI and NDRE to preserve the reliability of the vegetation index that unified reconstruction losses cannot enforce (Liu et al., 2022; Gao et al., 2022).

All four components are contributions to RGB-to-multispectral synthesis, and they differ in how much they build on existing mechanisms. SDA reframes the channel-then-spatial attention of CBAM (Jin et al., 2022; Woo et al., 2018) as spectral-then-spatial band selection and integrates it with VFM cross-attention in every encoder and decoder block, which replaces the uniform cross-attention of Pix2Next. The Split-Head Decoder extends the multi-task gradient-separation principle (Standley et al., 2020) into an asymmetric three-branch design with branch-specific depths for the visible, Red Edge, and NIR band groups. SIB introduces a latent-level reflectance constraint that operates without target spectral reference inputs, unlike the physics-informed fusion methods that require them (Liu et al., 2022; Xiao et al., 2021; Q. Zhu et al., 2024; Zheng et al., 2021). The spectral consistency loss introduces direct supervision on the NDVI and NDRE band ratios that per-band reconstruction losses leave unconstrained (Liu et al., 2022; Gao et al., 2022). The objectives of this study were to: (i) develop and evaluate a synthesis method that generates Blue, Green, Red, Red Edge, and NIR bands from RGB UAV imagery of cocoa agroforestry in Côte d'Ivoire (Lammoglia et al., 2024); (ii) assess the reliability of synthesized bands as environmental monitoring indicators through eight vegetation indices validated against multispectral ground truth

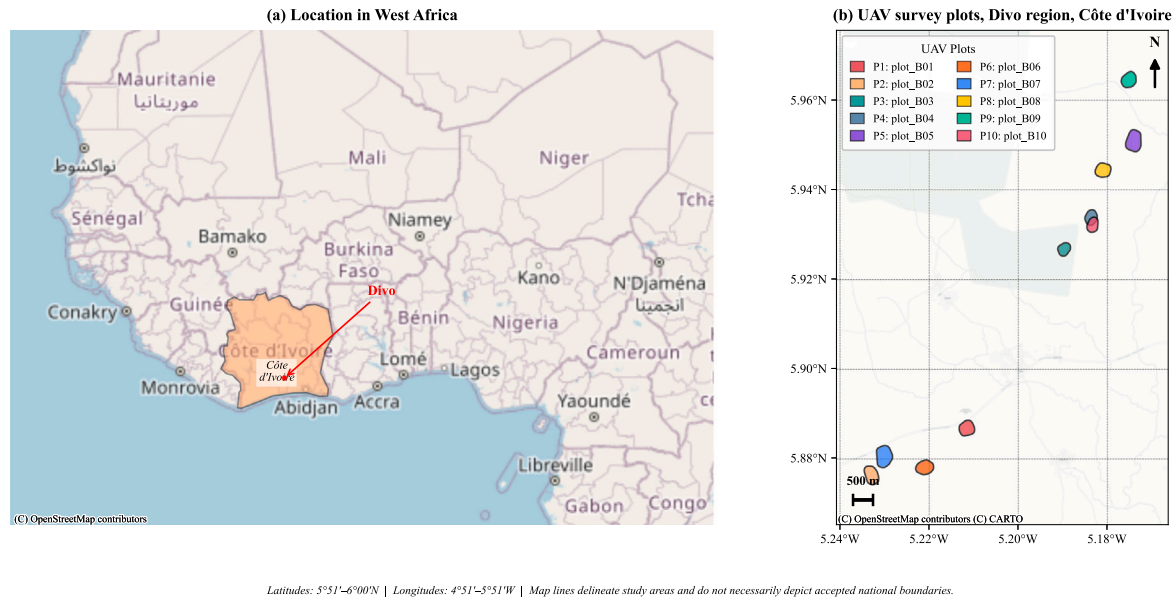


Fig. 1. Study area location. (a) Location of Divo in Côte d'Ivoire, West Africa. (b) Spatial distribution of the ten UAV survey plots (B01–B10) in the Divo region. Latitudes: 5°51'–6°00'N; Longitudes: 4°51'–5°51'W. Map lines delineate study areas and do not necessarily depict accepted national boundaries.

measurements on ten agroforestry plots; and (iii) investigate the contribution of each architectural component to the quality of the synthesis through controlled ablation. The study formulated four hypotheses, one per component. SDA would raise reconstruction quality most on the band with heterogeneous RGB correlations, namely NIR, because spectral-then-spatial attention reweights the informative channels before spatial localization. SIB would improve the bands governed by leaf cellular structure, namely Red Edge and NIR, because the latent reflectance constraint preserves their relationship to the visible bands. The Split-Head Decoder would raise NIR reconstruction because a dedicated branch removes the gradient interference between visible and non-visible bands. The spectral consistency loss would improve the preservation of the NDVI and NDRE band ratios because it adds ratio-level supervision that per-band reconstruction losses do not provide.

2. Study area and data

The geographic context of the study and the UAV dataset used for training and evaluation are described in this section. The characteristics of the dataset, sensor specifications, and preprocessing procedures are presented to establish the experimental basis for the proposed method.

2.1. Study area

The study area is located in the Divo region, Central-West Côte d'Ivoire, West Africa, between latitudes 5°51'–6°00'N and longitudes 4°51'–5°51'W (Fig. 1). The region lies within the cocoa belt of Côte d'Ivoire, a country that accounts for around 40% of global cocoa production and where most output is managed by smallholder farmers (Kouassi et al., 2023). The landscape is highly fragmented, characterized by a mix of forests, cocoa plantations, and food crops alongside high rates of deforestation driven by agricultural expansion (Lammoglia et al., 2024). These conditions make the region relevant for vegetation health monitoring as both an agricultural productivity indicator and a land cover sustainability indicator.

2.2. UAV dataset

The UAV multispectral dataset employed in this study (Lammoglia et al., 2024) is publicly available in the CIRAD Dataverse repository at <https://doi.org/10.18167/DVN1/MK2ZRG>. The dataset comprises aerial imagery from ten cocoa-based agroforestry plots (B01–B10) acquired in May 2022 using a DJI Phantom 4 Multispectral (P4M) drone at 80 m Above Ground Level (AGL). For each acquisition the P4M sensor captured one RGB image together with five co-registered narrow-band multispectral channels, namely Blue (450 nm), Green (560 nm), Red (650 nm), Red Edge (730 nm), and Near-Infrared (NIR; 840 nm). The dataset contains 7638 RGB and multispectral image files at a native resolution of 1600 × 1300 pixels with ground sampling distances of 4.2–4.6 cm. Spatial alignment grouped each RGB image with its five multispectral bands into one RGB-to-multispectral pair and produced 1271 complete pairs across the ten plots. Detailed specifications of the data collection procedure are documented in the original data publication (Lammoglia et al., 2024). The RGB image came from the integrated RGB camera of the P4M and served as the network input. The dataset did not include imagery from a separate RGB device. Each pair consisted of one RGB input image and its five multispectral target bands. The UAV survey acquired images with overlapping coverage, so adjacent images within a plot shared spatial content (Lammoglia et al., 2024). Fig. 2 shows representative sample imagery across all spectral channels.

3. Methodology

Expanding multispectral monitoring access to smallholder agroforestry systems requires synthesis methods that are both technically reliable and practically deployable under resource constraints. This section presents the proposed RGB-to-multispectral synthesis method, covering the preprocessing procedures, network architecture, loss function design, and implementation details.

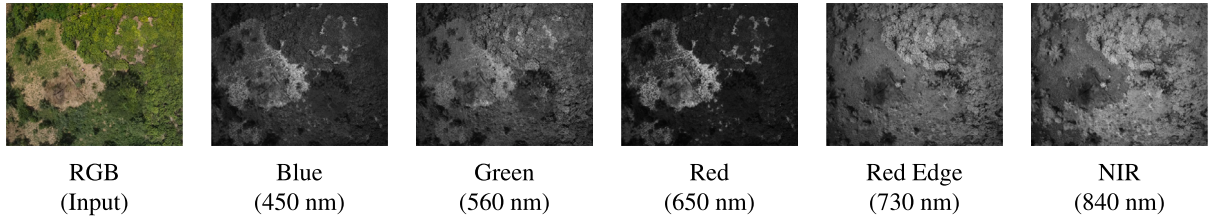


Fig. 2. Representative sample imagery from one cocoa agroforestry plot across all spectral channels. The RGB composite served as network input. The five multispectral bands constitute the synthesis target: Blue (450 nm), Green (560 nm), Red (650 nm), Red Edge (730 nm), and NIR (840 nm).

3.1. Data preprocessing

Three preprocessing steps were applied to prepare the dataset for training: spatial alignment to ensure geometric consistency between RGB and multispectral bands, radiometric calibration to convert raw digital numbers to reflectance values, and normalization with dataset partitioning for model training and evaluation.

3.1.1. Spatial alignment

Spatial alignment was performed with the sensor's onboard calibration parameters and metadata embedded in the image files. The manufacturer's factory calibration ensured sub-pixel alignment accuracy between spectral bands under nominal operating conditions.

3.1.2. Radiometric calibration

Raw Digital Numbers (DN) were converted to reflectance values with a simplified Empirical Line Method (ELM) (Fawcett et al., 2020). A global calibration factor K_b for each spectral band b was derived from the mean DN of the training subset and a per-band reference reflectance value. The dataset provided no calibration panels (Lammoglia et al., 2024), so the reference value was set to a nominal level typical of vegetation canopy. The factor was then applied to all pixel values to convert them to reflectance.

3.1.3. Normalization and partitioning

Both RGB and multispectral images were resized to 256×256 pixels with bilinear interpolation and normalized to the $[-1, 1]$ range. The pipeline resized full images without patch extraction, and no data augmentation was applied during training.

The pairs were partitioned at the plot level so that every pair from a plot belonged to a single subset. This plot-level assignment kept the strong spatial autocorrelation among overlapping UAV images within a plot from crossing the training, validation, and test boundaries, which prevented train-to-test leakage. Six plots (B04 to B09) formed the training set with 905 pairs (905 RGB and 4525 multispectral images), two plots (B01 and B02) formed the validation set with 265 pairs (265 RGB and 1325 multispectral images), and two plots (B03 and B10) formed the test set with 101 pairs (101 RGB and 505 multispectral images). The test plots resided in a separate directory and were excluded from both training and validation, so the test set consisted entirely of previously unseen plots. Table 1 reports the plot-to-subset assignment and the pair count per plot. The plot assignment was fixed across all experiments for reproducibility.

3.2. Proposed architecture

The proposed architecture extended Pix2Next (Jin et al., 2025) for multi-band multispectral synthesis. As illustrated in Fig. 3, the network comprised four main components: a Vision Foundation Model (VFM) feature extractor, an encoder-decoder with Serial Dual Attention (SDA) modules, a Spectral-Informed Bottleneck (SIB), and a Split-Head Decoder. The generator processed the RGB input in four stages. The VFM feature extractor produced multi-scale features, the encoder applied Serial Dual Attention at each stage, the Spectral-Informed Bottleneck refined the latent representation at the encoder to decoder transition, and the Split-Head Decoder produced the visible bands, the Red Edge band, and the NIR band through separate branches.

Table 1

Plot-to-subset assignment and image-pair counts. The partitioning was performed at the plot level, so no plot contributed images to more than one subset and the two test plots remained unseen during training and validation. Each pair consists of one RGB image and five multispectral bands.

Subset	Plot	Pairs
Training	B04	67
	B05	244
	B06	160
	B07	204
	B08	121
	B09	109
Validation	B01	133
	B02	132
Test	B03	49
	B10	52
Total		1271

3.2.1. Vision foundation model feature extractor

SwinV2-Tiny was selected as the pre-trained backbone for the VFM feature extractor. The hierarchical window-based self-attention design of SwinV2-Tiny captures both local texture patterns and long-range spatial context across the full image, which is relevant for synthesis tasks where non-visible band reconstruction depends on distributed RGB cues (Jin et al., 2025). The backbone was initialized with ImageNet pre-trained weights and kept frozen during training to preserve the learned visual representations. Features extracted at multiple encoder stages were projected to match the decoder channel dimensions through 1×1 convolution layers. These features were then integrated into the encoder-decoder via cross-attention in each SDA module and the SIB, as described in Eqs. (4) and (5).

3.2.2. Serial dual attention module

Pix2Next applied uniform cross-attention across all spectral bands without distinguishing between channel-wise and spatial feature relationships. This design was suboptimal for multi-band synthesis. Visible bands required spatial localization of RGB-correspondent regions, while NIR inference required channel-wise reasoning across all three RGB inputs to identify spectral combinations correlated with leaf cellular structure. SDA addressed this limitation by separating spectral and spatial attention into sequential stages, allowing the network to first select informative channel combinations and then localize spatially relevant regions within the selected feature representation (Woo et al., 2018; Jin et al., 2022; Putri et al., 2026).

The Serial Dual Attention (SDA) module applied spectral and spatial attention sequentially within each encoder and decoder block. Let $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$ denote the input feature tensor, where C is the number of channels, H the feature map height, and W the feature map width. Spectral attention $\mathbf{A}_{spec} \in \mathbb{R}^{C \times 1 \times 1}$ was computed as Eq. (1).

$$\mathbf{A}_{spec} = \sigma(\text{MLP}(\text{AvgPool}(\mathbf{F})) + \text{MLP}(\text{MaxPool}(\mathbf{F}))) \quad (1)$$

where σ denoted the sigmoid activation function, AvgPool and MaxPool denoted global average and maximum pooling operations over spatial

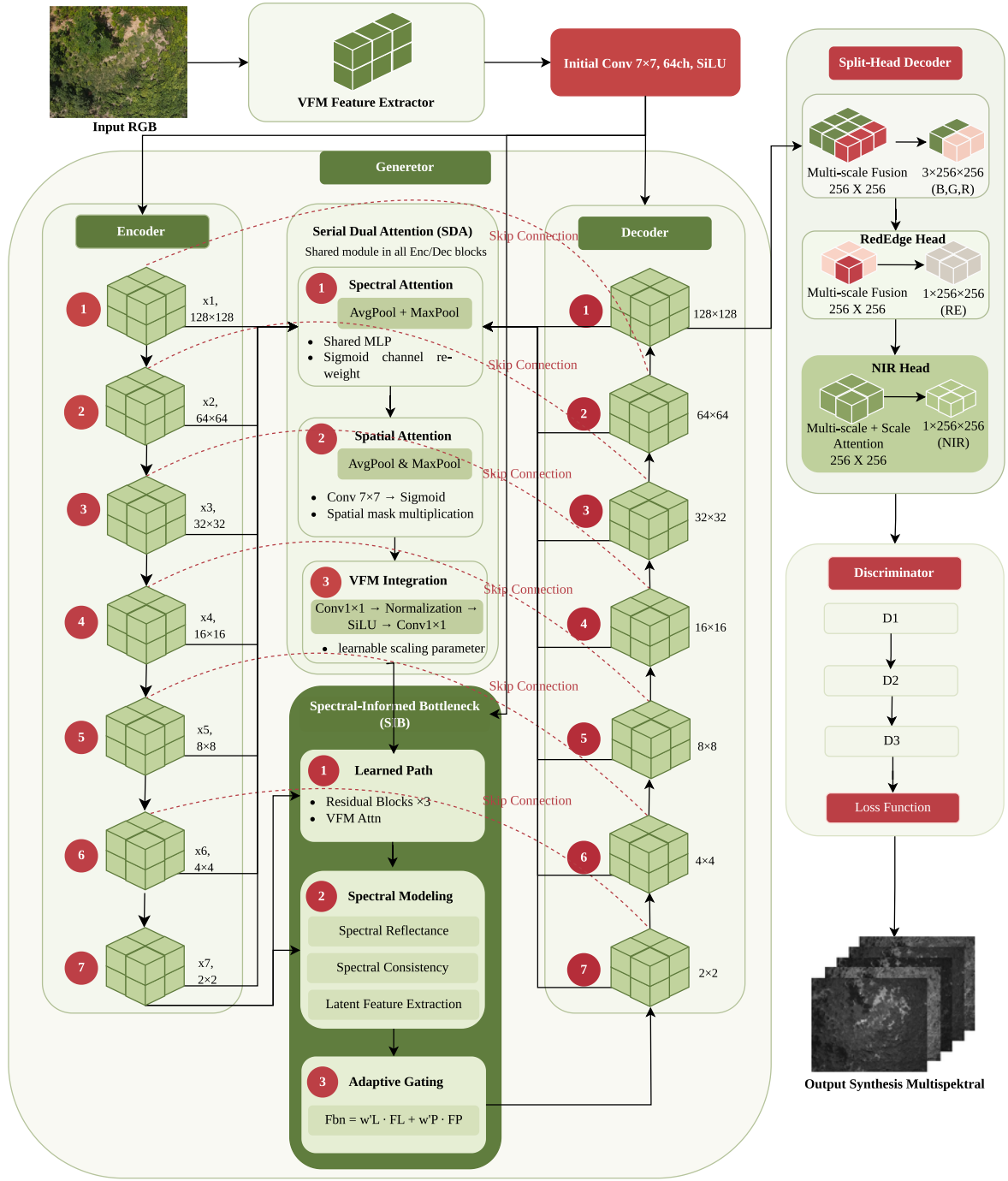


Fig. 3. Overall architecture of the proposed SMH-GAN. The generator comprised a VFM feature extractor (SwinV2), encoder blocks with Serial Dual Attention modules, a Spectral-Informed Bottleneck with dual-path processing, and a Split-Head Decoder with dedicated branches for visible bands (B, G, R), Red Edge, and NIR. The discriminator employed a multi-scale PatchGAN architecture (D1–D3).

dimensions, and MLP consisted of two fully connected layers with channel reduction ratio $r = 16$ and SiLU activation. Spectral-refined features $\mathbf{F}_{spec} \in \mathbb{R}^{C \times H \times W}$ were obtained by element-wise multiplication as Eq. (2) (Rasyidah et al., 2026).

$$\mathbf{F}_{spec} = \mathbf{F} \otimes \mathbf{A}_{spec} \quad (2)$$

where $\otimes$ denoted element-wise multiplication with broadcast along spatial dimensions.

