A Physics-Inspired Lightweight Multimodal Network for Robust Winter Wheat LAI Estimation under Spectral Saturation Conditions

Shihao Mei
Fuyang Normal University
Yusheng Cheng
Fuyang Normal University
Chundu Wu
Jiangsu University
Lulu Zhang
zhangll_py@163.com

Fuyang Normal University
Xiaowen Wang
Jiangsu University

Research Article

Keywords: winter wheat, leaf area index, lightweight neural network, multimodal fusion, precision agriculture

Posted Date: September 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10626981/v1

Additional Declarations: No competing interests reported.

A Physics-Inspired Lightweight Multimodal Network for Robust Winter Wheat LAI Estimation under Spectral Saturation Conditions

Shihao Mei[1], Yusheng Cheng[1,2], Chundu Wu[3,4,5], Lulu Zhang[1*], Xiaowen Wang[3,6]

[1]School of Artificial Intelligence and Computer Science, Anqing Normal University, Anqing, 246133, China.
[2]The University Key Laboratory of Intelligent Perception and Computing of Anhui Province, Anqing Normal University, Anqing, 246133, China.
[3]School of Agricultural Engineering, Jiangsu University, Zhenjiang, 212013, China.
[4]Key Laboratory for Theory and Technology of Intelligent Agricultural Machinery and Equipment, Jiangsu University, Zhenjiang, 212013, China.
[5]Jiangsu Province and Education Ministry Cosponsored Synergistic Innovation Centre of Modern Agricultural Equipment, Jiangsu University, Zhenjiang, 212013, China.
[6]Key Laboratory of Modern Agricultural Equipment and Technology (Jiangsu University), Ministry of Education, Zhenjiang, 212013, China.

*Corresponding author(s). E-mail(s): zhangll_py@163.com;

Abstract

Context: Accurate leaf area index (LAI) estimation is essential for UAV-based winter wheat growth monitoring, but optical saturation effects under dense canopies remain a major limitation.
Aims: This study aims to develop a lightweight dual-stream network (HFI-Net) that improves LAI estimation under high-LAI conditions where saturation effects commonly occur in conventional vegetation indices for efficient UAV-based LAI mapping.

Methods: A field dataset containing 637 paired UAV image patches and ground LAI measurements was collected from 21 winter wheat cultivars across seven phenological stages. HFI-Net was proposed, integrating RGB texture and vegetation-index (VI) features derived from multispectral imagery through attention-guided star-shaped multiplicative feature interactions, and five-fold cross-validation with data augmentation was applied.
Key Results: HFI-Net achieved a coefficient of determination ($\boldsymbol{R^2}$) of 0.9023 and an RMSE of 0.4822 on the test set. Compared with ResNet50, the proposed model reduced parameters by 98.3% to only 0.40 M, while providing improved prediction performance under saturation-prone high-LAI conditions ($> \mathbf{4.0}$).
Conclusion: HFI-Net enables reliable winter wheat growth monitoring throughout the growing season with reduced performance degradation under high-LAI conditions and low computational cost.
Implications and Impacts: These results indicate the potential of HFI-Net for efficient UAV-based LAI mapping and precision agriculture applications.

Keywords: winter wheat, leaf area index, lightweight neural network, multimodal fusion, precision agriculture

Highlights

- HFI-Net improves LAI estimation under dense canopy conditions by integrating complementary structural and spectral cues.
- A dual-stream architecture effectively integrates RGB texture information and vegetation-index features.
- Star-shaped interactions provide physics-inspired nonlinear feature interactions for multimodal fusion.
- The model achieves R^2=0.9023, maintaining reliable estimation under saturation-prone high-LAI conditions.
- The 0.40M lightweight design provides potential for UAV-based edge applications.

Impact: This study developed a lightweight dual-stream network that effectively mitigates the impact of saturation effects on LAI estimation in dense canopies, aligning with the scope of Precision Agriculture by providing an efficient approach for UAV-based LAI mapping for precision agriculture applications.

1 Introduction

Leaf area index (LAI) is a key biophysical parameter characterizing vegetation canopy development and radiation interception capacity. Its spatiotemporal variations are closely associated with land surface energy exchange and ecosystem productivity. For major grain crops such as winter wheat, LAI not only reflects population growth status but also serves as an important monitoring indicator for precision irrigation, variable-rate fertilization (particularly nitrogen diagnosis), and yield prediction. Conventional methods for LAI acquisition, primarily relying on manual sampling or ground-based instrumentation, suffer from low efficiency and poor timeliness, limiting their applicability to large-scale dynamic agricultural monitoring. In recent years, low-altitude unmanned aerial vehicle (UAV)-based remote sensing platforms, leveraging their operational flexibility and multi-sensor integration capabilities, have emerged as a mainstream approach for field-scale phenotyping information acquisition [1, 41]. Therefore, developing computationally efficient and physics-inspired UAV-based LAI estimation models is important for practical digital agriculture systems.

However, during the middle and late growth stages of winter wheat, from jointing to heading, rapid biomass accumulation often forms highly dense canopies. Under this condition, optical remote sensing suffers from spectral saturation effects [35]. Many early studies relied on vegetation indices, such as NDVI and EVI, to construct empirical regression models. When LAI reaches moderate-to-high levels, typically around 3–4, depending on crop structure and sensor configuration, red-band absorption tends to approach saturation. This often leads to reduced estimation accuracy under dense canopy conditions. To alleviate saturation, some studies introduced red-edge bands or three-dimensional radiative transfer models such as PROSAIL [8, 17]. Yet physical inversion models depend heavily on prior parameters that are difficult to obtain, such as leaf angle distribution. They are also computationally expensive and cannot easily support large-scale real-time monitoring [13].

With the development of artificial intelligence, data-driven deep learning methods have been widely introduced into agricultural remote sensing. To improve LAI estimation under saturation-prone high-LAI conditions, multimodal fusion has become an important research direction. Some studies have fused canopy texture from RGB images with vegetation-index (VI) information to improve feature representation under complex canopy structures [32, 38, 40]. Although these studies have achieved promising results, several limitations remain in complex agricultural scenarios: First, most feature-level fusion methods still use simple concatenation or addition. This simple stacking approach does not explicitly model potential nonlinear interactions between structural and spectral features. In this context, canopy structural information captured by RGB imagery can provide complementary information to multispectral observations. Existing fusion networks often improve feature representation but rarely incorporate explicit nonlinear interactions between canopy structural and spectral features into the fusion process. Ignoring this relationship may limit effective feature complementarity between heterogeneous data sources [30]. Second, deep models such as ResNet and Transformer can improve estimation accuracy, but their large parameter counts and high computational cost do not fit resource-constrained edge devices in agricultural IoT systems, such as onboard UAV chips [36].

To address these issues, this study proposes HFI-Net, a lightweight two-stream feature interaction network for winter wheat LAI estimation. HFI-Net integrates RGB texture features and vegetation-index information through a hybrid feature interaction mechanism, improving estimation accuracy in dense canopy conditions while maintaining computational efficiency. The main contributions are as follows: (i) A lightweight dual-stream encoder with depthwise separable convolutions is developed to reduce model complexity while preserving complementary spatial and spectral representations. (ii) A hybrid feature interaction module based on star-shaped multiplicative interaction is introduced to model nonlinear interactions between canopy structural and spectral features, thereby improving LAI estimation under saturation-prone conditions. (iii) Five-fold cross-validation, ablation experiments, and high-LAI evaluation are conducted to comprehensively assess the effectiveness and efficiency of the proposed method.

2 Materials and Methods

2.1 Experimental Design

The study was conducted at the Runguo Agricultural Ecology Base in Zhenjiang, Jiangsu Province, China (32°8′15.54″N, 119°43′38.19″E). The site is located in the middle and lower Yangtze River plain and belongs to the subtropical monsoon climate zone. Its fertile soil is suitable for winter wheat production. The experiment covered seven major phenological stages during the 2023–2024 winter wheat growing season. Seven key phenological stages were monitored, including regreening, jointing, booting, heading, pre-anthesis, post-anthesis, and grain filling.

The total experimental area was about 23,700 m^2. Rice-wheat rotation was the main planting system, and the soil type was yellow-brown loam under rainfed conditions. The site included 21 wheat cultivars, mainly from the Yangmai and Zhenmai series, and all plots were managed under a small-plot planting design. Sampling was conducted from November 2023 to June 2024. A total of 91 sampling points were randomly distributed across the study area. For the first 20 cultivars, four sampling points were selected for each cultivar. For the remaining cultivar, 11 sampling points were selected because of its larger planting area and plot availability.

A synchronized air-ground observation scheme was adopted. On the same day that UAV flights acquired remote sensing imagery, usually between 10: 00 and 14: 00 under stable illumination, the ground survey team measured LAI within the corresponding plots [9].

By combining continuous observations across the seven key stages from regreening to grain filling, the study finally constructed a multimodal spatiotemporal dataset of winter wheat. The dataset contained 637 independent samples, corresponding to 91 spatial sampling points across seven phenological stages.

2.2 Data Acquisition

2.2.1 UAV Remote Sensing Data Acquisition

A DJI P4M UAV (DJI, Shenzhen, China) was used as the data acquisition platform. It integrates one high-resolution RGB camera and five multispectral cameras, which enabled synchronous collection of RGB and multispectral imagery. The central wavelengths of the five multispectral bands were blue (450 nm ± 6 nm), green (560 nm ± 6 nm), red (650 nm ± 6 nm), red edge (730 nm ± 6 nm), and near infrared (840 nm ± 26 nm). Detailed parameters of the DJI P4M and its sensors are listed in Supplementary Table S1. UAV image acquisition was conducted at each sampling date between 11: 00 and 14: 00 under clear weather and low wind speed. The flight speed was 5.2 m/s, and the camera was triggered every 2 s in time-interval mode. The flight altitude

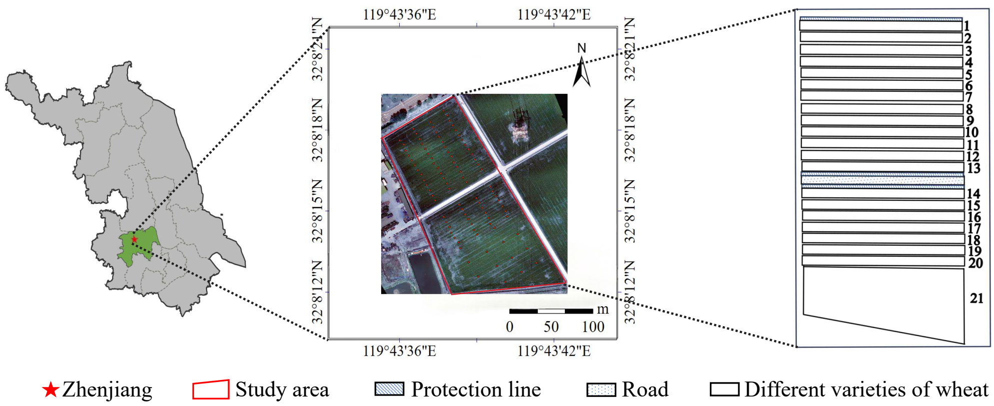

Fig. 1: Overview of the research area.

was set to 100 m, resulting in a ground sampling distance (GSD) of approximately 53 mm/pixel for both RGB and multispectral images. To obtain full coverage of the study area, the forward and side overlaps were set to 85% and 80%, respectively. To improve geometric correction accuracy, 10 ground control points (GCPs) were deployed, and reflectance calibration was performed with a reflectance gray panel before each flight.

RGB imagery provided high spatial resolution information on canopy color and texture. It captured crop cover, canopy structure characteristics, lodging status, and soil background texture.

Multispectral imagery provided spectral reflectance information from five narrow bands. The red-edge band is highly sensitive to subtle chlorophyll variation, while the near-infrared band strongly reflects cell structure and biomass accumulation. Together, they form the core data source for LAI estimation in dense canopies.

2.2.2 Remote Sensing Data Preprocessing

To reduce computational redundancy and noise caused by the small field of view and high overlap of raw UAV images, multispectral imagery was preprocessed with Pix4D Mapper (Pix4D, Lausanne, Switzerland). The RGB and multispectral orthomosaics were co-registered to the same coordinate system, and image patches were extracted using identical plot boundaries. First, low-quality images were removed, and feature matching was performed to generate a sparse point cloud for initial mosaicking with UAV GPS/IMU data. Second, GCPs were imported for geometric refinement, and accumulated errors were reduced by optimizing exterior orientation parameters. Third, a digital surface model (DSM) was generated from the dense point cloud and used for orthorectification, which produced a complete digital orthomosaic (DOM). Finally, standard gray panels with known reflectance values of 12% and 36% were used for radiometric calibration before and after field flights. Radiometric calibration was performed using panel-based reflectance conversion in Pix4D Mapper (Pix4D, Lausanne, Switzerland), after which each band was exported as surface reflectance. This converted digital numbers (DNs) into true surface reflectance and provided reliable input for subsequent vegetation-index construction.

2.2.3 Ground-Based LAI Measurement

Leaf area index (LAI) is a key eco-physiological parameter for evaluating canopy structure and function. It directly affects photosynthesis, transpiration, and yield formation. In this study, winter wheat LAI was measured with a SunScan canopy analysis system produced by Delta-T Devices, UK. The instrument estimates LAI from the ratio between transmitted canopy photon flux density and incident photon flux density above the canopy. It offers fast, nondestructive, and high-accuracy measurements [24, 28].

To maintain consistent measurement conditions, all ground measurements were conducted between 11: 00 and 14: 00 under sunny, windless conditions, thereby reducing the influence of illumination variation. For each plot, LAI was measured at five representative positions, including the center and four corners, to characterize canopy conditions. The average of these five measurements was used as the plot-level LAI label for model training.

2.3 Data Processing

To fully exploit physiological information in the multispectral data, remove dimensional differences among sensors, and satisfy deep learning input requirements, the raw data were subjected to detailed preprocessing and feature engineering.

2.3.1 LAI Data Preprocessing

To remove random noise introduced by abrupt illumination changes or instrument operation errors during field measurements, the Savitzky-Golay (S-G) filter was applied to the LAI time series throughout the major growth stages of winter wheat [5]. The S-G filter is a time-domain filtering method based on local polynomial least-squares fitting. Its main advantage is that it suppresses high-frequency noise while preserving local shape features of the growth curve, such as peak width and height [18]. The window length and polynomial order were set to 21 and 2, and the LAI sequence of each sampling plot from regreening to grain filling was smoothed in turn. The S-G filter was used to suppress abrupt local fluctuations while preserving the seasonal growth trajectory. No valid high-LAI observations were removed solely based on magnitude. The procedure thus provided high-quality labels with strong biophysical consistency for HFI-Net.

2.3.2 Construction of Vegetation Index Features

Although deep learning can extract features automatically, introducing prior agronomic knowledge, namely vegetation indices, can markedly accelerate convergence and improve robustness in agricultural remote sensing tasks with limited samples. For the VI feature branch, the original five bands were transformed into eight VI channels rather than directly used as raw spectral bands. These indices were stacked into an eight-channel feature map and used as the input of the VI feature branch.

The calculation formulas for various vegetation indices are shown in Supplementary Table S2. In these formulas, B, G, R, RE, and NIR represent the radiation-corrected surface reflectances of the blue, green, red, red edge, and near-infrared wavelength bands, respectively.

2.3.3 Data Standardization and Tensor Conversion

To accelerate network convergence and prevent gradient vanishing or explosion, all inputs were standardized strictly:

RGB images: the mean and standard deviation of the R, G, and B channels were computed separately, and Z-score normalization was applied as:

$$X_{\text{norm}} = \frac{X - \mu}{\sigma + \epsilon}, \tag{1}$$

where $\epsilon = 10^{-6}$ prevents division-by-zero errors. This operation centers the input distribution near zero and facilitates network weight updates.

Multispectral VI data: because different VIs have very different physical ranges, for example NDVI is bounded while MCARI may take much larger values, each of the eight channels was standardized independently. This normalized all input features to comparable numerical ranges and prevented features with large magnitudes from dominating gradient updates.

Tensor conversion: a ToTensor operation was used to transform image data (H, W, C) into the PyTorch input format (C, H, W) and convert them to floating-point tensors.

2.4 Data Augmentation and Validation Strategy

2.4.1 Spatially Synchronized Geometric Transformations

To simulate UAV observations under different flight passes, headings, and altitudes and to increase spatial diversity, a strict spatially synchronized geometric augmentation strategy was introduced. Because HFI-Net uses a dual-stream architecture, all geometric transformations had to be applied to RGB images and multispectral (VI) images in strict synchrony to preserve spatial alignment in the pixel coordinate system. Specifically, horizontal or vertical flipping was applied with a probability of 0.5 during training to simulate different field row orientations. Rigid rotations by multiples of $90°$ were combined with small-angle random rotations, where the rotation range was limited to $\pm 15°$, to simulate random heading deviations of the UAV. Affine transformations, including translation and scaling, were also introduced to mimic geometric distortions caused by altitude fluctuations or sensor pose tilting.