Spatial attention $\mathbf{A}_{spat} \in \mathbb{R}^{1 \times H \times W}$ was then applied to the spectral-refined features as Eq. (3) (Rasyidah et al., 2026).

$$\mathbf{A}_{spat} = \sigma \left(\text{Conv}_{7 \times 7} \left([\text{AvgPool}_c(\mathbf{F}_{spec}); \text{MaxPool}_c(\mathbf{F}_{spec})] \right) \right) \quad (3)$$

where $[\cdot; \cdot]$ denoted concatenation along the channel dimension, subscript c indicated pooling over the channel dimension to produce single-channel maps, and $\text{Conv}_{7 \times 7}$ denoted a convolutional layer with kernel size 7×7 . Pre-trained SwinV2 features $\mathbf{F}_{VFM}$ were then integrated through learned projection as Eq. (4).

$$\mathbf{F}_{out} = (\mathbf{F}_{spec} \otimes \mathbf{A}_{spat}) + \beta \cdot \text{Proj}(\mathbf{F}_{VFM}) \quad (4)$$

where $\mathbf{F}_{out} \in \mathbb{R}^{C \times H \times W}$ was the final output feature tensor, $\text{Proj}(\cdot)$ denoted a projection consisting of Conv-GroupNorm-SiLU-Conv layers that aligned $\mathbf{F}_{VFM}$ to the decoder channel dimension, and $\beta \in \mathbb{R}$ was a

learnable scale parameter initialized to 0.1 to control the contribution of VFM features at the start of training.

3.2.3. Spectral-informed bottleneck

Pix2Next processed the bottleneck layer through standard residual blocks without incorporating constraints based on vegetation spectral characteristics. The latent representation at the encoder–decoder transition therefore captured data-driven features without any domain-informed regularization, which provided no explicit mechanism to preserve inter-band spectral relationships governed by pigment absorption and leaf cellular structure (Gitelson et al., 2003). SIB addressed this limitation by introducing a dual-path bottleneck that processed features through both a data-driven learned path and a spectral-informed path in parallel. Adaptive gating then determined the contribution of each path based on the content of the latent representation. This design enabled the network to incorporate spectral consistency constraints at the bottleneck level without requiring target band reference inputs.

The Spectral-Informed Bottleneck (SIB) incorporated domain-informed modeling at the encoder–decoder transition through dual-path processing with adaptive gating.

Learned Path: Let $\mathbf{F}_{learned}^{(i)} \in \mathbb{R}^{C \times H \times W}$ denote the learned path feature at the i th residual block, where C , H , and W denote channel count, height, and width respectively. Three standard residual blocks received cross-attention inputs from VFM backbone features $\mathbf{F}_{VFM}$ as Eq. (5).

$$\mathbf{F}_{learned}^{(i)} = \mathbf{F}_{learned}^{(i-1)} + \text{Conv-GN-SiLU}^{(2)}(\mathbf{F}_{learned}^{(i-1)}) + \text{CrossAttn}(\mathbf{F}_{VFM}) \quad (5)$$

where $\text{Conv-GN-SiLU}^{(2)}$ denoted two sequential convolution-GroupNorm-SiLU blocks and $\text{CrossAttn}(\mathbf{F}_{VFM})$ denoted cross-attention between the current feature and the pre-trained VFM representation.

Spectral-Informed Path: Three sequential modules processed the input features $\mathbf{F}_{in} \in \mathbb{R}^{C \times H \times W}$. The Spectral Reflectance Module employed $M = 8$ learnable prototype vectors $\mathbf{P}_m \in \mathbb{R}^{C \times 1 \times 1}$, where $m \in \{1, \dots, M\}$ indexed each prototype, trained end-to-end to capture material-specific spectral patterns as Eq. (6).

$$\mathbf{F}_{ref} = \mathbf{F}_{in} + \alpha_{ref} \sum_{m=1}^M w_m \cdot \mathbf{P}_m \quad (6)$$

where $\mathbf{F}_{ref} \in \mathbb{R}^{C \times H \times W}$ was the reflectance-refined feature, $w_m \in \mathbb{R}$ were attention weights computed through a convolutional network with softmax normalization over m , and $\alpha_{ref} \in \mathbb{R}$ was a learnable scale parameter initialized to 0.1. The model applied no temperature scaling to this softmax, which followed the standard formulation. The residual scaling by α_{ref} bounded the contribution of the softmax-weighted prototype path by construction, so the path contributed a small fraction of the latent representation at the start of training. The Spectral Consistency Module applied channel mixing via convolutional bottleneck as Eq. (7).

$$\mathbf{F}_{cons} = \mathbf{F}_{ref} + \alpha_{cons} \cdot \text{Conv}_{1 \times 1}(\text{GN-SiLU}(\text{Conv}_{1 \times 1}(\mathbf{F}_{ref}))) \quad (7)$$

where $\mathbf{F}_{cons} \in \mathbb{R}^{C \times H \times W}$ was the consistency-refined feature and $\alpha_{cons} \in \mathbb{R}$ was a learnable scale parameter initialized to 0.1. The Latent Feature Extraction Module compressed features through a channel reduction sequence ($C \rightarrow C/2 \rightarrow 3$) with Tanh activation as Eq. (8).

$$\mathbf{I}_{latent} = \tanh(\text{Conv}_{1 \times 1}(\text{SiLU}(\text{Conv}_{3 \times 3}(\mathbf{F}_{cons})))) \quad (8)$$

where $\mathbf{I}_{latent} \in \mathbb{R}^{3 \times H \times W}$ was the compressed latent representation encoding dominant spectral patterns in a three-channel space. These representations were fused with $\mathbf{F}_{cons}$ through 1×1 convolution as Eq. (9).

$$\mathbf{F}_{SIB} = \mathbf{F}_{cons} + \alpha_{latent} \cdot (\text{Conv}_{1 \times 1}([\mathbf{F}_{cons}; \mathbf{I}_{latent}]) - \mathbf{F}_{cons}) \quad (9)$$

where $\mathbf{F}_{SIB} \in \mathbb{R}^{C \times H \times W}$ was the final spectral-informed path output, $[\mathbf{F}_{cons}; \mathbf{I}_{latent}]$ denoted concatenation along the channel dimension, and $\alpha_{latent} \in \mathbb{R}$ was a learnable scale parameter initialized to 0.1.

Adaptive Gating: The two paths were combined through normalized gating weights. Let $w_L \in \mathbb{R}^{C \times 1 \times 1}$ and $w_P \in \mathbb{R}^{C \times 1 \times 1}$ denote the unnormalized gate values for the learned and spectral-informed paths, computed as Eqs. (10)–(11).

$$w_L = \sigma(\text{Conv}_{1 \times 1}(\text{AvgPool}(\mathbf{F}_{learned}))) \quad (10)$$

$$w_P = \sigma(\text{Conv}_{1 \times 1}(\text{AvgPool}(\mathbf{F}_{SIB}))) \quad (11)$$

where σ denoted the sigmoid activation function and AvgPool denoted global average pooling over spatial dimensions. Normalized weights w'_L and w'_P were obtained as Eq. (12).

$$w'_L = \frac{w_L}{w_L + w_P + \epsilon}, \quad w'_P = \frac{w_P}{w_L + w_P + \epsilon} \quad (12)$$

where $\epsilon = 10^{-8}$ prevented division by zero and ensured $w'_L + w'_P = 1$ for each channel. The final bottleneck output $\mathbf{F}_{bn} \in \mathbb{R}^{C \times H \times W}$ was computed as Eq. (13).

$$\mathbf{F}_{bn} = w'_L \odot \mathbf{F}_{learned} + w'_P \odot \mathbf{F}_{SIB} \quad (13)$$

where $\odot$ denoted element-wise multiplication with broadcast along spatial dimensions.

3.2.4. Split-head decoder

Pix2Next employed a unified decoder that generated all output bands through a single output layer. This design did not account for the divergent optimization characteristics across spectral bands: visible band reconstruction relied on direct correspondence with RGB input channels and produced small, stable loss gradients, whereas NIR reconstruction required inference from indirect cues and produced larger, more variable gradients. Under joint optimization, these gradient distributions interfered with each other, consistent with observations in multi-task learning where tasks with conflicting loss landscapes produced suboptimal convergence for both objectives (Standley et al., 2020). The Split-Head Decoder addressed this limitation by assigning dedicated branches to band groups with distinct reconstruction characteristics, so that gradient updates for each group did not interfere with the optimization of other groups.

The Split-Head Decoder employed three dedicated branches. Let $\mathbf{d}_1 \in \mathbb{R}^{C_{base} \times H \times W}$ denote the decoder output feature and $\mathbf{x}_0 \in \mathbb{R}^{C_{base} \times H \times W}$ the initial encoder feature, where C_{base} is the base channel count, H the feature map height, and W the feature map width. Each branch received the concatenation $[\mathbf{d}_1; \mathbf{x}_0] \in \mathbb{R}^{2C_{base} \times H \times W}$ as input.

The visible band branch (Blue, Green, Red) processed input features through two Conv-GroupNorm-SiLU blocks. Multi-scale features were extracted through three parallel convolutions with kernel sizes 3×3 , 5×5 , and 7×7 , and fused through a 1×1 convolution to produce the visible band output $\mathbf{Y}_{vis} \in \mathbb{R}^{3 \times H \times W}$, where the three channels corresponded to Blue (450 nm), Green (560 nm), and Red (650 nm).

The Red Edge branch processed input features through three Conv-GroupNorm-SiLU blocks with the same multi-scale extraction strategy to produce the Red Edge output $\mathbf{Y}_{RE} \in \mathbb{R}^{1 \times H \times W}$ at 730 nm. The additional depth relative to the visible band branch reflected the greater reconstruction difficulty of this transition band.

The NIR branch processed input features through three residual blocks, each consisting of two convolutional layers (six convolutional layers total). Let $\mathbf{S}_{fine}$, $\mathbf{S}_{medium}$, and $\mathbf{S}_{coarse} \in \mathbb{R}^{C' \times H \times W}$ denote multi-scale feature maps at three receptive field sizes, where C' is the intermediate channel count. Scale attention weights $\mathbf{W}_{scale} \in \mathbb{R}^{3 \times H \times W}$ were computed as Eq. (14).

$$\mathbf{W}_{scale} = \text{Softmax}(\text{Conv}_{1 \times 1}([\mathbf{S}_{fine}; \mathbf{S}_{medium}; \mathbf{S}_{coarse}])) \quad (14)$$

where $[\mathbf{S}_{fine}; \mathbf{S}_{medium}; \mathbf{S}_{coarse}]$ denoted concatenation along the channel dimension and Softmax was applied along the scale dimension to produce normalized weights that summed to 1. The weighted combination

produced the NIR output $\mathbf{Y}_{NIR} \in \mathbb{R}^{1 \times H \times W}$ at 840 nm. The final five-band multispectral output $\mathbf{Y}$ was formed by concatenation as Eq. (15).

$$\mathbf{Y} = [\mathbf{Y}_{vis}, \mathbf{Y}_{RE}, \mathbf{Y}_{NIR}] \in \mathbb{R}^{5 \times H \times W} \quad (15)$$

where the five output channels corresponded to Blue (450 nm), Green (560 nm), Red (650 nm), Red Edge (730 nm), and NIR (840 nm).

3.3. Loss functions

The training objective combined four complementary terms to address structural fidelity, perceptual quality, adversarial realism, and spectral consistency.

3.3.1. Structural similarity loss

The Structural Similarity Index (SSIM) (Wang et al., 2004) was adopted as the primary reconstruction loss due to its sensitivity to structural distortions rather than pixel-wise differences. The SSIM loss was computed per band and averaged as Eq. (16).

$$\mathcal{L}_{SSIM} = 1 - \frac{1}{B} \sum_{b=1}^B \text{SSIM}(\hat{\mathbf{Y}}_b, \mathbf{Y}_b) \quad (16)$$

where $B = 5$ was the number of output bands, $\hat{\mathbf{Y}}_b$ the synthesized band, and $\mathbf{Y}_b$ the ground truth.

3.3.2. Feature matching loss

Feature matching loss stabilized GAN training by comparing feature activations at multiple discriminator layers (Salimans et al., 2016) as Eq. (17).

$$\mathcal{L}_{FM} = \sum_{i=1}^L \frac{1}{N_i} \left\| D^{(i)}(\mathbf{X}, \hat{\mathbf{Y}}) - D^{(i)}(\mathbf{X}, \mathbf{Y}) \right\|_1 \quad (17)$$

where $D^{(i)}$ denoted the feature map at the i th discriminator layer and N_i the number of elements in each feature map.

3.3.3. Adversarial loss

The Least Squares GAN (LSGAN) formulation (Mao et al., 2017) was adopted for training stability. A multi-scale discriminator with three scales (D1, D2, D3) captured both fine details and global structure. Each scale was a PatchGAN that received the channel-wise concatenation of the RGB input and the five multispectral bands as an eight-channel input, and the three scales operated on the input downsampled by average pooling. Each PatchGAN applied five convolutional layers with 4×4 kernels and channel widths of 64, 128, 256, and 512 followed by a single-channel output, with stride two in the first three layers and stride one in the last two. Instance normalization followed each convolutional layer except the first and the output, and a LeakyReLU activation with slope 0.2 followed each layer except the output. The intermediate activations supplied the feature-matching loss of Eq. (17).

3.3.4. Spectral ratio consistency loss

Per-band reconstruction losses optimized each band output independently and provided no mechanism to preserve the inter-band ratios on which vegetation index computation depended (Liu et al., 2022; Gao et al., 2022). A spectral ratio consistency loss $\mathcal{L}_{spec}$ was introduced to constrain these relationships during training.