2.4.2 Modality-Specific Pixel Perturbations

To improve robustness against reduced image quality, modality-specific pixel-level augmentation strategies were designed. Gaussian noise with σ in the range of 0.2-0.44 and motion blur with kernel sizes ranging from 3 to 7 pixels were applied in a OneOf manner with an overall probability of 0.3. The former simulates sensor thermal noise under low-light or high-temperature conditions, and the latter simulates imaging blur caused by airflow disturbances.

Notably, no pixel-level noise was added to the multispectral (VI) data. This is because indices such as NDVI and NDRE have explicit quantitative physical meanings that directly correspond to vegetation biochemical states. Random noise would damage the radiometric fidelity and physical consistency of spectral features and therefore introduce unrealistic spectral variations.

2.4.3 Mixup Augmentation and Manifold Regularization

In addition to conventional augmentation, Mixup was introduced to address data sparsity under the small-sample scenario [37]. Mixup is a data-agnostic augmentation strategy based on vicinal risk minimization.

It constructs a new virtual training sample by linearly interpolating two randomly selected training samples and their labels. Let (x_i, y_i) and (x_j, y_j) denote two randomly selected samples, including RGB data, VI data, and measured LAI values. The mixed sample is generated as:

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1-\lambda)x_j, \\ \tilde{y} &= \lambda y_i + (1-\lambda)y_j. \end{aligned} \tag{2}$$

Here, $\lambda \in [0, 1]$ is a random mixing coefficient drawn from a Beta distribution, $\mathrm{Beta}(\alpha, \alpha)$. In this study, the hyperparameter α was set to 0.4, which makes λ tend to values close to 0 or 1. This preserves the dominant characteristics of the original samples while still introducing moderate perturbation.

The same mixing coefficient λ was applied simultaneously to the RGB patch, the VI patch, and the LAI label.

Introducing Mixup into winter wheat LAI estimation produced a clear dual manifold-regularization effect [31]. On one hand, it forced the model to behave linearly in the feature space between training samples, which smoothed the decision surface and reduced sensitivity to outliers and noisy samples. On the other hand, by generating virtual samples in a continuous feature space, Mixup effectively filled the distribution gaps between discrete field sampling points [34]. This allowed HFI-Net to learn smoother feature representations in unsampled regions and thus improved robustness and generalization when facing the evaluated dataset.

2.4.4 Validation Strategy

Given the limited sample size, a single hold-out split, such as 7: 3, may lead to unstable evaluation results due to sample randomness or extreme observations. Therefore, five-fold cross-validation was adopted to assess the estimation accuracy and robustness of HFI-Net [2]. Specifically, 637 winter wheat samples collected from 91 sampling points across seven growth stages were grouped by sampling location and divided into five approximately equal subsets to prevent spatial correlation leakage. In each fold, one subset was used for validation, while the remaining four subsets were used for model training. This procedure was repeated five times, ensuring that each sample was used for validation exactly once.

After completing all folds, the coefficient of determination (R^2) and root mean square error (RMSE) were averaged across the five validation results and reported as the final quantitative metrics. Combined with the data augmentation strategy

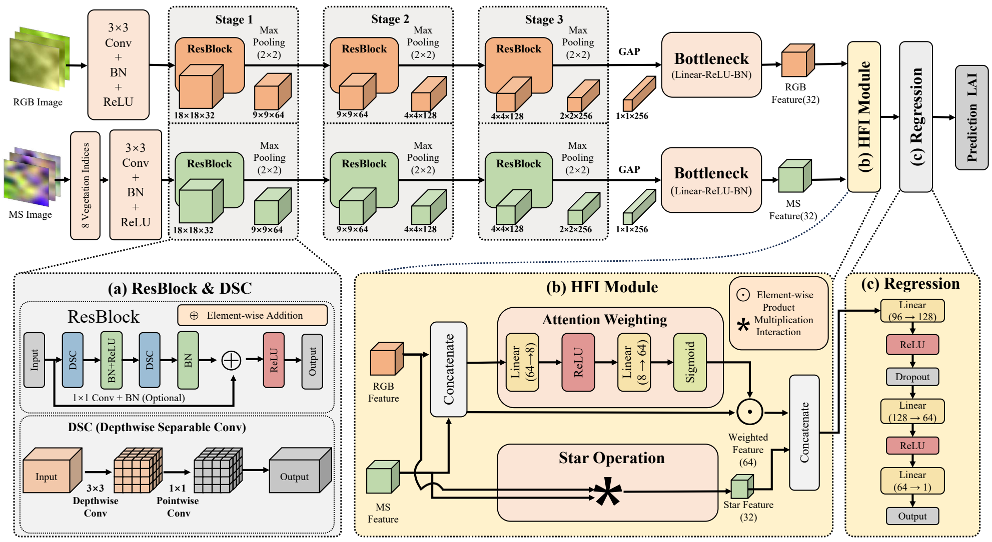

Fig. 2: The architecture of the proposed dual-branch convolutional neural network for leaf area index (LAI) prediction. The main network consists of a dual-stream backbone for extracting features from RGB images and VI features derived from multispectral data, followed by feature fusion and regression. (a) Detailed structure of the Residual Block (ResBlock) with Depthwise Separable Convolution (DSC). (b) Details of the HFI Module for multi-source feature interaction. (c) Architecture of the Regression module for the final LAI prediction.

described above, this validation protocol made full use of the limited field samples without increasing measurement costs, while providing a more reliable evaluation of HFI-Net performance across different growth stages and field plots.

2.5 HFI-Net Model

The proposed model is a lightweight dual-stream network designed for multimodal feature extraction and interaction. It adopts a dual-stream encoder architecture to process RGB texture information and vegetation-index information separately, and it performs feature interaction and regression through the proposed HFI module.

2.5.1 Overall Architecture and Dual-Stream Feature Extraction

To facilitate future deployment on agricultural edge devices, a lightweight dual-stream network with attention enhancement and feature interaction, HFI-Net, was constructed, as shown in Figure 2. HFI-Net mainly consists of a dual-stream feature extractor, a hybrid feature interaction (HFI) module, and a regression predictor. Each module will be introduced in detail below.

To reduce computational complexity while maintaining sufficient feature extraction depth for agricultural edge devices, a lightweight CNN encoder was designed. The RGB branch and VI feature branch share the same architecture but use independent parameters. The RGB branch received a $3 \times H \times W$ patch, whereas the VI feature branch received an $8 \times H \times W$ patch, where $H = W = 18$. In the initial feature extraction stage, a standard 3×3 convolution maps the input channels, three for RGB and eight for VI, to 32 feature channels. Batch normalization and a ReLU activation function are then applied to complete the initial feature-space mapping.

2.5.2 Residual Block Based on Depthwise Separable Convolution

The core feature extraction unit of the encoder is a residual block based on depthwise separable convolutions. To achieve a lightweight design, the conventional ResNet block was modified by replacing standard convolutions with depthwise separable convolutions [6]. This operation is decomposed into two consecutive steps: depthwise convolution and pointwise convolution. First, depthwise convolution independently performs spatial filtering on each input channel without changing the number of channels. Assuming the input features are X, and the depth convolutional kernel is $K^{\mathrm{dw}} \in \mathbb{R}^{D_k \times D_k \times C_{\mathrm{in}}}$, the calculation formula for its output $M \in \mathbb{R}^{H \times W \times C_{\mathrm{in}}}$ is as follows:

$$M_{h,w,c} = \sum_{i=1}^{D_k} \sum_{j=1}^{D_k} K^{\mathrm{dw}}_{i,j,c} X_{h+i-1,w+j-1,c}, \tag{3}$$

where the subscript c denotes the independent operation for the c-th channel.

Following this, pointwise convolution employs a 1×1 convolution kernel $K^{\mathrm{pw}} \in \mathbb{R}^{1 \times 1 \times C_{\mathrm{in}} \times C_{\mathrm{out}}}$ to linearly combine the outputs of the depth convolution across the channel dimension, thereby generating a new set of feature representations $Y \in \mathbb{R}^{H \times W \times C_{\mathrm{out}}}$. The calculation formula for this process is as follows:

$$Y_{h,w,n} = \sum_{c=1}^{C_{\mathrm{in}}} K^{\mathrm{pw}}_{1,1,c,n} M_{h,w,c}, \tag{4}$$

where n denotes the n-th output channel.

To further enhance feature propagation and stabilize optimization, a residual shortcut connection was introduced into the block. Let $F(X)$ denote the residual mapping learned by the main branch, consisting of two successive depthwise separable convolution layers with Batch Normalization. To preserve negative values before addition, a ReLU activation is applied only after the first layer. Let $S(X)$ denote the shortcut mapping. The output of the residual block can then be formulated as:

$$Z = \mathrm{ReLU}(F(X) \oplus S(X)) \tag{5}$$

where $\oplus$ denotes element-wise addition, and $S(X) = X$ when the input and output dimensions are consistent; otherwise, a 1×1 convolution followed by Batch Normalization is applied to match the channel dimensions. This residual design helps preserve low-level information, facilitates gradient propagation, and improves training stability while maintaining a lightweight architecture.