The loss combined four terms with distinct and non-overlapping supervisory scopes. $\mathcal{L}_{NDVI}$ and $\mathcal{L}_{NDRE}$ constrained non-linear normalized ratios between non-adjacent band pairs, NIR–Red and NIR–Red Edge respectively. $\mathcal{L}_{spectral}$ did not supervise, as it targeted linear differences between adjacent pairs only. Both involved the NIR band, which carried the highest reconstruction difficulty among all five output bands. $\mathcal{L}_{NDVI}$ targeted the ratio associated with canopy vigor assessment (Boiarskii and Hasegawa, 2019), while $\mathcal{L}_{NDRE}$ targeted the ratio associated with chlorophyll sensitivity in conditions where

NDVI saturated (Boiarskii and Hasegawa, 2019; Gitelson et al., 2003). Neither ratio term supervised the Blue–Green or Green–Red transitions, as these band pairs did not appear in either formulation. $\mathcal{L}_{spectral}$ addressed this gap through L1 consistency on all four adjacent inter-band difference profiles (M. Zhu et al., 2024; Gao et al., 2022). The Red Edge–NIR pair appeared in both $\mathcal{L}_{spectral}$ and $\mathcal{L}_{NDRE}$, but the two terms were not redundant: $\mathcal{L}_{spectral}$ penalized linear differences between adjacent bands, whereas $\mathcal{L}_{NDRE}$ penalized deviations in the non-linear normalized ratio between the same pair. $\mathcal{L}_{spatial}$ operated on a different dimension from the three spectral terms. Per-band losses measured patch-level structural similarity, whereas $\mathcal{L}_{spatial}$ penalized pixel-level spatial gradient deviations within each synthesized band. The per-band weights $\mathbf{w} = \{2.0, 0.5, 0.5, 1.5, 2.0\}$ in Eq. (22) were set before training based on the RGB–multispectral correlation structure rather than through a search. Bands without strong direct single-channel RGB correspondence depend on multi-channel inference and carry a greater risk of spatial gradient artifacts, so they received higher weights. The Blue band reached only $r = 0.49$ with RGB Blue and relied on RGB Red and Green at $r = 0.58$ and $r = 0.60$, and the NIR band reached between $r = -0.003$ and $r = 0.50$ across the RGB channels, while the Green and Red bands held strong direct correspondence at $r = 0.84$ and $r = 0.78$. Blue and NIR were therefore assigned a weight of 2.0, Green and Red a weight of 0.5, and Red Edge an intermediate weight of 1.5, in line with its transitional correspondence of $r = 0.13$ with RGB Blue and $r = 0.67$ with RGB Green. The four terms were combined as Eq. (18).

$$\mathcal{L}_{spec} = \mathcal{L}_{NDVI} + \mathcal{L}_{NDRE} + \mathcal{L}_{spectral} + \mathcal{L}_{spatial} \quad (18)$$

where $\mathcal{L}_{NDVI}$ penalized deviations in the NIR–Red normalized difference ratio, $\mathcal{L}_{NDRE}$ penalized deviations in the NIR–Red Edge normalized difference ratio, $\mathcal{L}_{spectral}$ penalized deviations in reflectance transitions across adjacent band pairs, and $\mathcal{L}_{spatial}$ penalized deviations in spatial gradients within each synthesized band.

The NDVI ratio loss was defined as Eq. (19).

$$\mathcal{L}_{NDVI} = \left\| \frac{\hat{\mathbf{Y}}_{NIR} - \hat{\mathbf{Y}}_{Red}}{\hat{\mathbf{Y}}_{NIR} + \hat{\mathbf{Y}}_{Red} + \epsilon} - \frac{\mathbf{Y}_{NIR} - \mathbf{Y}_{Red}}{\mathbf{Y}_{NIR} + \mathbf{Y}_{Red} + \epsilon} \right\|_1 \quad (19)$$

where $\hat{\mathbf{Y}}_{NIR}, \hat{\mathbf{Y}}_{Red} \in \mathbb{R}^{H \times W}$ denoted the synthesized NIR and Red band outputs, $\mathbf{Y}_{NIR}, \mathbf{Y}_{Red} \in \mathbb{R}^{H \times W}$ the corresponding ground truth bands, and $\epsilon = 10^{-6}$ prevented division by zero. The NDRE ratio loss followed the same formulation for the NIR–Red Edge ratio (Gitelson et al., 2003) as Eq. (20).

$$\mathcal{L}_{NDRE} = \left\| \frac{\hat{\mathbf{Y}}_{NIR} - \hat{\mathbf{Y}}_{RE}}{\hat{\mathbf{Y}}_{NIR} + \hat{\mathbf{Y}}_{RE} + \epsilon} - \frac{\mathbf{Y}_{NIR} - \mathbf{Y}_{RE}}{\mathbf{Y}_{NIR} + \mathbf{Y}_{RE} + \epsilon} \right\|_1 \quad (20)$$

where $\hat{\mathbf{Y}}_{RE}, \mathbf{Y}_{RE} \in \mathbb{R}^{H \times W}$ denoted the synthesized and ground truth Red Edge band outputs. The spectral gradient loss was defined as Eq. (21).

$$\mathcal{L}_{spectral} = \sum_{b=1}^{B-1} \left\| (\hat{\mathbf{Y}}_{b+1} - \hat{\mathbf{Y}}_b) - (\mathbf{Y}_{b+1} - \mathbf{Y}_b) \right\|_1 \quad (21)$$

where $b \in \{1, \dots, B-1\}$ indexed adjacent band pairs ordered by central wavelength (Blue, Green, Red, Red Edge, NIR), $B = 5$ was the total number of output bands, $\hat{\mathbf{Y}}_b \in \mathbb{R}^{H \times W}$ was the synthesized band at index b , and $\mathbf{Y}_b \in \mathbb{R}^{H \times W}$ the corresponding ground truth band. The spatial gradient matching loss was defined as Eq. (22).

$$\mathcal{L}_{spatial} = \frac{1}{\sum_{c=1}^B w_c} \sum_{c=1}^B w_c \left(\mathcal{E}_c^h + \mathcal{E}_c^w \right) \quad (22)$$

where the per-band gradient matching errors were

$$\mathcal{E}_c^h = \frac{1}{(H-1)W} \sum_{i=2}^H \sum_{j=1}^W \omega_{c,i,j}^h \left| \Delta_{c,i,j}^h(\hat{\mathbf{Y}}) - \Delta_{c,i,j}^h(\mathbf{Y}) \right| \quad (23)$$

with $\Delta_{c,i,j}^h(\mathbf{Y}) = \mathbf{Y}_{c,i,j} - \mathbf{Y}_{c,i-1,j}$ the vertical spatial gradient of band c at pixel (i, j) , and $\mathcal{E}_c^w$ defined analogously along the horizontal direction. The edge-aware weight was

$$\omega_{c,i,j}^h = \exp\left(-\alpha \left| \Delta_{c,i,j}^h(\mathbf{Y}) \right| \right) \quad (24)$$

where $\alpha = 10$ controlled the degree of edge suppression in the penalty weighting.

3.3.5. Total training objective

The total generator loss combined all four terms as Eq. (25).

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{SSIM} + \lambda_2 \mathcal{L}_{FM} + \lambda_3 \mathcal{L}_{GAN} + \lambda_4 \mathcal{L}_{spec} \quad (25)$$

Loss weights were set as $\lambda_1 = 10.0$, $\lambda_2 = 10.0$, $\lambda_3 = 1.0$, and $\lambda_4 = 10.0$. The adversarial term received unit weight, and the structural, feature matching, and spectral consistency terms received an order-of-magnitude larger weight to prioritize reconstruction fidelity over adversarial realism. Section 4.4 reports a sensitivity analysis of the spectral consistency weight λ_4 across reconstruction quality and vegetation index consistency.

3.4. Implementation details

The model was implemented in PyTorch 2.0 and trained on an NVIDIA T4 GPU (16 GB) for 50 epochs with a batch size of 4. The generator and discriminator were each optimized with Adam ($\beta_1 = 0.5$, $\beta_2 = 0.999$). The generator used an initial learning rate of 2×10^{-4} and the discriminator used 2×10^{-5} , and both followed a cosine-annealing schedule with $T_{max} = 50$ and $\eta_{min} = 1 \times 10^{-6}$. The SwinV2-Tiny backbone was initialized with ImageNet-pretrained weights and held frozen, and the remaining layers used the default PyTorch initialization. Training ran to completion without early stopping, and the checkpoint with the highest validation SSIM was retained for evaluation. A fixed random seed of 42 was applied across all experiments. At inference, only the generator was used, in evaluation mode and without gradient computation, and it produced the five multispectral bands in a single forward pass from the RGB input.

4. Results

The proposed method was evaluated through seven analyses: (1) spectral correlation analysis to characterize synthesis difficulty, (2) comparison with state-of-the-art methods, (3) statistical significance testing, (4) ablation study, (5) VFM backbone comparison, (6) qualitative evaluation, and (7) validation of vegetation indices as indicators of practical utility for sustainable agroforestry monitoring.

4.1. Spectral correlation analysis

Pearson correlations were computed between the three RGB input channels and the five multispectral target bands across 300,000 pixel pairs sampled from the six training plots. Fig. 4 presents the resulting correlation matrix, which revealed three distinct regimes that shaped the synthesis task difficulty.

First, RGB channels exhibited strong correlations with the visible-range multispectral targets. MS Green (560 nm) correlated with RGB Green at $r = 0.84$, the highest RGB–MS correspondence among visible bands. MS Red (650 nm) correlated with RGB Red at $r = 0.78$, which reflected direct overlap between the broadband RGB Red channel and the narrowband multispectral Red channel. MS Blue (450 nm) correlated with RGB Red at $r = 0.58$ and with RGB Green at $r = 0.60$, higher than its correlation with the corresponding RGB Blue channel ($r = 0.49$). This pattern indicated that Blue band reconstruction relied on multi-channel RGB cues rather than direct channel correspondence alone, consistent with the spectral overlap between the RGB Red channel and shorter wavelengths in typical consumer camera sensors.

Second, correlations degraded substantially for non-visible target bands. RGB Blue–NIR correlation reached $r = -0.003$, indicating the absence of linear correlation between these channels. RGB Red–NIR correlation was $r = 0.25$, and RGB Green–NIR reached $r = 0.50$. These values indicated that NIR inference relied predominantly on the Green channel among RGB inputs, with Blue contributing no information and

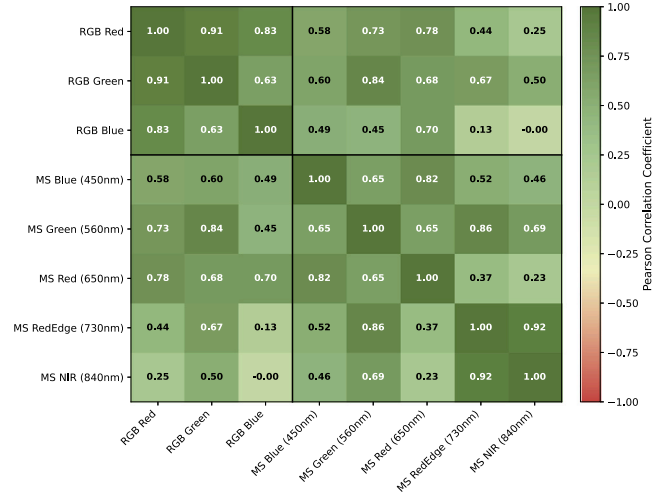


Fig. 4. Pearson correlation matrix between three RGB input channels (rows/columns 1–3) and five multispectral target bands (rows/columns 4–8), computed across 300,000 pixel pairs from the training plots.

Red providing minimal signal. MS RedEdge showed an intermediate pattern, with RGB Blue–RedEdge correlation of only $r = 0.13$ alongside RGB Green–RedEdge of $r = 0.67$. This heterogeneity was consistent with the transitional position of the 730 nm wavelength between chlorophyll absorption and the NIR reflectance plateau.

Third, strong inter-band correlations persisted within the multi-spectral domain despite the weak RGB–NIR relationships. MS RedEdge correlated with MS NIR at $r = 0.92$, reflecting the shared dependence of both bands on leaf cellular structure and mesophyll scattering (Gitelson et al., 2003). MS Green correlated with MS RedEdge at $r = 0.86$, and MS Blue correlated with MS Red at $r = 0.82$. The target spectral domain therefore exhibited strong internal structure even where input–target correspondence was weak.

The correlation patterns established the fundamental challenge of five-band synthesis. Non-visible band reconstruction required inference from weakly-correlated RGB inputs, while inter-band target correlations remained strong ($r = 0.82$ – 0.92) across both visible and non-visible regimes. This asymmetry motivated two architectural decisions evaluated in subsequent experiments. The Split-Head Decoder addressed weak RGB–NIR input correspondence through dedicated inference capacity for non-visible bands. The Spectral Consistency Loss exploited strong inter-band target correlations through ratio preservation constraints during training.

4.2. Comparison with state-of-the-art methods

The proposed method, a generative adversarial network, was compared against three baseline methods, two adversarial networks, Pix2Pix (Isola et al., 2017) and Pix2Next (Jin et al., 2025), and one non-adversarial cycle-consistent network, TaijiGNN (Liu et al., 2021). Table 2 presents the quantitative results across all spectral bands. All competing methods used the identical plot-level partition, the 256×256 input resolution, the Adam optimizer, the same number of training epochs, and the same evaluation metric set, which ensured a fair comparison.

SMH-GAN outperformed all three baseline methods across every spectral band. Average SSIM rose from 0.836 (Pix2Next) to 0.876, a 4.8% relative gain. The NIR band yielded the largest absolute improvement (+0.127 SSIM, a 21.2% relative gain over Pix2Next), which is consistent with the contribution of the Split-Head Decoder in providing a dedicated optimization pathway for non-visible band reconstruction. Visible bands showed more modest absolute gains, consistent with

Table 2
Quantitative comparison with state-of-the-art methods.

Band	Method	SSIM↑	PSNR↑	RMSE↓	MAE↓	LPIPS↓	DISTS↓	Corr↑
Blue	Pix2Pix (Isola et al., 2017)	0.909 ± 0.033	38.94 ± 2.01	0.0116 ± 0.003	0.0091 ± 0.003	0.246 ± 0.027	0.139 ± 0.026	0.739 ± 0.110
	TaijiGNN (Liu et al., 2021)	0.886 ± 0.038	35.88 ± 1.68	0.0164 ± 0.003	0.0127 ± 0.003	0.275 ± 0.027	0.172 ± 0.030	0.416 ± 0.162
	Pix2Next (Jin et al., 2025)	0.913 ± 0.031	38.69 ± 2.24	0.0120 ± 0.003	0.0095 ± 0.003	0.268 ± 0.027	0.156 ± 0.028	0.757 ± 0.098
	SMH-GAN	0.936 ± 0.037	40.87 ± 2.28	0.0094 ± 0.003	0.0071 ± 0.002	0.206 ± 0.055	0.111 ± 0.040	0.811 ± 0.128
Green	Pix2Pix (Isola et al., 2017)	0.902 ± 0.016	35.54 ± 1.55	0.0170 ± 0.003	0.0127 ± 0.003	0.145 ± 0.016	0.106 ± 0.015	0.908 ± 0.008
	TaijiGNN (Liu et al., 2021)	0.874 ± 0.024	32.11 ± 1.40	0.0251 ± 0.005	0.0198 ± 0.004	0.222 ± 0.022	0.156 ± 0.021	0.794 ± 0.030
	Pix2Next (Jin et al., 2025)	0.904 ± 0.013	35.42 ± 1.25	0.0171 ± 0.003	0.0130 ± 0.002	0.151 ± 0.017	0.107 ± 0.012	0.909 ± 0.010
	SMH-GAN	0.913 ± 0.012	36.27 ± 0.84	0.0154 ± 0.002	0.0115 ± 0.001	0.141 ± 0.016	0.099 ± 0.012	0.916 ± 0.009
Red	Pix2Pix (Isola et al., 2017)	0.946 ± 0.011	39.79 ± 1.90	0.0105 ± 0.003	0.0080 ± 0.002	0.139 ± 0.008	0.075 ± 0.009	0.920 ± 0.020
	TaijiGNN (Liu et al., 2021)	0.921 ± 0.016	36.95 ± 1.51	0.0144 ± 0.003	0.0109 ± 0.002	0.224 ± 0.017	0.132 ± 0.012	0.803 ± 0.045
	Pix2Next (Jin et al., 2025)	0.943 ± 0.024	39.23 ± 2.89	0.0116 ± 0.005	0.0092 ± 0.005	0.142 ± 0.012	0.077 ± 0.018	0.927 ± 0.018
	SMH-GAN	0.962 ± 0.008	42.11 ± 1.53	0.0080 ± 0.002	0.0058 ± 0.001	0.124 ± 0.008	0.060 ± 0.008	0.939 ± 0.015
Red edge	Pix2Pix (Isola et al., 2017)	0.829 ± 0.034	29.48 ± 0.93	0.0338 ± 0.004	0.0260 ± 0.003	0.198 ± 0.027	0.174 ± 0.025	0.898 ± 0.022
	TaijiGNN (Liu et al., 2021)	0.754 ± 0.039	24.84 ± 0.89	0.0576 ± 0.006	0.0446 ± 0.005	0.304 ± 0.032	0.270 ± 0.028	0.678 ± 0.062
	Pix2Next (Jin et al., 2025)	0.820 ± 0.035	29.62 ± 0.59	0.0331 ± 0.002	0.0257 ± 0.002	0.225 ± 0.037	0.186 ± 0.028	0.898 ± 0.026
	SMH-GAN	0.845 ± 0.033	30.11 ± 0.68	0.0313 ± 0.003	0.0240 ± 0.002	0.186 ± 0.027	0.160 ± 0.025	0.907 ± 0.026
NIR	Pix2Pix (Isola et al., 2017)	0.617 ± 0.098	20.93 ± 0.89	0.0904 ± 0.011	0.0705 ± 0.008	0.334 ± 0.036	0.339 ± 0.060	0.819 ± 0.047
	TaijiGNN (Liu et al., 2021)	0.547 ± 0.066	18.47 ± 0.76	0.1198 ± 0.012	0.0945 ± 0.009	0.348 ± 0.035	0.387 ± 0.045	0.658 ± 0.047
	Pix2Next (Jin et al., 2025)	0.598 ± 0.084	21.22 ± 0.78	0.0873 ± 0.009	0.0684 ± 0.006	0.387 ± 0.036	0.364 ± 0.054	0.829 ± 0.043
	SMH-GAN	0.725 ± 0.080	22.39 ± 0.86	0.0763 ± 0.009	0.0590 ± 0.006	0.261 ± 0.034	0.254 ± 0.052	0.873 ± 0.035

Table 3
Wilcoxon signed-rank test: Overall metrics and per-band SSIM comparing Pix2Next and SMH-GAN.

Level	Metric/Band	p-value	Effect size (r)
Overall	SSIM	2.66×10^{-18}	0.868
	PSNR	4.29×10^{-18}	0.863
	RMSE	2.66×10^{-18}	0.868
	LPIPS	2.66×10^{-18}	0.868
Per-band (SSIM)	Blue	5.75×10^{-11}	0.652
	Green	2.83×10^{-18}	0.867
	Red	1.53×10^{-17}	0.848
	Red edge	2.66×10^{-18}	0.868
	NIR	2.66×10^{-18}	0.868

the strong RGB–MS correlations reported in Section 4.1 (MS Red–RGB Red: $r = 0.78$; MS Green–RGB Green: $r = 0.84$), which constrained the headroom for architectural improvement. The NIR band, by contrast, required inference from weakly-correlated inputs (RGB–NIR: $r = 0.00$ – 0.50) and thus benefited most from the dedicated optimization pathway of the Split-Head Decoder. TaijiGNN performed below all other methods on the NIR band (0.547 SSIM), consistent with its MLP-based generator design, which prioritizes spectral mapping over spatial feature extraction and was developed for a different synthesis context.

4.3. Statistical significance analysis

A Wilcoxon signed-rank test was conducted between Pix2Next and SMH-GAN on paired test samples to assess whether the observed improvements were statistically significant. Table 3 presents results for overall evaluation metrics and per-band SSIM.

All improvements were statistically significant ($p < 0.001$) with large effect sizes across all metrics and spectral bands. The Blue band yielded the smallest effect size ($r = 0.652$), consistent with its modest absolute SSIM gain of +0.023 over Pix2Next, a band where direct RGB correspondence limits the contribution of architectural modifications. All other bands exceeded $r = 0.848$, which suggested that the performance gains held across the 101 test RGB-to-multispectral pairs from the two held-out plots rather than arising from a small number of favorable cases.

The Wilcoxon signed-rank test treated each test RGB-to-multispectral pair as one observation. Adjacent UAV images within a plot shared spatial content, so the effective number of independent observations was smaller than the 101 test pairs. The reported p-values

therefore overstate the statistical precision, and the analysis interpreted the results together with the effect sizes in Table 3 and the cross-band consistency of the improvements instead of relying on the p-values alone.

4.4. Ablation study

An incremental ablation study was conducted to evaluate the contribution of each proposed component. Starting from the Pix2Next baseline, components were added sequentially to construct the proposed model. A removal experiment was then performed for the Spectral-Informed Bottleneck, the one component that admitted a clean single-component removal, to assess its contribution within the full model. Table 4 presents SSIM, PSNR, and LPIPS results for all configurations across five spectral bands.

The full model reached the best SSIM and LPIPS across all five bands, with NIR showing the largest SSIM gain (+0.127) and LPIPS improvement (−0.126) over the Pix2Next baseline. The spectral loss traded a small amount of PSNR on Red Edge and NIR for the LPIPS and SSIM gains, so the 3-Head configuration held the highest PSNR on those two bands. The following subsections discuss the contribution of each proposed component through incremental or leave-one-out analyses where applicable.

4.4.1. Effect of serial dual attention

Rows 1–4 of Table 4 isolate the contribution of attention mechanism design. Spectral attention alone improved NIR SSIM by +0.055 while the Green band showed no change (+0.000). This asymmetric response reflected the difference in reconstruction difficulty across bands. Green band synthesis depends on a direct channel correspondence with the RGB Green input, so channel-wise reweighting offered limited selectivity benefit. NIR synthesis, in contrast, required the network to identify which combinations of all three RGB channels correlated with leaf cellular structure. This task benefited from explicit spectral guidance that redirected the network toward more informative feature combinations. Spatial attention showed the strongest effect on Red Edge (+0.016 SSIM), consistent with the role of boundary-level spatial patterns in the transition region of the vegetation reflectance spectrum between Red absorption and the NIR plateau.

The serial combination SDA improved Red SSIM to 0.953 from 0.946 (spectral only) and 0.943 (spatial only). For NIR, SDA matched spatial attention alone (0.664) and exceeded spectral attention alone (0.653). This result suggested that spatial processing contributed more

Table 4
Incremental ablation study.

Configuration	SSIM $\uparrow$					PSNR $\uparrow$					LPIPS $\downarrow$				
	B	G	R	RE	NIR	B	G	R	RE	NIR	B	G	R	RE	NIR
Baseline (Pix2Next)	0.913	0.904	0.943	0.820	0.598	38.69	35.42	39.23	29.62	21.22	0.268	0.151	0.142	0.225	0.387
+ Spectral att. only	0.919	0.904	0.946	0.827	0.653	39.03	35.43	39.38	29.65	21.44	0.263	0.147	0.137	0.207	0.325
+ Spatial att. only	0.918	0.910	0.943	0.836	0.664	38.88	35.66	39.50	29.93	21.62	0.257	0.143	0.132	0.200	0.313
+ SDA	0.923	0.908	0.953	0.832	0.664	39.64	35.86	40.64	29.73	21.57	0.264	0.145	0.127	0.203	0.314
+ SDA + SIB	0.918	0.909	0.945	0.837	0.667	38.87	35.65	39.23	29.88	21.75	0.290	0.144	0.132	0.201	0.318
+ SDA + SIB + 2-Head dec.	0.936	0.909	0.960	0.838	0.713	40.66	35.96	41.67	29.85	22.39	0.242	0.145	0.123	0.197	0.273
+ SDA + SIB + 3-Head dec.	0.936	0.913	0.960	0.844	0.723	40.59	36.18	41.77	30.13	22.50	0.236	0.142	0.124	0.188	0.264
Full (+ Spectral Loss)	0.936	0.913	0.962	0.845	0.725	40.87	36.27	42.11	30.11	22.39	0.206	0.141	0.124	0.186	0.261
Full (w/o SIB)	0.936	0.911	0.961	0.838	0.717	40.64	35.98	41.82	29.84	22.27	0.240	0.143	0.123	0.191	0.268

to NIR improvement than spectral reweighting in the serial design. The attention stage was the one part of the ablation with isolated configurations, because spectral attention and spatial attention were each trained alone and in combination. On the Red band the serial SDA reached the highest SSIM among the three attention configurations, and on the NIR band the serial SDA matched spatial attention alone and did not exceed the larger single-path gain. These comparisons were not converted into a formal interaction estimate, because each configuration was trained once with a fixed seed and a single run provides no measure of the training variability needed to separate an interaction from noise. Green band SSIM decreased slightly under SDA compared to spatial attention alone (0.908 vs. 0.910). For bands with direct RGB correspondence, the additional spectral stage introduced minimal benefit and a small regularization cost.

The correlation structure reported in Section 4.1 supports this interpretation. RGB Green correlated with MS NIR at $r = 0.50$, the highest among RGB channels, while RGB Blue showed complete independence from MS NIR ($r = 0.00$). Spectral reweighting provided the largest gain on the band with heterogeneous RGB correlations because channel-wise attention redistributed emphasis toward the most informative input channels. For bands with concentrated correlation from a single RGB source, channel reweighting had less room to improve feature selection.

4.4.2. Effect of spectral-informed bottleneck

Row 5 of Table 4 shows the effect of adding SIB to the SDA configuration. SIB decreased SSIM for Blue (0.923 $\rightarrow$ 0.918) and Red (0.953 $\rightarrow$ 0.945) while improving Red Edge (0.832 $\rightarrow$ 0.837) and NIR (0.664 $\rightarrow$ 0.667). The LPIPS score for Blue increased from 0.264 to 0.290, which indicated a perceptual quality reduction alongside the structural metric decline.

The trade-off pattern (Blue $\downarrow$ Red $\downarrow$ but RedEdge $\uparrow$ NIR $\uparrow$) was consistent with the learned spectral regularization in SIB emphasizing cross-band relationships at the expense of individual band accuracy for spectrally distant bands. Blue reflectance (450 nm) is governed by atmospheric scattering and pigment absorption, while NIR reflectance (840 nm) depends on internal leaf cellular structure and mesophyll scattering (Gitelson et al., 2003; Vermote et al., 2016). The spectral distance between these bands means that channel mixing optimized for NIR–Red Edge vegetation relationships introduced competing gradient directions for Blue and Red reconstruction. When combined with the Split-Head Decoder in subsequent configurations, this trade-off was reduced because band-specific heads isolated the optimization objectives. A leave-one-out comparison measured the contribution of SIB within the full model. The last row of Table 4 reports the full model with SIB removed. Removal of SIB reduced Red Edge SSIM from 0.845 to 0.838 and NIR SSIM from 0.725 to 0.717, while the visible bands stayed within 0.002. The contribution of SIB therefore concentrated on the Red Edge and NIR bands, consistent with the intended role of the module. The leave-one-out reduction on NIR (0.008) exceeded the increment that SIB produced when first added to the SDA configuration (0.003), which indicated that the contribution of SIB was more evident within the complete architecture. SIB and the spectral loss were the two components that admitted a clean single-component removal, because each

Table 5

Effect of spectral ratio consistency loss components (SSIM).

Configuration	B	G	R	RE	NIR
w/o $\mathcal{L}_{spec}$	0.936	0.913	0.960	0.844	0.723
+ $\mathcal{L}_{NDVI}$ only	0.932	0.914	0.962	0.845	0.720
+ $\mathcal{L}_{NDRE}$ only	0.935	0.913	0.959	0.848	0.719
+ $\mathcal{L}_{spec}$ w/o $\mathcal{L}_{spatial}$	0.932	0.914	0.962	0.845	0.721
Full $\mathcal{L}_{spec}$	0.936	0.913	0.962	0.845	0.725

one was removed without altering any other module. This configuration was trained once with a fixed seed, so the observed difference was not separated from training variability.

4.4.3. Effect of split-head decoder

Rows 5–7 of Table 4 compare the unified and split decoder configurations. The Split-Head architecture produced the largest improvement among all components, with NIR SSIM increasing by +0.056 from the unified to the 3-Head configuration. The LPIPS score for NIR fell from 0.318 to 0.264, which indicated that the improvement extended beyond structural similarity to perceptual quality.

In a unified decoder, gradient updates were shared across all output channels. Visible band reconstruction, which benefited from direct RGB correspondence, produced smaller and more stable loss gradients. NIR reconstruction, which required inferring information absent from the RGB input, produced larger and more variable gradients that interfered with the optimization of visible band parameters. This gradient interference pattern was consistent with findings in multi-task learning (Standley et al., 2020), where tasks with divergent loss landscapes produced conflicting gradient directions and suboptimal convergence for both objectives. The 3-Head configuration separated these optimization paths, and the dedicated NIR head with six convolutional layers provided greater capacity for the harder inference task compared to two layers for visible bands. The separation of Red Edge into its own head further improved Red Edge SSIM from 0.838 to 0.844, consistent with the spectral distinctiveness of this transition band relative to both visible and NIR regions.

4.4.4. Effect of spectral loss components

Table 5 presents the contribution of each spectral loss component evaluated from the 3-Head configuration. Row 7 of Table 4 served as the reference for this analysis.

$\mathcal{L}_{NDVI}$ improved Red SSIM (+0.002) and reduced Blue (−0.004). The Red band received gradient supervision through the NIR–Red ratio, whereas the Blue band received no direct constraint from the NDVI objective and its gradients arose only through shared encoder parameters, which introduced a competing update direction. $\mathcal{L}_{NDRE}$ produced the largest improvement on Red Edge (+0.004) because the NIR–Red Edge ratio directly constrained the Red Edge output, and Red Edge reconstruction benefited from explicit ratio-level supervision that per-band losses did not provide. Both individual ratio components reduced NIR SSIM below the no-loss baseline (0.720 and 0.719 vs. 0.723). Each term constrained a different ratio involving NIR as the

Table 6

Sensitivity of reconstruction quality and vegetation index consistency to the spectral consistency loss weight λ_4 . All other settings were held fixed. $\lambda_4 = 0$ denotes the w/o $\mathcal{L}_{spec}$ configuration in Table 5.

λ_4	Mean SSIM	NIR SSIM	NDVI r	NDRE r
0	0.875	0.723	0.909	0.614
1	0.874	0.717	0.910	0.613
5	0.875	0.718	0.908	0.613
10	0.876	0.725	0.910	0.621
20	0.876	0.722	0.911	0.621

shared variable, and when applied in isolation, the gradient from each ratio objective pulled the NIR output toward different targets. $\mathcal{L}_{NDVI}$ optimized NIR relative to Red, while $\mathcal{L}_{NDRE}$ optimized NIR relative to Red Edge, and no balancing mechanism existed between the two objectives. The combination of spectral terms without $\mathcal{L}_{spatial}$ partially recovered NIR SSIM to 0.721, as $\mathcal{L}_{spectral}$ enforced L1 consistency on adjacent inter-band difference profiles across all band pairs and provided a regularizing signal that reduced the conflict between the two ratio objectives. The full $\mathcal{L}_{spec}$ with $\mathcal{L}_{spatial}$ recovered NIR SSIM to 0.725 and Blue to 0.936. The Blue and NIR bands carried the highest band-specific weights in $\mathcal{L}_{spatial}$, and pixel-level spatial gradient supervision provided a reconstruction constraint orthogonal to the spectral ratio terms, which addressed residual spatial inconsistencies not captured by inter-band objectives.