2.5.3 Hierarchical Encoder Architecture

At the macro level, the encoder contains three residual stages, and each stage includes the residual block above and a 2×2 max-pooling layer. As network depth increases, the number of feature channels grows progressively from 32 to 64, 128, and 256, while spatial resolution is halved stage by stage. This allows the network to extract increasingly abstract semantic information from low-level textures. At the end of the encoder, global adaptive average pooling compresses the two-dimensional feature map into a one-dimensional vector of length 256. A bottleneck layer consisting of a fully connected layer, ReLU, and BatchNorm then performs strong dimensionality reduction. Through this encoder, high-dimensional image data are converted into compact 32-dimensional feature representations ($F_{\text{rgb}} \in \mathbb{R}^{32}$ and $F_{\text{vi}} \in \mathbb{R}^{32}$) with discriminative information and reduced spatial redundancy, thereby laying the foundation for later cross-modal feature interaction.

After deep feature extraction, the outputs of the two branches are sent to the HFI module for modality fusion, as detailed in Section 2.5.4. The fused high-dimensional feature vector is then fed into a multilayer perceptron (MLP) for final LAI regression. The predictor consists of three fully connected layers that map the feature dimension from 96 to 128, then to 64, and finally to a single neuron that outputs a continuous LAI estimate. The 96-dimensional fused vector consisted of a 64-dimensional attention-weighted feature and a 32-dimensional multiplicative interaction feature.

To prevent overfitting on the limited agricultural sample set, Dropout regularization was introduced between fully connected layers with a dropout rate of 0.3. This strategy randomly deactivates some neuron connections during training, which forces the network to learn more robust feature representations and thus improves prediction stability in complex field conditions.

2.5.4 Hybrid Feature Interaction Module

The HFI module is the core component of HFI-Net. It is designed to fuse RGB structural features and VI features beyond simple linear concatenation. Specifically, it uses a dual-branch interaction strategy: one branch adaptively reweights global channel responses, while the other captures nonlinear cross-modal interaction.

Given the extracted RGB feature F_{rgb} and VI feature F_{vi}, the two features are first concatenated along the channel dimension to obtain:

$$F_{\text{concat}} = [F_{\text{rgb}}, F_{\text{vi}}] \in \mathbb{R}^{64}. \tag{6}$$

This joint representation preserves the original spatial and spectral information and serves as the input to the following interaction process.

In Branch A, an SE-style channel attention mechanism is used to estimate the relative importance of different channels. The concatenated feature F_{cat} is passed through fully connected layers and nonlinear activations to generate a channel-wise weight vector $W \in \mathbb{R}^{64}$, where $W_i \in (0, 1)$. The weighted feature is then obtained by:

$$F_{\text{weighted}} = W \odot F_{\text{concat}}, \tag{7}$$

where $\odot$ denotes element-wise product. This branch enables the model to adaptively adjust the contributions of different feature channels during multimodal fusion.

In Branch B, a star-shaped multiplicative interaction [23] is introduced to model nonlinear coupling between heterogeneous modalities:

$$F_{\text{star}} = F_{\text{rgb}} * F_{\text{vi}}, \tag{8}$$

where $*$ denotes the element-wise multiplication operation:

$$(F_{\text{rgb}} * F_{\text{vi}})_i = F_{\text{rgb},i} \cdot F_{\text{ms},i}. \tag{9}$$

Unlike linear feature stacking, this multiplicative interaction explicitly captures second-order cross-modal dependencies between canopy structural and spectral features.

Finally, the weighted linear feature and nonlinear interaction feature are concatenated to form the fused representation:

$$F_{\text{fusion}} = [F_{\text{weighted}}, F_{\text{star}}] \in \mathbb{R}^{96}, \tag{10}$$

which is then fed into the downstream MLP regressor for LAI estimation. In this design, F_{weighted} preserves the global spectral and structural responses, while F_{star} provides a nonlinear interaction pathway. From a physics-inspired perspective, the multiplicative interaction provides an intuitive analogy to the nonlinear relationship between canopy structural characteristics and spectral responses [25]. This helps alleviate the adverse effects of spectral saturation on LAI estimation in dense winter wheat canopies [11].

2.5.5 Experimental Environment and Parameter Settings

All experiments were conducted on a Windows 10 64-bit platform equipped with an AMD Ryzen 7 5800H CPU and an NVIDIA GeForce RTX 3050 (NVIDIA, Santa Clara, CA, USA) Laptop GPU. The software environment was configured with Python 3.12 and PyTorch 2.5.1. CUDA 12.1 and the corresponding cuDNN library were used to accelerate model training and inference. To facilitate fair comparison and reproducibility, fixed random seeds were used, and all compared methods followed the same data partitioning strategy.

The model was trained for 50 epochs without early stopping. The batch sizes for training and testing were set to 16 and 8, respectively. The RGB and vegetation index (VI) patches were cropped to 18×18 pixels, and the random seed was fixed at 42. Smooth L1 Loss was adopted as the objective function. Compared with mean squared error (MSE), Smooth L1 Loss behaves similarly to L2 loss when the prediction error is small, which helps maintain stable convergence. When the error is large, it behaves closer to L1 loss, thereby limiting the gradient magnitude and improving robustness to possible abnormal field measurements.

AdamW with weight decay was used to optimize the network parameters. The initial learning rate was set to 0.005, and the weight decay coefficient was set to

10^{-4}. The learning rate was determined through preliminary validation experiments and kept consistent across all folds. In addition, a ReduceLROnPlateau scheduling strategy was adopted, with the validation loss used as the monitoring indicator. When the validation loss did not decrease for three consecutive epochs, the learning rate was reduced by a factor of 0.5. This strategy enabled rapid optimization in the early training stage and more fine-grained parameter adjustment in the later stage.

2.5.6 Model Evaluation Metrics

To evaluate estimation accuracy and model performance comprehensively and objectively, the coefficient of determination (R^2) and root mean square error (RMSE) were used as the core quantitative metrics. Specifically, R^2 quantifies the proportion of variance in the measured LAI that can be explained by the model, and is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{11}$$

where y_i and $\hat{y} * i$ denote the measured and predicted LAI values of the i-th sample, respectively, $\bar{y}$ represents the mean of the measured LAI values, and n is the number of samples. A higher R^2 value indicates stronger explanatory ability and better agreement between predicted and measured LAI. RMSE was further employed to measure the average prediction error between the predicted and measured values, which is more sensitive to large deviations and has the same unit as LAI:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{12}$$

A lower RMSE value indicates smaller prediction errors and better estimation accuracy.

3 Results

3.1 Variability Analysis of LAI Data

Before model training, descriptive statistical analysis was conducted on the ground-measured winter wheat LAI data. The dataset covered seven key time points from regreening to grain filling and included 637 samples. The analysis results are shown in Figure 3.

The results showed that winter wheat LAI followed a typical single-peak temporal pattern during the whole season [19]. LAI was low during regreening, with a mean value of about 1.215 and a relatively high coefficient of variation (CV), which reflected uneven emergence and early growth in the field. During jointing and booting, LAI increased rapidly with rising temperature and increasing water and fertilizer supply. At heading, LAI reached its seasonal peak, and some plots under high density and high nitrogen treatment exceeded 6.0, forming highly dense canopies. During flowering

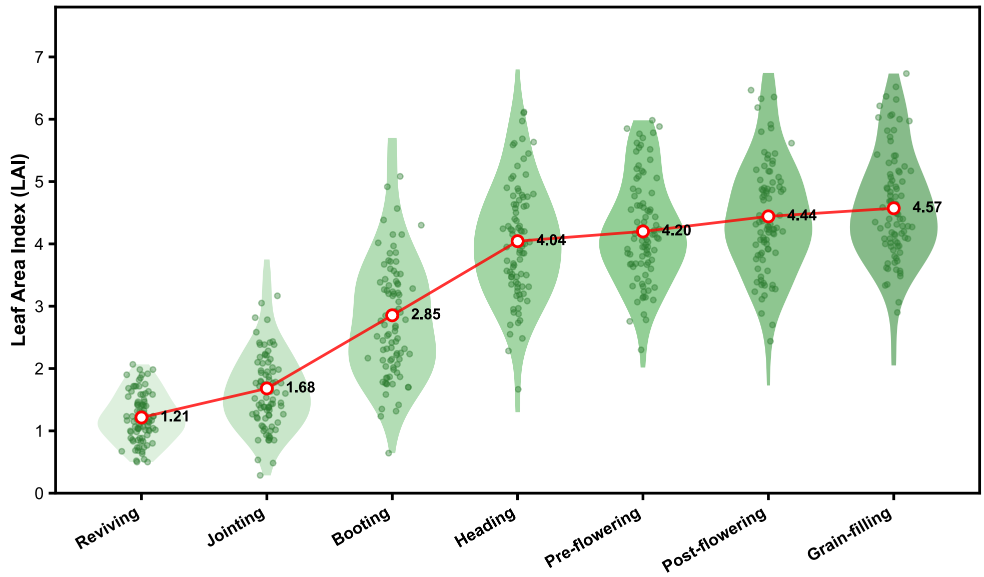

Fig. 3: Temporal evolution and distribution of winter wheat LAI across different growth stages.

and grain filling, lower leaves began to senesce and yellow, so effective green leaf area decreased slightly and then stabilized.