Beyond the on-off contribution of each spectral loss component, the magnitude of the spectral consistency weight λ_4 in Eq. (25) was examined through a sensitivity analysis across $\{0, 1, 5, 10, 20\}$ with all other settings held fixed, where $\lambda_4 = 0$ corresponds to the w/o $\mathcal{L}_{spec}$ configuration in Table 5. Each configuration used the same plot-level split, the same random seed, and the same schedule. Table 6 reports mean SSIM and NIR SSIM for reconstruction quality, together with NDVI and NDRE correlation for vegetation index consistency. Reconstruction quality stayed within a narrow band across the full range. Mean SSIM ranged from 0.874 to 0.876 and NIR SSIM from 0.717 to 0.725. Vegetation index consistency showed the same pattern, with NDVI correlation from 0.908 to 0.911 and NDRE correlation from 0.613 to 0.621. The synthesized output was therefore stable with respect to λ_4 on both reconstruction and vegetation index consistency. The value $\lambda_4 = 10$ was adopted so the spectral consistency term matched the magnitude of the structural and feature matching terms, each set to 10. Increasing the weight from 10 to 20 produced no meaningful improvement on any band or index, so a larger spectral weight was not adopted.

The attention distributions of the two softmax operations in the bottleneck and the NIR branch were examined on the trained model across the 101 test pairs to assess whether the softmax normalization produced peaked distributions that suppressed alternative feature contributions. The normalized Shannon entropy of the spectral prototype softmax (eight prototypes) reached a mean of 0.999 with a standard deviation of 0.000, and the normalized entropy of the NIR scale-attention softmax (three scales) reached a mean of 0.987 with a standard deviation of 0.001. Both distributions stayed close to the uniform upper bound across every test image, so neither softmax collapsed onto a single component. The NIR scale-attention entropy differed by only 0.001 between the highest and the lowest reconstruction tercile, so the attention distributions remained stable across the evaluated samples. As an additional check, the softmax temperature was varied at inference across $\{0.5, 1.0, 2.0, 4.0\}$ without retraining, which produced sharper and flatter attention distributions around the learned operating point. The mean SSIM remained within 0.0028 of the reference across this range. This result indicated limited sensitivity of the reconstruction to these temperature-induced changes in the attention distributions.

Table 7

VFM backbone comparison (SSIM).

Backbone	B	G	R	RE	NIR
ResNet-50	0.937	0.911	0.961	0.841	0.717
ConvNeXt-T	0.929	0.913	0.958	0.851	0.710
SwinV2-T	0.936	0.913	0.962	0.845	0.725
EfficientNet	0.927	0.914	0.955	0.850	0.713
MobileViT	0.933	0.911	0.958	0.841	0.720

4.5. VFM backbone comparison

Five Vision Foundation Model backbones were evaluated as feature extractors. Table 7 presents the SSIM comparison results across all spectral bands.

SwinV2-Tiny achieved the highest SSIM on Red (0.962) and NIR (0.725). ConvNeXt-T led on Red Edge (0.851) and EfficientNet on Green (0.914), but neither maintained competitive NIR performance. The performance differences across backbones reflected their architectural characteristics. SwinV2-Tiny's hierarchical window-based attention captured long-range spatial dependencies across the full image, which was relevant for NIR synthesis where leaf structural patterns had to be inferred from spatially distributed RGB cues rather than local features alone. The per-band differences among the close backbones came from a single training run and were small relative to the variability that repeated runs would be needed to quantify. The Red Edge value for ConvNeXt-T, the Blue value for ResNet-50, and the NIR value for SwinV2-Tiny each led their band by a narrow margin, and none of these margins was separated from stochastic training variability through statistical testing or repeated runs. The architectural interpretations of these small differences are therefore tentative. The five backbones produced similar and high reconstruction across all bands, which indicates that the method is robust to the choice of backbone and that no single backbone was required to reach this level of reconstruction. SwinV2-Tiny reached high reconstruction across all five bands without a marked weakness on any band.

SwinV2-Tiny was selected as the feature extractor for the final model on the basis of its consistent performance across the five bands, the consistency between its window-based attention and the spatially distributed cues involved in NIR synthesis, and its role as the backbone of the validated end-to-end results in Section 4.2, rather than on any single per-band margin.

4.6. Qualitative evaluation

Visual comparison across methods is presented in Fig. 5. The proposed method produced sharper band images with better texture preservation compared to baseline methods. TaijiGNN exhibited noticeable blurring across all bands, which was consistent with its MLP-based generator design, which extracts spectral features without spatial convolution and tends to smooth local spatial detail. Pix2Pix produced visible artifacts in the NIR band, where the absence of spectral consistency constraints allowed the adversarial objective to generate perceptually plausible but spectrally inconsistent outputs. Pix2Next and SMH-GAN produced visually similar visible band outputs, with differences most evident in Red Edge and NIR.

A representative synthesis result with per-band histograms is shown in Fig. 6. The histogram distributions showed close alignment between synthesized and ground truth values across all bands. Visible bands exhibited near-identical distributions, which reflected their direct RGB channel correspondence. Red Edge and NIR showed slight shifts in peak positions but maintained similar overall shapes, consistent with the inference-based nature of non-visible band reconstruction from RGB input. The histogram comparison characterized the global distribution of synthesized values and served as a supplementary check rather than as evidence of spatial fidelity. The density scatter plots in Fig. 7, the

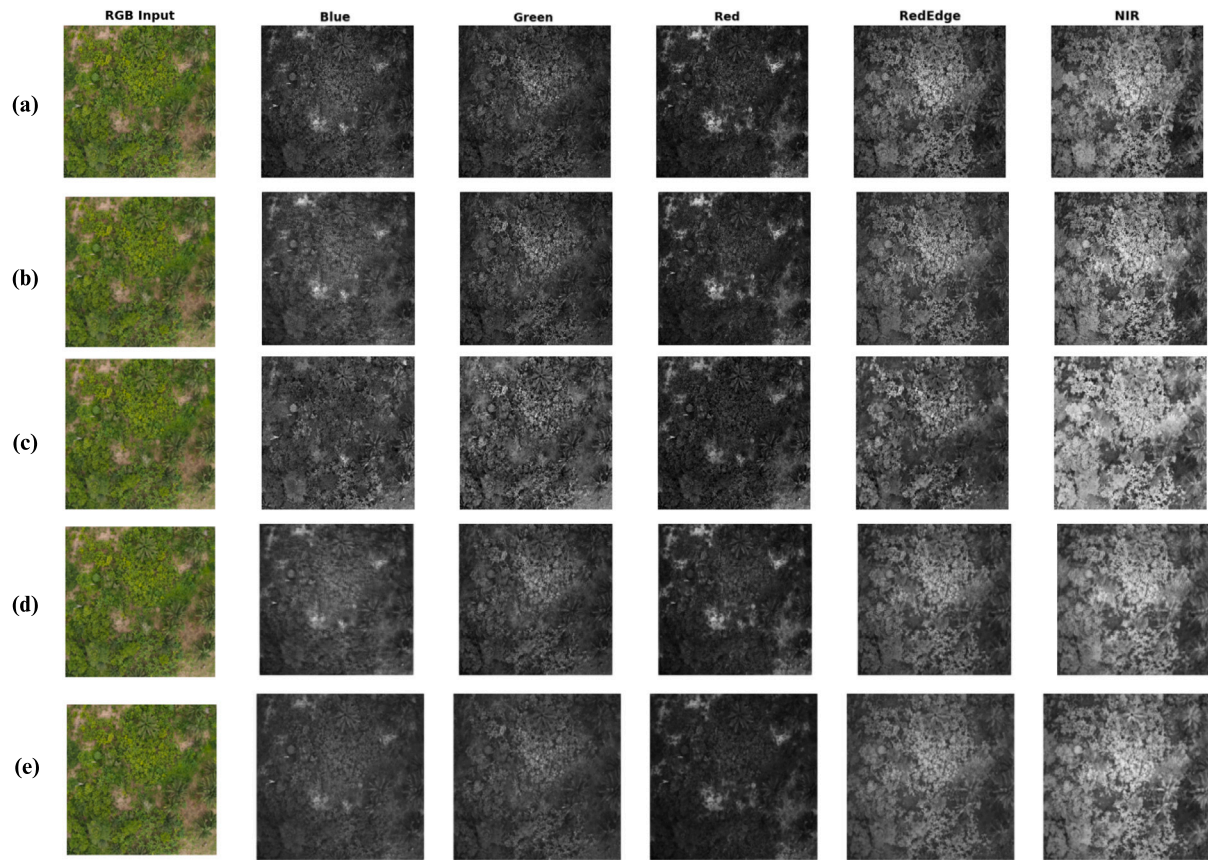


Fig. 5. Visual comparison of synthesized multispectral bands. (a) Ground truth. (b) Pix2Pix. (c) TaijiGNN. (d) Pix2Next. (e) Proposed method.

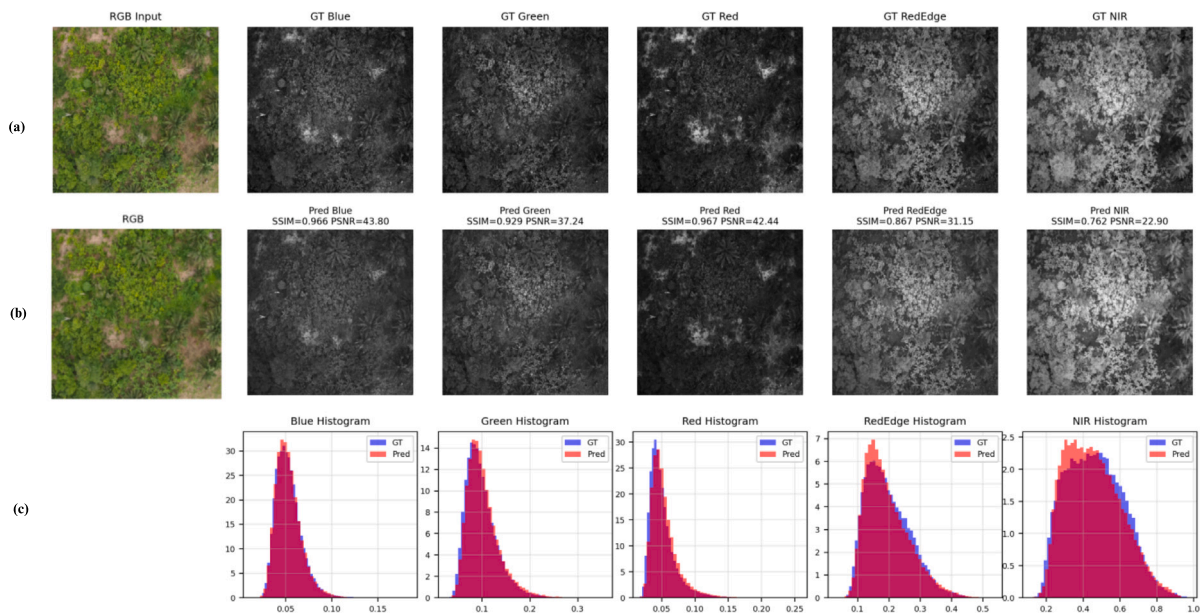


Fig. 6. Representative synthesis result. (a) Ground truth bands with value ranges. (b) Synthesized bands with per-band SSIM and PSNR. (c) Histogram comparison between ground truth (blue) and synthesized (red) for each band.

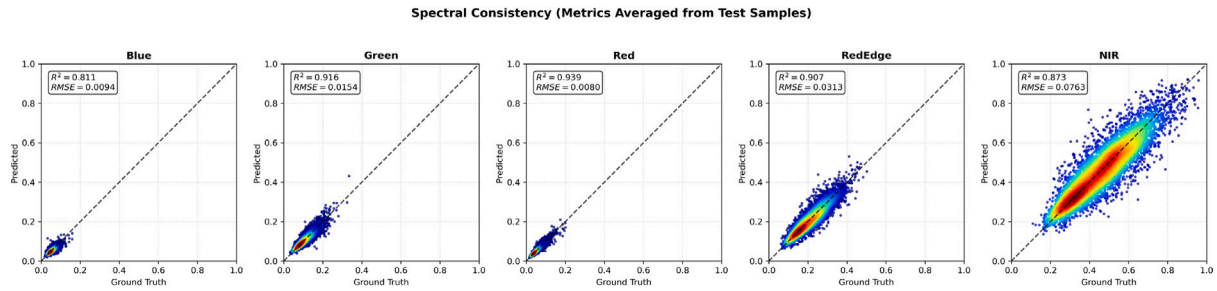


Fig. 7. Density scatter plots of predicted versus ground-truth reflectance for each spectral band across the test set. Points are colored by density (blue: sparse, red/yellow: dense). The diagonal dashed line indicates perfect prediction.

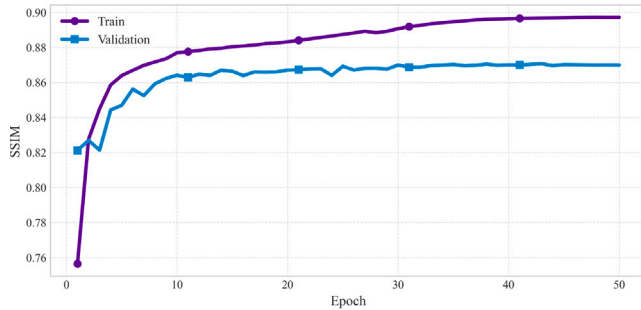


Fig. 8. Training and validation SSIM curves over 50 epochs. The model achieved stable convergence after approximately 30 epochs.

per-band SSIM in Table 2, and the vegetation index maps in Fig. 9 provided the pixel-level and spatial evidence of spectral consistency in heterogeneous canopy regions.

Spectral consistency across all bands was further examined through density scatter plots (Fig. 7). Visible bands showed tight clustering along the diagonal ($R^2 = 0.81\text{--}0.94$), which reflected the availability of direct spectral mapping cues from RGB input. Red Edge ($R^2 = 0.907$) and NIR ($R^2 = 0.873$) showed increased scatter but maintained strong linear correlation, consistent with the absence of direct correspondence for these bands and the reliance on indirect RGB cues. The RMSE range of 0.0094–0.0763 in the reflectance domain indicated small reconstruction errors across all bands, with the narrowest error distribution observed for the Blue band.

Training dynamics are summarized in Fig. 8. The model showed rapid SSIM improvement in the first 10 epochs, followed by gradual convergence. Validation SSIM stabilized around 0.87 after epoch 30 and showed no further gain through epoch 50, which indicated that training had converged and supported the choice of a 50-epoch schedule. The training-validation gap of approximately 0.03 indicated mild overfitting within acceptable bounds given the training-set size of 905 pairs.

4.7. Vegetation index validation

Vegetation indices were computed from both ground truth and synthesized bands to evaluate the practical utility of the synthesized multispectral data for sustainable agroforestry monitoring. Eight indices were analyzed, covering NDVI, Green NDVI (GNDVI), NDRE, Enhanced Vegetation Index (EVI), Soil Adjusted Vegetation Index (SAVI), Optimized SAVI (OSAVI), Modified SAVI (MSAVI), and Atmospherically Resistant Vegetation Index (ARVI). Three metrics were reported for each index, namely the Pearson correlation coefficient (r), the coefficient of determination (R^2), and the Mean Absolute Error (MAE). Pearson correlation was selected as the primary metric because vegetation index values served as relative indicators of canopy condition rather than absolute reflectance quantities, and the practical utility

Table 8

Vegetation index correlation between ground truth and synthesized bands.