The broad LAI range from 0.5 to 6.8 covered canopy conditions from very sparse to very dense and thus provided an ideal basis for testing model robustness under different cover levels. Analysis also showed that preprocessing removed a small number of outliers caused by measurement errors. The overall distribution covered a broad LAI range and was suitable for training regression models.

3.2 Correlation Analysis between Vegetation Indices and LAI

To explore the intrinsic relationships between vegetation indices (VIs) and leaf area index (LAI), as well as their data distribution characteristics, Spearman rank correlation analysis and significance testing were performed. As shown in Figure 4, the correlation matrix used multidimensional visualization to reveal the interaction mechanism among features. The diagonal displays the feature names. The upper-right triangle presents Spearman correlation coefficients and significance levels through color gradients and starred values. The lower-left triangle uses Gaussian kernel density estimation (KDE) to show the joint density distribution of two variables.

Several important agronomic and remote sensing conclusions can be drawn from Figure 4.

First, the red-edge band showed a clear advantage in saturation-prone conditions. The data indicate that NDRE, which is based on the red-edge band, had the strongest correlation with LAI (r = 0.83), clearly outperforming the traditional NDVI (r = 0.71) [16]. As a result, it cannot easily detect leaf information deep inside the canopy.

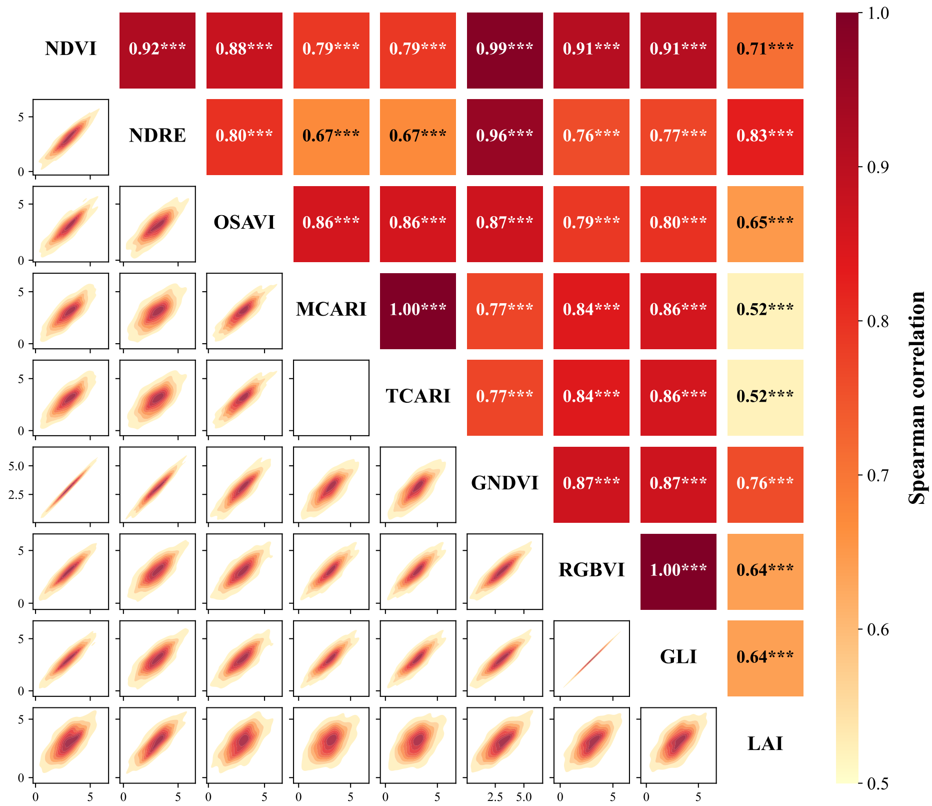

Fig. 4: Correlation of vegetation indices with LAI. *, **, and *** indicate significance at $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively.

In contrast, the red-edge band lies in the transition zone between the chlorophyll absorption peak and the scattering peak. It therefore has stronger canopy penetration and can effectively alleviate signal saturation in dense canopies.

Second, the green band showed strong sensitivity. GNDVI (r = 0.76) performed only slightly below NDRE and better than NDVI.

It is also noteworthy that structural and biochemical indices played highly complementary roles in the model. MCARI and TCARI had relatively low linear correlations with LAI, both with r = 0.52. Through nonlinear mapping, HFI-Net can use the physiological stress signals provided by MCARI to correct LAI estimation bias that would arise from structural information alone.

In addition, visible-light indices have inherent physical limitations. RGBVI and GLI, which rely only on visible bands, both showed correlations of 0.64 and were generally weaker than indices that include near-infrared information.

Even NDRE, which performed best, did not reach a correlation coefficient of 0.9. Therefore, constructing a multidimensional feature space that combines highly correlated structural indices, such as NDRE, with highly sensitive biochemical indices,

Table 1: Comparison of model performance.

Model	Params (M)	Floating-point operations (FLOPs) (G)	R^2	RMSE
RF	–	–	0.8155±0.0102	0.6694±0.0224
SVR	–	–	0.7890±0.0105	0.7081±0.0217
MobileNetV3 [15]	1.52	0.0072	0.8164±0.0129	0.6994±0.0206
ResNet50 [14]	23.51	0.5044	0.8705±0.0127	0.5549±0.0189
ShuffleNetV2 [22]	1.26	**0.0081**	0.8014±0.0132	0.6870±0.0208
EfficientNet [29]	4.01	0.0213	0.7007±0.0143	0.8434±0.0245
Ours	**0.4033**	0.0168	**0.9023±0.0096**	**0.4822±0.0175**

Notes: FLOPs are reported based on single-sample inference and are counted using the MAC convention.

such as MCARI, is a necessary choice for improving the robustness of deep learning estimation models under complex field conditions.

3.3 Performance Evaluation of HFI-Net

3.3.1 Overall Model Performance Evaluation

To evaluate the effectiveness of HFI-Net for winter wheat LAI estimation, it was compared with six representative mainstream algorithms, including traditional machine learning models (RF and SVR) and deep learning networks with different parameter scales (MobileNetV3, ShuffleNetV2, EfficientNet, and ResNet50). All models were trained and tested on the same multi-source dataset composed of RGB and multispectral imagery. For random forest (RF) and support vector regression (SVR), the input features were handcrafted statistical fusion features, consisting of the mean and standard deviation of each RGB channel and each VI channel (22 dimensions in total). For fair comparison, all deep learning baselines were modified to accept the same RGB and VI-derived information as the proposed model, which was provided as a concatenated multi-channel input. The quantitative results are shown in Table 1.

Model complexity was evaluated in terms of trainable parameters and computational cost. The number of trainable parameters was calculated as

$$\text{Params} = \sum_{l=1}^{L} N_l, \tag{13}$$

where N_l denotes the number of parameters in the l-th trainable layer, and L is the total number of trainable layers.

Computational cost was estimated by accumulating the multiply-accumulate operations (MACs) of convolutional and fully connected layers during single-sample inference. For a convolutional layer, the MACs were computed as

$$\text{MACs}_{\text{conv}} = \frac{K_h K_w C_{\text{in}} C_{\text{out}} H_{\text{out}} W_{\text{out}}}{G}, \tag{14}$$

where K_h and K_w are the kernel height and width, C_{in} and C_{out} denote the input and output channels, H_{out} and W_{out} are the spatial dimensions of the output feature map, and G is the number of groups. For a fully connected layer, the MACs were calculated as:

$$\text{MACs}_{\text{fc}} = C_{\text{in}} C_{\text{out}}. \tag{15}$$

In this study, MACs were reported as FLOP-equivalent units for relative complexity comparison, following common practice in lightweight model analysis. The statistical results are summarized in Table 1.

Several observations can be drawn from the results. First, deep learning models generally improved LAI estimation compared with traditional regressors, but their performance was strongly related to the suitability of the network architecture for small UAV patches. HFI-Net achieved RMSE reductions of 28.0% and 31.9% compared with RF and SVR, respectively.

Second, commonly used lightweight networks showed clear limitations in this task. MobileNetV3 and ShuffleNetV2 achieved R^2 values of only about 0.80-0.81 and did not achieve higher accuracy than RF. In contrast, HFI-Net combines reduced downsampling with depthwise separable convolutions, achieving a favorable trade-off between accuracy and computational efficiency with only 0.40 M parameters, approximately one quarter of MobileNetV3.

Finally, HFI-Net achieved a better balance between accuracy and efficiency than heavier networks. Although ResNet50 obtained the second-best performance with $R^2 = 0.8705$, it required 23.51 M parameters and 0.5044 G FLOPs. By comparison, HFI-Net achieved a higher R^2 of 0.9023 using only about 1/58 of the parameters and 1/30 of the computation.

To further evaluate prediction accuracy and robustness, a scatter-density plot of measured and predicted LAI values across the five validation folds was generated, as shown in Figure 5.

Figure 5 shows that the sample points are highly concentrated around the 1: 1 reference line, and the fitted line almost overlaps the reference line. This indicates that the model exhibited limited systematic bias across the observed full LAI range of 0.5 to 6.8.

It is worth noting that the red-to-orange high-density region in the figure is mainly distributed within the core LAI interval of 1.0-5.0, which corresponds to the main growth stages from jointing to heading. This indicates reliable prediction performance for most field samples. At the same time, even in the high-value region with LAI $>$ 6.0 and the low-value region with LAI $<$ 1.0, the scatter points generally remained close to the reference trend without obvious systematic deviation. This indicates that HFI-Net effectively demonstrates consistent prediction performance across high and low LAI ranges and maintains stable performance across samples involving different cultivars and nitrogen treatments.

3.3.2 Evaluation of High-LAI Prediction Capability

To quantitatively evaluate the prediction capability of HFI-Net under saturation-prone high-density canopy conditions, the 168 samples in the test set were divided into two subsets based on measured LAI values: LAI $\leq$ 4.0 (low-to-medium density canopy,

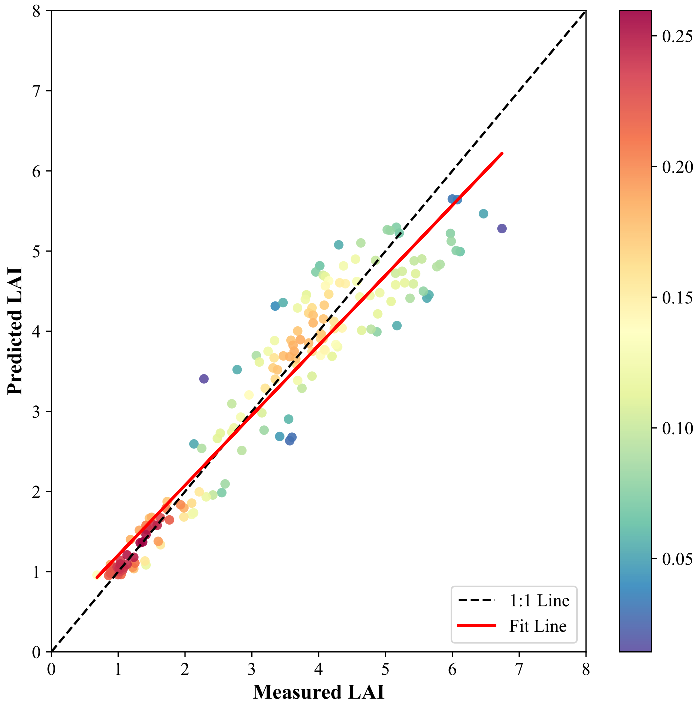

Fig. 5: Scatter-density plot of measured and predicted values on the test set.

104 samples) and LAI > 4.0 (high-density canopy, 64 samples), with R^2, RMSE, and mean absolute error (MAE) metrics calculated separately. To mitigate the impact of training randomness, all comparison models underwent three independent replicate experiments using fixed seed values (42, 123, 456), and subset metrics were reported as mean $\pm$ standard deviation. For the LAI > 4.0 subset, where LAI values are concentrated within a narrow high-range interval (approximately 4.0$\sim$6.8), the total sum of squares (SST) of the dependent variable is significantly smaller than that observed across the full range. Since R^2 is a normalization metric highly sensitive to SST variations, minor fluctuations in prediction residuals across the three replicates could lead to substantial differences or even sign reversals in R^2 values, making the reported standard deviation difficult to interpret and potentially misleading readers regarding model stability. Therefore, the R^2 for this subset was computed as the pooled result from all three experiments without additional standard deviation. The results are presented in Table 2.

The stratified evaluation results in Table 2 show clear differences in model performance under different canopy densities.

First, all models showed reduced accuracy when moving from the low-density subset to the high-density subset, but the degree of degradation varied substantially. For samples with LAI $\leq$ 4.0, most models maintained acceptable performance. HFI-Net achieved an R^2 of 0.8805, followed by ResNet50 with 0.8417, while ShuffleNetV2 and MobileNetV3 obtained 0.7563 and 0.7583, respectively. However, for samples with LAI

Table 2: Hierarchical comparison of model performance on different LAI density intervals.

Model	Subset	R^2	RMSE	MAE
HFI-Net (Ours)	All	0.9023 ± 0.0096	0.4822 ± 0.0175	0.3719
	≤4.0	0.8805 ± 0.0094	0.3659 ± 0.0133	0.2719
	>4.0	0.2251	0.6352 ± 0.0231	0.5344
ResNet50 [14]	All	0.8705 ± 0.0127	0.5549 ± 0.0189	0.4111
	≤4.0	0.8417 ± 0.0095	0.4208 ± 0.0150	0.3186
	>4.0	0.0634	0.6911 ± 0.0999	0.5614
ShuffleNetV2 [22]	All	0.8014 ± 0.0132	0.6870 ± 0.0208	0.4861
	≤4.0	0.7563 ± 0.0100	0.5220 ± 0.0170	0.4037
	>4.0	−0.2037	0.7914 ± 0.0212	0.6200
MobileNetV3 [15]	All	0.8164 ± 0.0129	0.6994 ± 0.0206	0.4820
	≤4.0	0.7583 ± 0.0100	0.5184 ± 0.0170	0.3994
	>4.0	−0.2094	0.7906 ± 0.0676	0.6161

> 4.0, the performance gap became much larger. ShuffleNetV2 and MobileNetV3 produced negative R^2 values (-0.2037 and -0.2094), indicating that their predictions were less reliable than simply using the subset mean. ResNet50 still maintained a positive R^2 of 0.0634, but its RMSE increased to 0.6911, with a standard deviation of 0.0999 across three independent runs. This suggests that although a deeper architecture can partly improve feature extraction, it does not fully alleviate the performance degradation observed under high-LAI dense canopy conditions.

In contrast, HFI-Net showed more stable prediction performance in the high-density region. For LAI > 4.0, it achieved an R^2 of 0.2251 and an RMSE of 0.6352, reducing RMSE by 8.1% compared with ResNet50 on the LAI > 4.0 subset. More importantly, the RMSE standard deviation of HFI-Net was only 0.0231, much lower than that of ResNet50.

For HFI-Net, the gap between RMSE and MAE was moderate on the LAI ≤ 4.0 subset, indicating that most errors were relatively small and stable. On the LAI > 4.0 subset, the gap increased from 0.3659 versus 0.2719 to 0.6352 versus 0.5344, suggesting the presence of a few larger deviations, but the overall error remained controlled. By contrast, ResNet50 showed a larger RMSE-MAE gap in the high-density subset, indicating more severe outliers. The RMSE values of ShuffleNetV2 and MobileNetV3 approached 0.8, further confirming their limited reliability in the saturation range.

3.3.3 Ablation Experiments and Feature Interaction Analysis

To evaluate the contribution and effectiveness of the dual-stream architecture, attention mechanism, and star-shaped interaction module in HFI-Net, five variant models were constructed for ablation experiments. The results are shown in Table 3.

The ablation results provide important insight into the contributions of each model component.

First, heterogeneous modalities showed clearly different utility in estimation, as seen in M1 versus M2. The RGB-only model, M1, achieved relatively low accuracy

Table 3: Ablation experiments.