Vegetation index	Correlation (r)	R^2	MAE
NDVI	0.9103 ± 0.0255	0.7955 ± 0.0834	0.0344 ± 0.0059
ARVI	0.8925 ± 0.0277	0.7593 ± 0.0754	0.0537 ± 0.0072
OSAVI	0.8883 ± 0.0289	0.7613 ± 0.0773	0.0394 ± 0.0051
SAVI	0.8803 ± 0.0292	0.7496 ± 0.0729	0.0542 ± 0.0061
MSAVI	0.8766 ± 0.0294	0.7415 ± 0.0745	0.0643 ± 0.0073
EVI	0.8706 ± 0.0273	0.7302 ± 0.0674	0.0790 ± 0.0066
GNDVI	0.7791 ± 0.0392	0.5596 ± 0.0872	0.0515 ± 0.0075
NDRE	0.6214 ± 0.0491	0.2717 ± 0.0932	0.0559 ± 0.0083

of synthesized indices for monitoring depended on whether spatial patterns of vegetation vigor and stress were preserved rather than on absolute value accuracy (Fawcett et al., 2020). MAE complemented correlation by quantifying absolute value deviation, which was relevant for applications that relied on threshold-based vegetation classification. Table 8 presents the correlation analysis across the test set, and Fig. 9 presents spatial maps comparing ground truth and predicted index values.

The results in Table 8 revealed a clear performance hierarchy determined by band composition. Indices computed from NIR and Red bands (NDVI, ARVI, OSAVI, SAVI, MSAVI, EVI) achieved correlations exceeding 0.87, with NDVI obtaining the highest ($r = 0.9103$). This result is consistent with the Red band showing the highest absolute SSIM (0.962) and NIR showing the largest improvement over the baseline (+0.127 SSIM), as NDVI computation depends on the relative ratio of these two bands rather than their absolute accuracy. GNDVI achieved a lower correlation ($r = 0.7791$), which reflected the Green band's slightly wider spectral sensitivity relative to the narrowband multispectral sensor channel (Candiago et al., 2015). NDRE obtained the lowest correlation ($r = 0.6214$), consistent with error compounding from both constituent bands: Red Edge SSIM (0.845) and NIR SSIM (0.725) were the two lowest-performing bands, and the normalized difference formulation amplified their combined reconstruction error.

Fig. 10 adds the per-pixel error for NDVI and NDRE on a representative sample at the median NDVI correlation. The NDVI error stayed small across the canopy, and the NDRE error was larger and spread across the scene as fine-scale speckle, consistent with the lower NDRE correlation in Table 8.

The effect of the 256×256 resize on the vegetation indices was assessed on the ground-truth imagery, separate from the synthesis. The structural similarity between each native-resolution band and the same band after a resize to 256×256 pixels and back to the native grid was 0.96 for Blue, 0.94 for Red, 0.81 for Green, 0.76 for Red Edge, and 0.66 for NIR, so the resize reduced fine spatial detail most in the Red Edge and NIR bands. The eight vegetation indices were then computed at native resolution and from the resized imagery and compared at native resolution. The spatial correlation between the two ranged from 0.68 to 0.95 across the indices, and the mean absolute error stayed between 0.025 and 0.059. NDRE showed the lowest correlation under

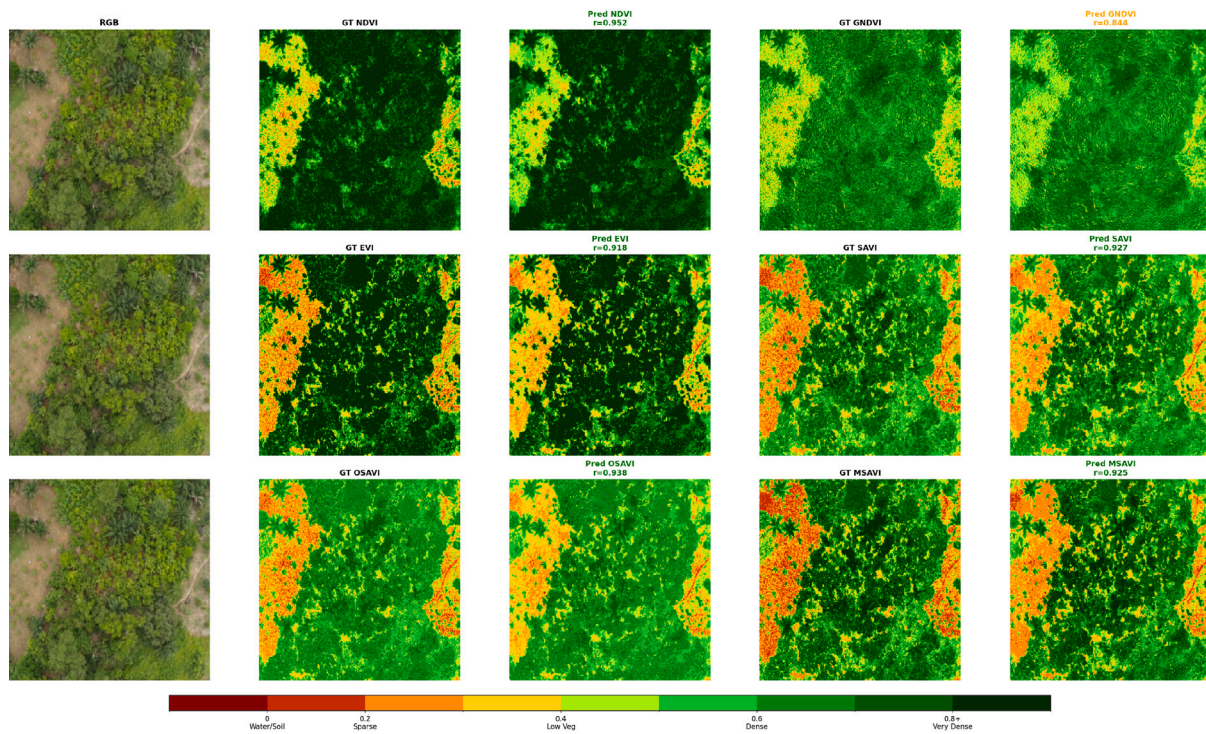


Fig. 9. Vegetation index maps comparing ground truth (GT) and predicted (Pred) values. The leftmost column shows RGB input for spatial reference. Color scale represents VI values: red-orange (<0.2, bare soil), yellow (0.2–0.4, sparse vegetation), and green (>0.5, dense vegetation). Correlation coefficients are displayed for each predicted index.

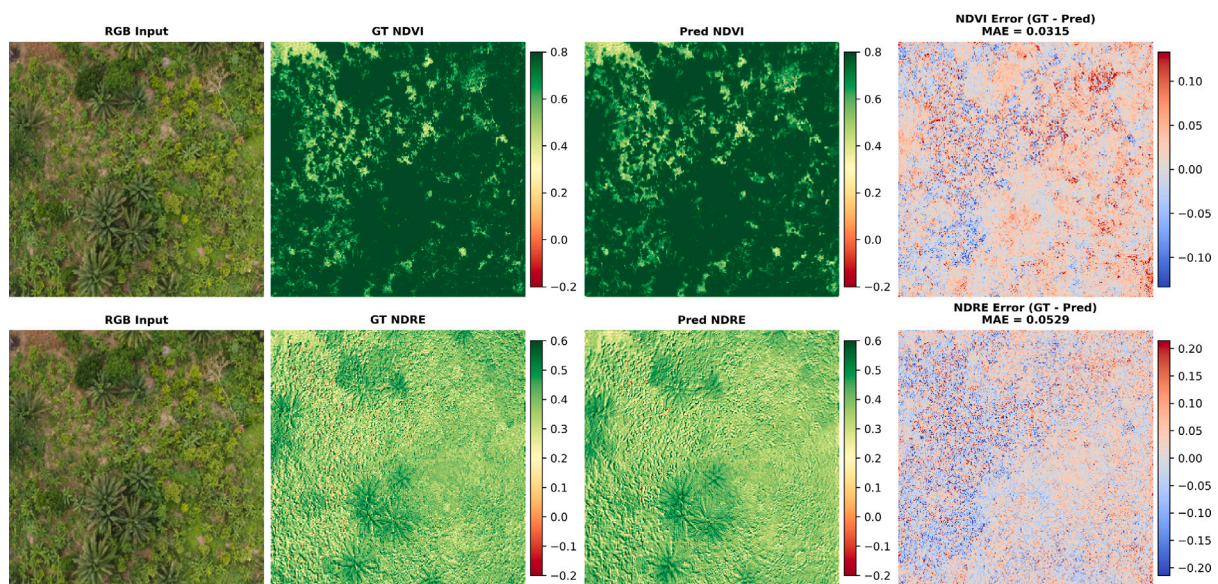


Fig. 10. Spatial comparison of vegetation indices for a representative test sample at the median NDVI correlation (plot B03). The top row shows the RGB input, the ground-truth NDVI, the synthesized NDVI, and the per-pixel NDVI error. The bottom row shows the same layout for NDRE. Each error map reports ground truth minus prediction with the mean absolute error.

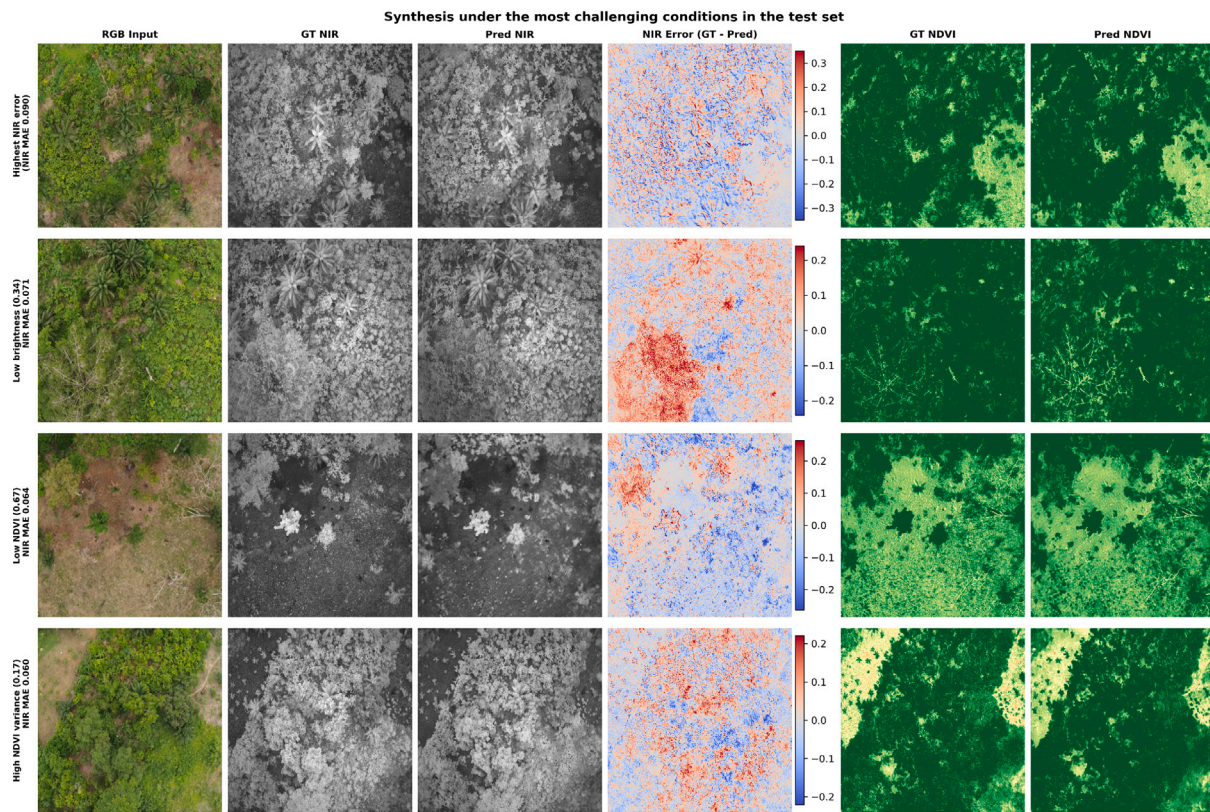


Fig. 11. Synthesis under the most challenging conditions in the test set. Each row shows, from left to right, the RGB input, the ground-truth NIR, the synthesized NIR, the per-pixel NIR error (ground truth minus prediction), the ground-truth NDVI, and the synthesized NDVI. The rows show the test image with the highest NIR error, the lowest-brightness image, the lowest-NDVI image, and the most heterogeneous image. Each case carried NIR error above the test-set mean of 0.059. The two test plots are dense cocoa canopy with NDVI between 0.67 and 0.84, so frames dominated by bare soil or deep shadow are rare in this dataset.

the resize as well, consistent with its dependence on the Near-Infrared to Red Edge contrast. The resize was applied to the ground truth and to the synthesized output through the same operation, so the index comparison in this section was performed at the resolution used throughout the study.

The spatial comparison in Fig. 9 showed consistent spatial patterns between ground truth and predicted maps, with bare soil regions correctly assigned low VI values and dense vegetation canopy regions assigned high values across all indices.

Fig. 11 shows the synthesis for the hardest test cases by NIR error. The largest residuals fell on fine canopy texture and on canopy and shadow boundaries, consistent with the limitations discussed in Section 5.6.

5. Discussion

The experimental findings are interpreted within the broader context of spectral synthesis research and their relevance to sustainable agroforestry monitoring. The analysis covers implications for smallholder access to vegetation indicators, positioning within prior work, band-specific performance patterns, vegetation index implications, model complexity trade-offs, interpretability via Grad-CAM attribution, and limitations of the current method.

Prior RGB-to-spectral synthesis methods include single-band RGB-to-Near-Infrared translation (Aslahishahri et al., 2021) and five-band aerial multispectral reconstruction (Zhang et al., 2022), which share the single-input setting of the present work. The proposed method is a generative adversarial network, and Section 4.2 evaluates it against two adversarial baselines, Pix2Pix and the GAN-based Pix2Next, on the identical dataset and task. The single-band study identified multi-band synthesis as an open direction, which the present five-band method

addresses (Aslahishahri et al., 2021). The five-band study separated the decoder into a visible head and a non-visible head and found that this separation improved reconstruction, which motivates the band-group separation in the proposed Split-Head Decoder (Zhang et al., 2022). The proposed method builds on a transformer backbone through the SwinV2 Vision Foundation Model feature extractor. Attention-driven and physics-informed fusion methods (Xing et al., 2022; Xiao et al., 2021; Gao et al., 2022) instead reconstruct a spectral output from two spectral inputs, so their reconstruction metrics are not on equal terms with single-input synthesis. The proposed method improved mean SSIM from 0.836 to 0.876 over the GAN-based Pix2Next, and Table 2 shows gains over Pix2Pix and TaijiGNN across all evaluation metrics, with the largest gain on the NIR band.

5.1. Implications for sustainable agroforestry monitoring

The vegetation index results carried practical implications for sustainable land management in West Africa. Smallholder farmers who lack access to multispectral sensors represent the primary beneficiary group. The NDVI correlation of $r = 0.91$ indicated that RGB UAVs operated with the proposed method reproduced vegetation indices that agreed with multispectral ground truth on the evaluated dataset (Fawcett et al., 2020). Lower sensor-cost requirements could expand precision agriculture access in comparable cocoa-producing settings, subject to validation under their acquisition conditions (Rashid and Kashani, 2025; Lammoglia et al., 2024).

The proposed method also addressed environmental sustainability monitoring beyond crop productivity. Canopy cover indicators derived from NDVI and GNDVI synthesized bands could serve as proxies for vegetation condition monitoring in agroforestry systems (Rashid and Kashani, 2025), where high-frequency UAV surveys with consumer

equipment could supplement satellite-based monitoring programs. Côte d'Ivoire's cocoa belt, characterized by high deforestation rates (Lammoglia et al., 2024), represented a context where accessible vegetation indicators contributed to both agricultural productivity and land cover sustainability goals.