Model	Modalities	Fusion strategy	R^2	RMSE
M1	RGB only	–	0.5962±0.0227	0.9797±0.0377
M2	VI only	–	0.8657±0.0122	0.5650±0.0270
M3	RGB + VI	Direct Concat	0.8651±0.0137	0.5663±0.0266
M4	RGB + VI	Attention	0.8540±0.0145	0.5890±0.0244
Ours	RGB + VI	Attn + Star	**0.9023±0.0096**	**0.4822±0.0175**

with $R^2 = 0.5962$, which indicates that multispectral information provides important spectral cues related to winter wheat growth status. However, the single-modality VI model, M2, also showed limited improvement under dense canopy conditions, possibly due to insufficient structural information.

Second, direct linear concatenation did not fully exploit the complementarity between heterogeneous modalities, as shown by M2 versus M3. A key observation was that directly concatenating RGB and VI features in M3 did not improve performance and instead produced slightly worse results than the VI-only model.

Finally, star-shaped interaction enhanced the utilization of spectral features under dense canopy conditions, as shown by M4 versus the proposed model. The results indicate that adding attention alone in M4 could not improve the prediction performance. Attention reweighting alone may be insufficient to model nonlinear interactions between heterogeneous modalities when no explicit nonlinear interaction is provided. In contrast, introducing star-shaped multiplicative interaction raised R^2 to 0.9023.

3.3.4 Evaluation of Prediction Capability under Dense Canopies

During the middle and late growth stages of winter wheat, from booting to grain filling, the canopy becomes highly closed. Under such conditions, optical remote sensing signals often suffer from saturation, which limits accurate LAI estimation in high-yield fields. To examine the performance of HFI-Net under saturation-prone conditions, the response patterns of NDVI and HFI-Net predictions were compared against measured LAI, as shown in Figure 6.

The two methods show clearly different response patterns. In the low-cover range with LAI < 3.0, NDVI increases rapidly with LAI, indicating good sensitivity to canopy variation. However, saturation effects commonly become more pronounced when LAI exceeds approximately 4.0, depending on vegetation type, sensor characteristics, and environmental conditions. As indicated by the gray shaded region, the NDVI response gradually flattens and remains mainly within 0.85–0.90. This plateau suggests that the spectral index becomes less sensitive to further increases in canopy density, making it difficult to distinguish structural differences among dense canopies.

By contrast, HFI-Net maintains a more consistent response across the full LAI range. Even in the high-density region with LAI > 4.0, the predicted values continue to follow the increasing trend of measured LAI, without a prominent prediction plateau. When VI responses become less distinguishable under dense canopies, RGB-derived

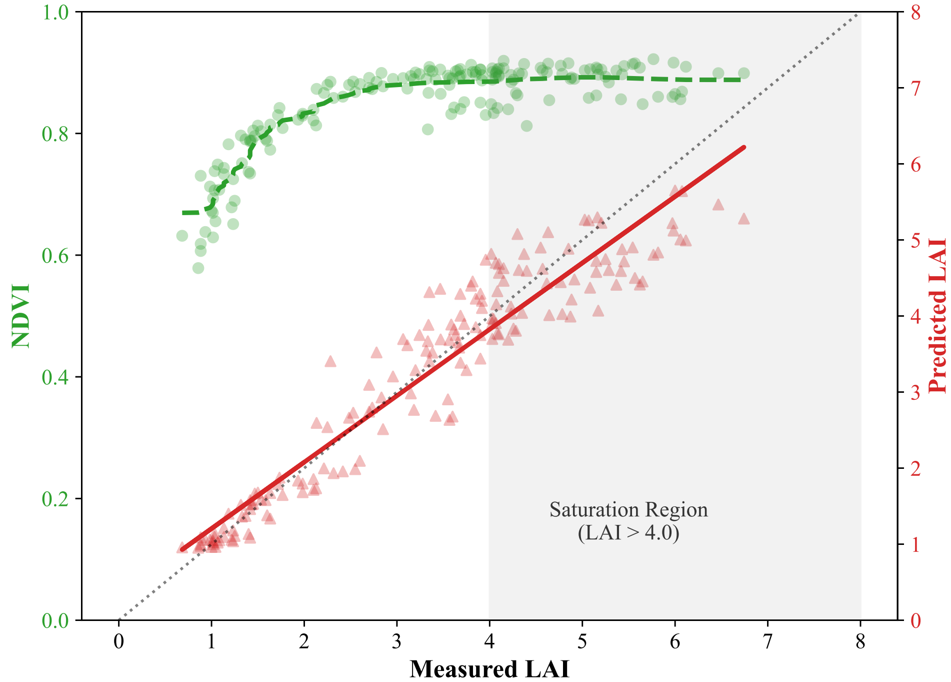

Fig. 6: Comparative analysis of prediction capability under saturation-prone conditions: NDVI versus HFI-Net.

texture features and spatial structural patterns provide additional complementary cues. Through nonlinear feature interaction, these cues enhance the spectral representation and help mitigate the loss of sensitivity under high-LAI conditions.

Compared with NDVI, which mainly depends on spectral contrast, HFI-Net benefits from the joint use of spectral information and canopy structural features, thereby improving the stability of LAI estimation in high-LAI regions.

3.3.5 Evaluation of Estimation Accuracy and Robustness

To evaluate the stability of HFI-Net under limited samples, five-fold cross-validation was conducted. The results showed highly consistent performance across different validation folds, with an average coefficient of determination (R^2) of 0.9023. At the same time, the fluctuation range of RMSE across folds was minimal. This low variance not only indicates consistent prediction performance of the model but also indicates that the lightweight architecture combined with manifold data augmentation may help improve robustness under limited samples. Statistically, the results indicate that HFI-Net achieves stable performance across validation folds with diverse field samples that include different cultivars, micro-topography, and nitrogen treatments. This study recorded the results of the five-fold cross-validation and shown in Supplementary Table S3.

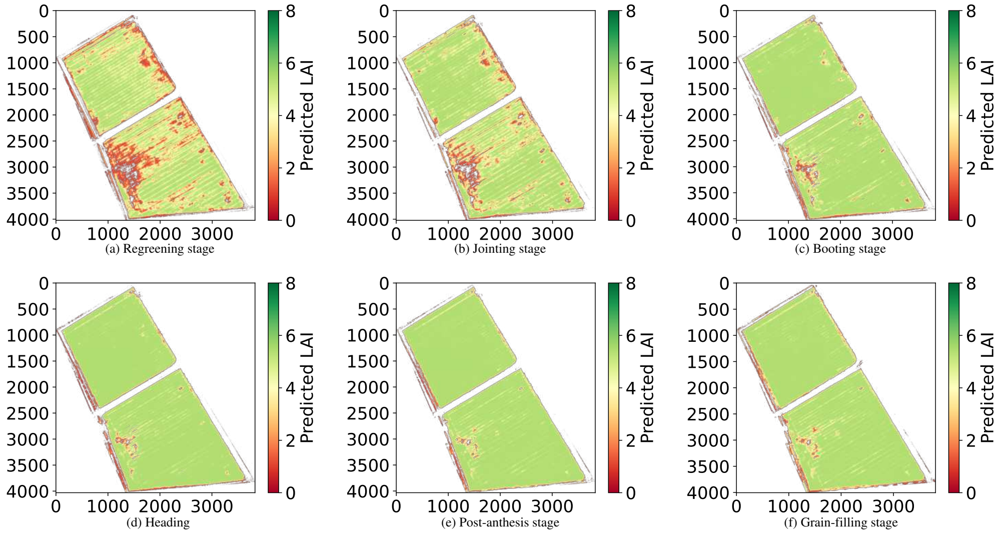

Fig. 7: LAI maps at different growth stages.

3.4 Full-Coverage LAI Mapping and Temporal Evolution Analysis

Based on the best-trained HFI-Net model, full-coverage LAI mapping was performed using UAV imagery acquired from regreening to grain filling. To reduce stitching artifacts and improve mapping efficiency, a sliding-window inference strategy was adopted, and overlapping predictions were merged by weighted averaging. Figure 7 presents the high-resolution LAI maps of winter wheat at six key growth stages.

Combined with field management conditions, the generated maps reveal clear spatial and temporal patterns. First, HFI-Net effectively characterized field-scale spatial heterogeneity. As shown in Figure 7, the maps not only reflect the overall LAI gradients among plots under different nitrogen treatments, but also preserve local variations caused by uneven sowing density, micro-topography, or within-plot growth differences. These results provide spatial information that could support precision agricultural management applications.

Second, the mapping results further confirm the enhanced prediction capability of HFI-Net under dense-canopy conditions. During booting and heading, when canopy closure reaches a high level, the LAI maps still retain clear value gradients rather than showing large homogeneous high-value areas. The star-shaped multiplicative interaction contributes to this behavior by leveraging RGB-derived structural cues to provide complementary structural information when spectral sensitivity weakens in dense canopies.

In addition, HFI-Net shows potential for efficient deployment. With only 0.40 M parameters, it provides efficient inference capability for UAV-based LAI mapping. The mapped results also show smooth transitions around field boundaries and non-vegetated regions, such as ridges, without noticeable artifacts.