National sustainability programs targeting deforestation reduction in the cocoa belt could incorporate consumer-grade UAV synthesis as a low-cost complement to satellite monitoring, particularly for plot-level assessment where satellite revisit periods are insufficient. With validation on their own sensors and conditions, extension services supporting smallholder farmers could use the proposed method to derive vegetation indices from RGB UAV imagery.

5.2. Band-specific performance analysis

The performance gradient documented in Section 4.2 reflects underlying physical mechanisms that the correlation analysis (Section 4.1) alone did not fully explain. Blue band performance (SSIM 0.936) exceeded expectations for an atmospherically sensitive wavelength, which is consistent with the 80 m acquisition altitude minimizing atmospheric path length relative to satellite platforms operating at several hundred kilometers (Fawcett et al., 2020; Vermote et al., 2016). Because the RGB input does not measure the Red Edge and NIR wavelengths, the model produced these bands from learned statistical relationships between the RGB channels and the target bands rather than from a physical measurement of non-visible reflectance, so the values reported for these bands are inferred estimates rather than recovered ground-truth reflectance. Red Edge occupied an intermediate position because its reflectance correlates with both Red (through chlorophyll absorption edge) and NIR (through leaf cellular structure onset) (Gitelson et al., 2003), which enabled indirect inference from RGB cues despite the absence of direct channel input. NIR variability (± 0.080) reflected differential reconstruction difficulty across vegetation conditions: dense canopy regions with strong mesophyll scattering produced consistent NIR responses, whereas mixed vegetation and bare soil transitions exhibited greater reconstruction variance (Aslahishahri et al., 2021; Zhang et al., 2022).

5.3. Vegetation index implications

For practical deployment, the NDVI MAE of 0.034 falls below the typical threshold boundaries used in vegetation classification schemes (e.g., 0.2 for bare soil, 0.5 for dense canopy) (Candiago et al., 2015). Systematic offsets of this magnitude are unlikely to shift classification assignments for most pixels, supporting the use of synthesized NDVI for vigor-based monitoring applications. The indices therefore fall into three reliability tiers on the evaluated dataset. The six NIR-Red indices, namely NDVI, ARVI, OSAVI, SAVI, MSAVI, and EVI, were the most reliable, with correlations above 0.87. GNDVI was intermediate at 0.78. NDRE was the least reliable at 0.62, so the indices that depend on the Red Edge to NIR contrast carry the largest synthesis error and warrant the most caution in downstream use. NDRE performance ($r = 0.62$, MAE 0.056) constrains its applicability for early stress detection, where pre-symptomatic chlorophyll changes require tighter ratio fidelity than canopy vigor assessment (Boiarskii and Hasegawa, 2019; Gitelson et al., 2003). Applications requiring NDRE-based stress detection would benefit from dedicated Red Edge–NIR ratio constraints beyond those in the current Spectral Loss formulation, or from hybrid deployment workflows that retain multispectral sensors for stress monitoring while leveraging synthesized bands for complementary indicators. Semantic and instance segmentation of UAV multispectral imagery offer a complementary refinement to the synthesized indices. Semantic segmentation isolates canopy pixels from soil and shadow through pixel-level classification (Tian et al., 2024), and instance segmentation delineates individual tree crowns (Dersch et al., 2023). Both restrict vegetation index computation to canopy pixels and reduce the

soil and shadow contributions that bias the index, which positions segmentation as a downstream step that operates on the synthesized bands rather than as a substitute for the synthesis. Plant-health applications such as disease or stress detection represent potential downstream uses of the synthesized vegetation indices, but would require validation against independent disease, stress, or chlorophyll measurements that this study did not include.

5.4. Model complexity and efficiency trade-off

Architectural modifications that improve reconstruction quality often increase computational cost, and the practical value of the proposed method depends on whether the added cost is proportionate to the performance gain. Table 9 reports parameter count, floating-point operations (FLOPs), and inference latency of SMH-GAN relative to the Pix2Next baseline. Both models shared the same frozen SwinV2-Tiny backbone (27.58 M parameters), so the comparison isolated the generator design.

SMH-GAN contained 110.13 M total parameters compared with 110.38 M for Pix2Next. The parameter-efficient Serial Dual Attention design reduced the cross-attention stack by approximately 22.35 M parameters, and this budget was reallocated to the Spectral-Informed Bottleneck (20.41 M) and the Split-Head Decoder (1.69 M). The net result was a slight parameter reduction rather than an increase.

Computational cost followed a different pattern. SMH-GAN required 194.36 G FLOPs per 256×256 image compared with 82.52 G for Pix2Next, a 2.36 $\times$ increase. Inference latency on an NVIDIA T4 GPU was 142.48 ± 8.38 ms for SMH-GAN versus 67.24 ± 3.68 ms for Pix2Next, a 2.12 $\times$ increase. Three design choices accounted for this overhead: SDA performed spectral then spatial attention in sequence, SIB enforced spectral constraints through dual-path processing with adaptive gating, and the Split-Head Decoder processed three output branches with multi-scale convolutions at kernel sizes 3×3 , 5×5 , and 7×7 . The near-alignment between the FLOPs ratio (2.36 $\times$) and the latency ratio (2.12 $\times$) indicated that the overhead reflected genuine computation rather than suboptimal kernel scheduling.

The target application of offline UAV-based agroforestry monitoring processes imagery in post-flight batch pipelines rather than real-time streams. At 142 ms per image, SMH-GAN synthesized seven images per second on commodity hardware, sufficient for the plot-scale monitoring workflow of the UAV dataset (Lammoglia et al., 2024). The parameter-neutral design also ensured that deployment memory footprint remained comparable to the baseline, easing integration into existing Pix2Next-based pipelines.

5.5. Explainable AI analysis

The ablation results showed that the Split-Head Decoder produced the largest performance gain, but architectural ablation alone did not verify whether each head attended to band-specific RGB cues or whether the gains arose from generic capacity expansion. Gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2020) was applied to each decoder head to examine the spatial attribution patterns for individual band outputs. For each synthesized band, the gradient of the summed band output with respect to activations at the final feature layer of the corresponding head was computed, and the resulting map was overlaid on the input RGB image.

Fig. 12 presents attribution maps for a representative test sample. The visible band maps (Blue, Green, Red) shared the same underlying head but produced distinct spatial distributions. The Blue map highlighted bare soil regions and shadow boundaries, consistent with Rayleigh scattering sensitivity and surface-level reflectance variation at 450 nm. The Green map concentrated on dense canopy regions where active photosynthesis drove reflectance at 560 nm. The Red map emphasized contrast boundaries between canopy and bare soil,

Table 9

Model complexity and inference efficiency. Latency measured on an NVIDIA T4 GPU with batch size 1 and input resolution 256×256 , averaged over 100 runs after 20 warmup iterations.

Model	Total params (M)	Trainable params (M)	FLOPs (G)	Latency (ms)	Throughput (img/s)
Pix2Next (Jin et al., 2025)	110.38	82.80	82.52	67.24 ± 3.68	14.87
SMH-GAN	110.13	82.56	194.36	142.48 ± 8.38	7.02
Ratio (SMH-GAN/Pix2Next)	0.998×	0.997×	2.36×	2.12×	0.47×

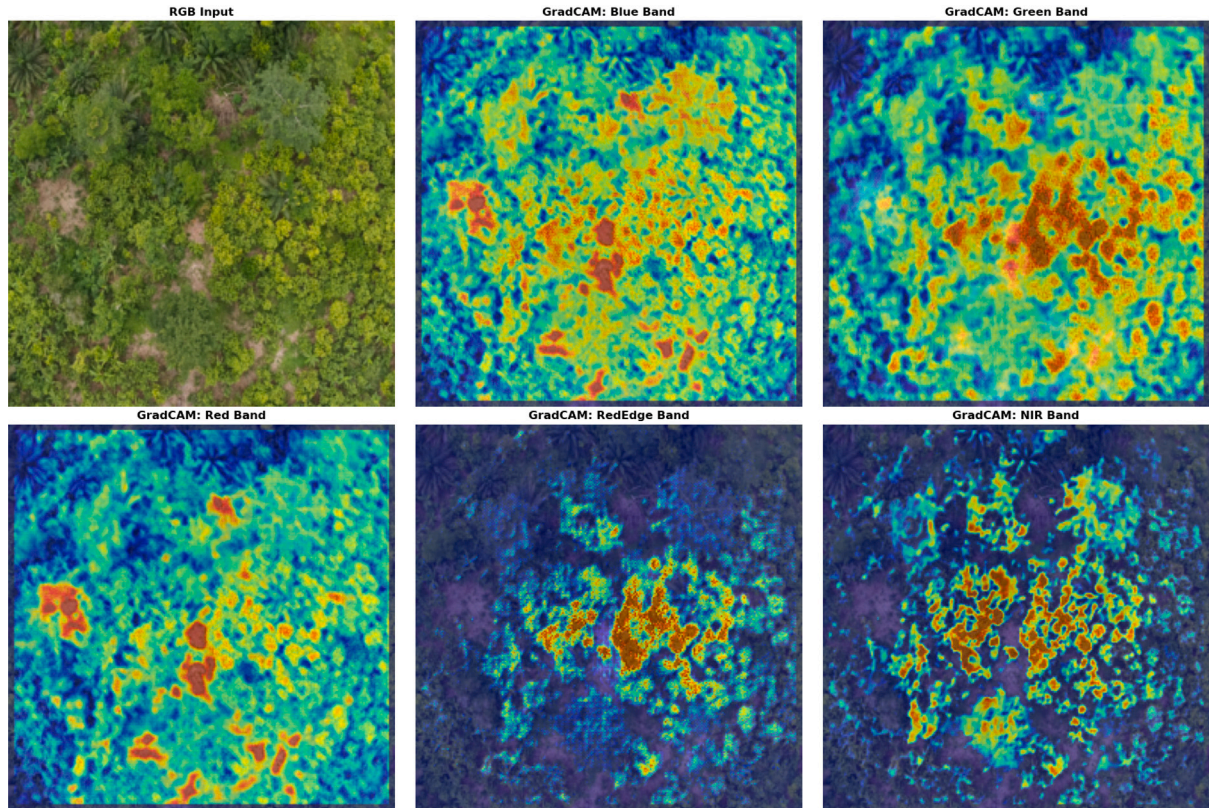


Fig. 12. Grad-CAM attribution maps showing RGB regions that influenced each synthesized band. The top-left panel shows the input RGB image from plot B03. The remaining panels show the attribution map for each synthesized band, overlaid on the RGB input. Warmer colors indicate higher influence on the corresponding band output.

reflecting chlorophyll absorption in vegetation and higher reflectance on exposed soil surfaces (Gitelson et al., 2003).

The Red Edge and NIR maps exhibited distinct attribution patterns that diverged from both each other and from the visible bands. The Red Edge map highlighted transition regions at canopy boundaries and areas of mixed vegetation density, consistent with the 730 nm wavelength position between chlorophyll absorption and internal leaf scattering regimes (Gitelson et al., 2003). The NIR map showed the most localized attribution, with high-intensity clusters concentrated on dense canopy pixels that drove the mesophyll scattering response at 840 nm.

The distinct attribution patterns provide empirical support for the Split-Head Decoder design rationale. The three heads developed band-group-specific spatial preferences during training, and the RGB cues that drove each head's output corresponded to the physical origins of reflectance in the target wavelength range. The interpretability evidence complements the quantitative ablation results by indicating that the performance gains arose from band-appropriate feature selection rather than from generic network capacity expansion.

5.6. Limitations

Seven limitations constrain the interpretation and generalization of the current findings. First, NIR band reconstruction ($SSIM = 0.725$) remains below the visible band range (0.913–0.962) despite the largest absolute gain over the Pix2Next baseline. The inherent difficulty of inferring non-visible spectral information from three-channel RGB input establishes an upper bound on reconstruction accuracy that architectural modifications alone cannot overcome, consistent with observations in prior RGB-to-NIR synthesis work (Aslahishahri et al., 2021; Zhang et al., 2022). The analysis also operated at a reduced spatial resolution of 256×256 pixels, which lowered the fine spatial detail of the Red Edge and NIR bands. The indices built on the Near-Infrared to Red Edge contrast showed the lowest agreement under the resize. NDRE showed the lowest agreement in Table 8. Sub-canopy variation in the vegetation indices was not resolved at this resolution.

Second, although the plot-level partition removed train-to-test leakage, the held-out evaluation used only two plots (B03 and B10) from

a single region and acquisition campaign (Lammoglia et al., 2024). Generalization to other crop systems, canopy structures, illumination conditions, or flight altitudes has not been assessed, and performance under different phenological stages, sensor platforms, or atmospheric conditions remains an open question. The reported performance therefore reflects only the conditions represented in this dataset, and broader operational deployment would require validation across additional regions, seasons, sensors, and acquisition conditions. In particular, the RGB input in this study came from the P4M RGB camera. Application of the learned mapping to RGB cameras with different spectral response functions would require additional validation.

Third, shadow regions produced reconstruction artifacts because the RGB input contained insufficient spectral information to distinguish shadowed vegetation from non-vegetation dark surfaces. This failure mode affected a small fraction of test samples but indicated a structural limitation of RGB-based synthesis under heterogeneous illumination.

Fourth, the computational overhead reported in Section 5.5 restricts deployment scenarios that require high-throughput processing. The $2.12\times$ latency increase relative to Pix2Next is acceptable for offline batch pipelines but would constrain real-time applications such as on-board UAV analysis or large-scale regional surveys with time-constrained revisit schedules.

Fifth, the radiometric calibration applied a global factor per band derived from training-set statistics and fixed nominal reflectance values. This factor did not model illumination variation, atmospheric effects, bidirectional reflectance, or flight-to-flight differences, and recalibration would be required for imagery acquired under other radiometric conditions.

Sixth, the proposed method requires paired multispectral ground truth to train, so the reduction in sensor requirements applies at inference rather than at model development.

Seventh, the experimental design of the ablation and the backbone comparison had two constraints. Each configuration was trained once with a fixed random seed, so the variance of the reported differences across repeated training runs was not characterized, and the small differences between close backbones were not separated from stochastic training variability through statistical testing. The ablation also followed an incremental design rather than a full factorial design over the four components. The attention stage had isolated configurations that allowed a direct comparison, and the Spectral-Informed Bottleneck and the spectral loss each admitted a clean single-component removal from the full model, reported in Section 4.4. The Split-Head Decoder did not admit a clean single-component removal, because a single-head replacement changes the output capacity together with the band separation, so its contribution was assessed through the incremental ablation. A factorial evaluation with repeated seeds would quantify the remaining interactions and their variability. The per-band weights of the spatial gradient loss were likewise fixed from the correlation structure rather than tuned, and a systematic exploration of this weight vector remains outside the scope of the present study.

6. Conclusions

The proposed work introduced modifications to the Pix2Next architecture for synthesizing five-band multispectral images from RGB UAV input. Four components were introduced: Serial Dual Attention for sequential spectral-spatial feature processing, a Spectral-Informed Bottleneck for domain-informed latent representation, a Split-Head Decoder for separating visible and non-visible band reconstruction pathways, and a Spectral Consistency Loss for constraining inter-band relationships. SMH-GAN outperformed Pix2Pix, TaijiGNN, and Pix2Next across all five spectral bands, with the NIR band showing the largest improvement ($+0.127$ SSIM over the Pix2Next baseline).

Vegetation index validation indicated that the synthesized bands preserved the spectral relationships needed for vegetation index computation on the evaluated dataset. NDVI achieved $r = 0.91$ correlation

with ground truth, and six NIR-Red based indices exceeded $r = 0.87$. NDRE correlation remained lower ($r = 0.62$), consistent with error compounding from the two lowest-performing bands (Red Edge: 0.845 SSIM, NIR: 0.725 SSIM).

Three principal limitations qualify these findings. NIR band performance (0.725 SSIM) remained below visible band levels, reflecting the inherent difficulty of inferring non-visible spectral information from three-channel RGB input. Evaluation was limited to one cocoa agroforestry dataset, and generalization to other crop systems, geographic regions, or flight conditions has not been assessed. The proposed method also incurred approximately $2.1\times$ inference latency relative to the Pix2Next baseline, which restricts deployment in real-time scenarios.

The results carried practical implications for sustainable land management in cocoa-producing regions. On the evaluated cocoa agroforestry dataset, the synthesized bands yielded vegetation indices consistent with multispectral ground truth. Validation beyond the evaluated dataset would be needed before the method could support operational monitoring for smallholder farmers in West Africa. The NDVI and GNDVI indicators derived from synthesized bands could complement satellite-based deforestation monitoring programs in Côte d'Ivoire's cocoa belt, contributing to land cover sustainability objectives alongside crop productivity goals.

Future work will investigate transfer learning approaches to enable generalization across crop types without full retraining. Uncertainty quantification methods will be examined to identify image regions where synthesis confidence is low, enabling selective use of synthesized versus actual measurements. Integration with downstream agricultural tasks through end-to-end training will be explored to optimize synthesis for objectives such as yield prediction or disease detection.

CRedit authorship contribution statement

Dalila Anwar Dini Rasyidah: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Riyanarto Sarno:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Kelly Rossa Sungkono:** Writing – review & editing, Validation, Methodology. **Agus Tri Haryono:** Writing – review & editing, Validation. **Rizqy Ahsana Putri:** Writing – review & editing, Validation. **Sholiq:** Writing – review & editing, Validation. **Dhimas Wiratmoko:** Resources, Investigation, Data curation. **Siti Halimah Larekeng:** Validation, Resources, Investigation. **Arif Abdullah Sagan:** Writing – review & editing, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was funded by the Institut Teknologi Sepuluh Nopember through the IRN Collaborative Research Type B (Kemitraan IRN Tipe B) Scheme, year 2026 (Grant No. 2657/PKS/ITS/2026).

Appendix A. Supplementary data

The UAV dataset used in this study is publicly available in the CIRAD Dataverse repository at <https://doi.org/10.18167/DVN1/MK2ZRG>.

Data availability

Data will be made available on request.

References

- Ahmed, M.T., Monjur, O., Khaliduzzaman, A., Kamruzzaman, M., 2025. A comprehensive review of deep learning-based hyperspectral image reconstruction for agri-food quality appraisal. *Artif. Intell. Rev.* 58 (4), 96. <http://dx.doi.org/10.1007/s10462-024-11090-w>.
- Antwi-Agyei, P., Wiafe, E.A., Amanor, K., Baffour-Ata, F., Codjoe, S.N.A., 2021. Determinants of choice of climate change adaptation practices by smallholder pineapple farmers in the semi-deciduous forest zone of Ghana. *Environ. Sustain. Indic.* 12, 100140. <http://dx.doi.org/10.1016/j.indic.2021.100140>.
- Aslahishahri, M., Stanley, K.G., Duddu, H., Shirliff, S., Vail, S., Bett, K., Pozniak, C., Stavness, I., 2021. From RGB to NIR: Predicting of near infrared reflectance from visible spectrum aerial images of crops. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 1312–1322.
- Boiarskii, B., Hasegawa, H., 2019. Comparison of NDVI and NDRE indices to detect differences in vegetation and chlorophyll content. *J. Mech. Continuum Math. Sci.* 4, 20–29. <http://dx.doi.org/10.26782/jmcs.spl.4/2019.11.00003>.
- Candiago, S., Remondino, F., De Giglio, M., Dubbini, M., Gattelli, M., 2015. Evaluating multispectral images and vegetation indices for precision farming applications from UAV images. *Remote. Sens.* 7 (4), 4026–4047. <http://dx.doi.org/10.3390/rs70404026>.
- De Souza, V.L.T., Marques, B.A.D., Batagelo, H.C., Gois, J.P., 2023. A review on generative adversarial networks for image generation. *Comput. Graph.*
- Dersch, S., Schoettl, A., Krzystek, P., Heurich, M., 2023. Towards complete tree crown delineation by instance segmentation with mask R-CNN and DETR using UAV-based multispectral imagery and lidar data. *ISPRS Open J. Photogramm. Remote. Sens.* 8, 100037. <http://dx.doi.org/10.1016/j.ojphoto.2023.100037>.
- Fajar, A., Sarno, R., Fatchah, C., Susilo, R.I., Pangestu, G., 2023. Cyclical learning rate optimization on deep learning model for brain tumor segmentation. *IEEE Access* 11, 119802–119810. <http://dx.doi.org/10.1109/ACCESS.2023.3326475>.
- Farokhah, L., Sarno, R., Fatchah, C., 2023. Cross-subject channel selection using modified relief and simplified CNN-based deep learning for EEG-based emotion recognition. *IEEE Access* 11, 110136–110150. <http://dx.doi.org/10.1109/ACCESS.2023.3322294>.
- Fawcett, D., Panigada, C., Tagliabue, G., Boschetti, M., Celesti, M., Evdokimov, A., Biriukova, K., Colombo, R., Miglietta, F., Rascher, U., et al., 2020. Multi-scale evaluation of drone-based multispectral surface reflectance and vegetation indices in operational conditions. *Remote. Sens.* 12 (3), 514. <http://dx.doi.org/10.3390/rs12030514>.
- Gao, H., Li, S., Dian, R., 2022. Hyperspectral and multispectral image fusion via self-supervised loss and separable loss. *IEEE Trans. Geosci. Remote Sens.* 60, 1–12. <http://dx.doi.org/10.1109/TGRS.2022.3204769>.
- Gitelson, A.A., Gritz, Y., Merzlyak, M.N., 2003. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *J. Plant Physiol.* 160 (3), 271–282.
- Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., Zhang, S.-H., Martin, R.R., Cheng, M.-M., Hu, S.-M., 2022. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* 8 (3), 331–368.
- Haryono, A.T., Sarno, R., Sungkono, K.R., 2023. Transformer-gated recurrent unit method for predicting stock price based on news sentiments and technical indicators. *IEEE Access* 11, 77132–77146.
- Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1125–1134.
- Jin, Y., Park, I., Song, H., Ju, H., Nalcakan, Y., Kim, S., 2025. Pix2Next: Leveraging vision foundation models for RGB to NIR image translation. *Technologies* 13 (4), 154. <http://dx.doi.org/10.3390/technologies13040154>.
- Jin, X., Xie, Y., Wei, X.-S., Zhao, B.-R., Chen, Z.-M., Tan, X., 2022. Delving deep into spatial pooling for squeeze-and-excitation networks. *Pattern Recognit.* 121, 108159. <http://dx.doi.org/10.1016/j.patcog.2021.108159>.
- Kouassi, J.-L., Diby, L., Konan, D., Kouassi, A., Bene, Y., Kouamé, C., 2023. Drivers of cocoa agroforestry adoption by smallholder farmers around the Taï national park in southwestern Côte d'Ivoire. *Sci. Rep.* 13 (1), 14309. <http://dx.doi.org/10.1038/s41598-023-41593-5>.
- Lammoglia, S.-K., Akpa, Y.L., Danumah, J.H., Brou, Y.L.A., Kassi, J.N., 2024. High-resolution multispectral and RGB dataset from UAV surveys of ten cocoa agroforestry typologies in Côte d'Ivoire. *Data Brief* 55, 110664. <http://dx.doi.org/10.1016/j.dib.2024.110664>.
- Liu, X., Gherbi, A., Li, W., Wei, Z., Cheriet, M., 2021. TaijiGNN: A new cycle-consistent generative neural network for high-quality bidirectional transformation between RGB and multispectral domains. *Sensors* 21 (16), 5394. <http://dx.doi.org/10.3390/s21165394>.
- Liu, L., Li, W., Shi, Z., Zou, Z., 2022. Physics-informed hyperspectral remote sensing image synthesis with deep conditional generative adversarial networks. *IEEE Trans. Geosci. Remote Sens.* 60, 1–15. <http://dx.doi.org/10.1109/TGRS.2022.3173532>.
- Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S., 2017. Least squares generative adversarial networks. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 2794–2802.
- Ndoumbe-Nkeng, M., Cilas, C., Nyemb, E., Nyassé, S., Bieysse, D., Flori, A., Sache, I., 2004. Impact of removing diseased pods on cocoa black pod caused by phytophthora megakarya and on cocoa production in Cameroon. *Crop. Prot.* 23 (5), 415–424. <http://dx.doi.org/10.1016/j.cropro.2003.09.010>.
- Okyer, F.G., Cudjoe, D.K., Virlet, N., Castle, M., Riche, A.B., Greche, L., Mohareb, F., Simms, D., Mhada, M., Hawkesford, M.J., 2024. Hyperspectral imaging for phenotyping plant drought stress and nitrogen interactions using multivariate modeling and machine learning techniques in wheat. *Remote. Sens.* 16 (18), 3446. <http://dx.doi.org/10.3390/rs16183446>.
- Putri, R.A., Sarno, R., Utomo, W.P., Taufany, F., Sungkono, K.R., Amri, T.C., Kamilah, A., Handayani, R., Lee, S.-S., Tjoa, A.M., et al., 2026. Cross-attention feature fusion network for robust estimation of Cd2+ and Pb2+ in water samples using cyclic voltammetry. *Chem. Eng. J. Adv.* 26, 101065. <http://dx.doi.org/10.1016/j.cej.2026.101065>.
- Rahmawati, D., Sarno, R., Fatchah, C., 2022. Enhancing marching cubes algorithms for 3D corpus callosum reconstruction from 2D MRI brain images segmentation. *Appl. Soft Comput.*
- Rashid, I., Kashani, S.D.R., 2025. Forest dynamics and above-ground forest biomass changes utilizing google earth engine, machine learning, and field-based observations in the Kashmir Himalaya, India. *Environ. Sustain. Indic.* 27, 100759. <http://dx.doi.org/10.1016/j.indic.2025.100759>.
- Rasyidah, D.A.D., Sarno, R., Sungkono, K.R., Taufany, F., Sagan, A.A., Al Fauzi, A., Hadi, S., 2026. Serial dual attention for VFM-guided RGB-to-multispectral synthesis. In: *2026 International Seminar on Intelligent Business and Edge-Computing Research*. ISIBER, IEEE, pp. 452–457. <http://dx.doi.org/10.1109/ISIBER68248.2026.11469817>.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 234–241. http://dx.doi.org/10.1007/978-3-319-24574-4_28.
- Sahoo, S., Singha, C., Govind, A., Moghimi, A., 2025. Review of climate-resilient agriculture for ensuring food security: Sustainability opportunities and challenges of India. *Environ. Sustain. Indic.* 25, 100544. <http://dx.doi.org/10.1016/j.indic.2024.100544>.
- Salehin, S.M.U., Poudyal, C., Rajan, N., Bagavathiannan, M., 2025. Cover crop types influence biomass estimation using unmanned aerial vehicle-mounted multispectral sensors. *Remote. Sens.* 17 (8), 1471. <http://dx.doi.org/10.3390/rs17081471>.
- Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans. 29,
- Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2020. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* 128 (2), 336–359. <http://dx.doi.org/10.1007/s11263-019-01228-7>.
- Sishodia, R.P., Ray, R.L., Singh, S.K., 2020. Applications of remote sensing in precision agriculture: A review. *Remote. Sens.* 12 (19), 3136. <http://dx.doi.org/10.3390/rs12193136>.
- Standley, T., Zamir, A., Chen, D., Guibas, L., Malik, J., Savarese, S., 2020. Which tasks should be learned together in multi-task learning? In: *International Conference on Machine Learning*. PMLR, pp. 9120–9132.
- Tian, B., Yu, H., Zhang, S., Wang, X., Yang, L., Li, J., Cui, W., Wang, Z., Lu, L., Lan, Y., et al., 2024. Inversion of cotton soil and plant analytical development based on unmanned aerial vehicle multispectral imagery and mixed pixel decomposition. *Agriculture* 14 (9), 1452. <http://dx.doi.org/10.3390/agriculture14091452>.
- Vermote, E., Justice, C., Claverie, M., Franch, B., 2016. Preliminary analysis of the performance of the landsat 8/OLI land surface reflectance product. *Remote Sens. Environ.* 185, 46–56. <http://dx.doi.org/10.1016/j.rse.2016.04.008>.
- Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13 (4), 600–612. <http://dx.doi.org/10.1109/TIP.2003.819861>.
- Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In: *Proceedings of the European Conference on Computer Vision*. ECCV, pp. 3–19. http://dx.doi.org/10.1007/978-3-030-01234-2_1.
- Xiao, J., Li, J., Yuan, Q., Jiang, M., Zhang, L., 2021. Physics-based GAN with iterative refinement unit for hyperspectral and multispectral image fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* 14, 6827–6841. <http://dx.doi.org/10.1109/JSTARS.2021.3075727>.
- Xing, Y., Yang, S., Zhang, Y., Zhang, Y., 2022. Learning spectral cues for multispectral and panchromatic image fusion. *IEEE Trans. Image Process.* 31, 6964–6975. <http://dx.doi.org/10.1109/TIP.2022.3215906>.
- Yang, G., Liu, J., Zhao, C., Li, Z., Huang, Y., Yu, H., Xu, B., Yang, X., Zhu, D., Zhang, X., et al., 2017. Unmanned aerial vehicle remote sensing for field-based crop phenotyping: Current status and perspectives. *Front. Plant Sci.* 8, 1111. <http://dx.doi.org/10.3389/fpls.2017.01111>.
- Yuan, Y., Wang, X., Shi, M., Wang, P., 2022. Performance comparison of RGB and multispectral vegetation indices based on machine learning for estimating hopea hainanensis SPAD values under different shade conditions. *Front. Plant Sci.* 13, 928953. <http://dx.doi.org/10.3389/fpls.2022.928953>.
- Zeng, T., Diao, C., Lu, D., 2021. U-net-based multispectral image generation from an rgb image. *IEEE Access* 9, 43387–43396. <http://dx.doi.org/10.1109/ACCESS.2021.3066472>.

- Zhang, Y., Yang, W., Zhang, W., Yu, J., Zhang, J., Yang, Y., Lu, Y., Tang, W., 2022. Two-step resup&down generative adversarial network to reconstruct multispectral image from aerial RGB image. *Comput. Electron. Agric.* 192, 106617. <http://dx.doi.org/10.1016/j.compag.2021.106617>.
- Zheng, Y., Li, J., Li, Y., Guo, J., Wu, X., Shi, Y., Chanussot, J., 2021. Edge-conditioned feature transform network for hyperspectral and multispectral image fusion. *IEEE Trans. Geosci. Remote Sens.* 60, 1–15. <http://dx.doi.org/10.1109/TGRS.2021.3108122>.
- Zhou, H., Yang, J., Lou, W., Sheng, L., Li, D., Hu, H., 2023. Improving grain yield prediction through fusion of multi-temporal spectral features and agronomic trait parameters derived from UAV imagery. *Front. Plant Sci.* 14, 1217448. <http://dx.doi.org/10.3389/fpls.2023.1217448>.
- Zhu, M., Liu, J., Wang, F., 2024. Blind deblurring method for casearth multispectral images based on inter-band gradient similarity prior. *Sensors* 24 (19), 6259. <http://dx.doi.org/10.3390/s24196259>.
- Zhu, Q., Zhang, M., Chen, Y., Zheng, G., Luo, J., 2024. Spectral correlation-based fusion network for hyperspectral image super-resolution. *IEEE Trans. Geosci. Remote Sens.* 63, 1–14. <http://dx.doi.org/10.1109/TGRS.2024.3423422>.