4 Discussion

This study proposed HFI-Net, a lightweight dual-stream network for winter wheat LAI estimation from UAV RGB and multispectral imagery. The main motivation was to improve LAI estimation under dense-canopy conditions while maintaining a low computational cost. The experimental results show that HFI-Net achieved higher accuracy than conventional regressors and deeper CNN models, despite using only 0.40 M parameters. This indicates that, for small-patch agricultural remote sensing tasks, model performance depends not only on network depth or parameter scale, but also on whether the architecture matches the characteristics of the specific remote sensing estimation problem.

4.1 Mechanism for Improved High-LAI Estimation in HFI-Net

Spectral saturation is a common limitation in LAI estimation, especially during the middle and late growth stages when the winter wheat canopy becomes highly closed. This occurs because in high-LAI conditions, changes in red-band reflectance become less sensitive due to strong chlorophyll absorption [16]. Consequently, traditional vegetation indices mainly relying on spectral contrast experience decreased sensitivity when canopy reflectance responses become saturated. The correlation analysis confirmed that no single index can fully capture LAI variation across the full growing season. While indices like NDVI saturate, structural and biochemical indices play highly complementary roles. Indices such as MCARI and TCARI, despite showing lower linear correlation with LAI, are highly sensitive to small variations in leaf chlorophyll concentration (LCC) instead of responding linearly to canopy-scale structure.

HFI-Net alleviates the impact of spectral saturation on LAI estimation by introducing star-shaped multiplicative interaction between RGB-derived structural features and VI features. When spectral differences become less distinguishable under dense canopy conditions, this nonlinear interaction provides complementary structural information. RGB texture features, such as canopy gaps, spatial structural cues, and local structural patterns, provide structural cues that guide multimodal feature interactions. Specifically, when vegetation-index responses become less sensitive in high-LAI regions, the element-wise multiplication encourages the model to incorporate structural variation during multimodal feature fusion. The improvement in high-LAI regions is consistent with the contribution of this module, suggesting that the observed advantage is not caused by random variation. Although this design does not explicitly model radiative transfer processes, it follows the physical intuition that canopy structure influences spectral reflectance, thereby better improving LAI estimation stability under dense canopy conditions.

4.2 Multimodal Fusion and Lightweight Deployment

The experimental results suggest that effective multimodal fusion is critical for UAV-based LAI estimation. RGB imagery alone cannot adequately describe the spectral variability associated with vegetation canopies, indicating the importance of introducing vegetation-index information. Conversely, directly concatenating RGB and VI features may cause limited feature complementarity, which may not be fully optimized for heterogeneous feature interaction. Their linear superposition may limit effective feature interaction and lead to suboptimal multimodal representation. HFI-Net reduces this problem through independent dual-stream encoding and attention-guided feature interaction. Within a deep learning framework, low-linear-correlation biochemical features provide information orthogonal to structural indices. The dual-stream design effectively preserves this complementary information and helps improve feature representation quality.

In addition, HFI-Net shows clear advantages in computational efficiency over both traditional regressors and generic deep networks. Traditional regressors based on handcrafted features are limited in capturing high-frequency spatial texture information. Meanwhile, standard lightweight models originally designed for large-scale natural images usually apply stride-2 downsampling at shallow layers, which can cause irreversible loss of texture details in small agricultural patches. The results highlight the importance of preserving spatial resolution in small-patch remote sensing estimation. The physically inspired feature interaction design can capture cross-modal relationships more efficiently than simply increasing model depth or parameter scale. This indicates the potential of HFI-Net for resource-constrained UAV applications.

4.3 Limitations and Future Work

Although HFI-Net achieved promising performance, several limitations remain. First, the dataset was collected from one experimental region and one growing season. Although five-fold cross-validation was used to evaluate internal stability, the model's performance consistency across years, regions, cultivars, and crop types still requires further validation. Future work should include multi-year and multi-site experiments to assess the robustness and transferability of the proposed method.

Second, HFI-Net is currently physics-inspired rather than fully physics-constrained. The star-shaped interaction implicitly models the relationship between canopy structural features and spectral representations, but it has not yet been directly linked to radiative transfer models such as PROSAIL. Future studies could incorporate physical constraints into the loss function or model structure to further improve interpretability and potentially reduce dependence on labeled field samples [20, 33, 39].

Finally, this study has not yet evaluated deployment performance on real UAV edge devices. Future work should test inference speed, energy consumption, and mapping stability on embedded platforms, so as to better support real-time LAI monitoring and field management applications.

5 Conclusion

This study presented HFI-Net, a lightweight dual-stream network for improving winter wheat LAI estimation under dense canopy conditions. By integrating RGB texture and VI features through attention-guided star-shaped multiplicative interactions, HFI-Net effectively captures nonlinear relationships between canopy structural and spectral representations. The results indicated improved prediction performance, particularly under high-LAI conditions where spectral saturation effects commonly reduce estimation sensitivity. The proposed physics-inspired fusion strategy achieved a favorable balance between accuracy and computational efficiency while maintaining stable performance across the evaluated phenological stages. Future studies will focus on validating the model across multiple regions, years, and crop species, and incorporating additional physical constraints to further improve interpretability and robustness.

Table S1: Parameters of the DJI P4M.

Parameter	Value
complementary metal-oxide-semiconductor (CMOS)	1/2.9-inch
Pixels	2.08 megapixels (effective)
field of view (FOV)	62.7°
Focal length	5.74 mm
Aperture	f/2.2
Maximum resolution	1600 × 1300
Photo format	JPEG (visible) + TIFF (multispectral)

Table S2: Selected vegetation indices used in this study.

Vegetation index	Formula
NDVI [27]	$(NIR - R)/(NIR + R)$
NDRE [4]	$(NIR - RE)/(NIR + RE)$
GNDVI [10]	$(NIR - G)/(NIR + G)$
OSAVI [26]	$(1 + 0.16)(NIR - R)/(NIR + R + 0.16)$
MCARI [7]	$[(RE - R) - 0.2(RE - G)] \times (RE/R)$
TCARI [12]	$3 \times [(RE - R) - 0.2(RE - G) \times (RE/R)]$
RGBVI [3]	$(G^2 - B \times R)/(G^2 + B \times R)$
GLI [21]	$(2G - R - B)/(2G + R + B)$

Table S3: The results of each fold.

Fold	R^2	RMSE
Fold 1	0.9112	0.4615
Fold 2	0.8945	0.4988
Fold 3	0.9087	0.4731
Fold 4	0.8892	0.5024
Fold 5	0.9097	0.4752
Mean	0.9023	0.4822
Standard deviation	0.0096	0.0175

Supplementary Material

Acknowledgements

The authors would like to thank the Runguo Agricultural Ecology Base for providing the experimental site and field support.

Declarations

Funding: This work was supported by the Major Project of Natural Science Foundation of the Educational Committee of Anhui Province (grant number 2024AH040175); the Key Project of Natural Science Foundation of the Educational Committee of Anhui Province (grant number 2024AH051099); the Key and General Projects of Jiangsu Province (grant number BE2022338); the Project of Faculty of Agricultural Engineering of Jiangsu University (grant number NZXB20200102); the Open Fund for Key Laboratory of Modern Agricultural Equipment and Technology, Ministry of Education of the People's Republic of China (grant number MAET202113); and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (grant number PAPD-2023-87).

Competing Interests: 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.

Author Contributions: Conceptualization, Methodology, Writing - original draft: Shihao Mei; Validation, Investigation, Data Curation: Yusheng Cheng; Writing - review & editing: Chundu Wu, Lulu Zhang, Xiaowen Wang; Supervision, Project administration, Funding acquisition: Lulu Zhang. All authors have read and agreed to the published version of the manuscript.

AI Assistant Declaration: AI-assisted technologies were used solely for linguistic refinement and academic tone improvement during the manuscript preparation. The authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

[1] Abba JH, Asheikh YA, Kumar A, et al (2026) Drone-based phenotyping and its utilization in crop improvement: A review. International Journal of Latest Technology in Engineering, Management & Applied Science 15(2):1–9. https://doi.org/10.51583/IJLTEMAS.2026.1502000001

[2] Beigaite R, Mechenich M, Zliobaite I (2022) Spatial cross-validation for globally distributed data. In: Discovery Science. Springer, pp 127–140, https://doi.org/10.1007/978-3-031-18840-4_10

[3] Bendig J, Yu K, Aasen H, et al (2015) Combining UAV-based plant height from crop surface models, visible, and near infrared vegetation indices for biomass monitoring in barley. International Journal of Applied Earth Observation and Geoinformation 39:79–87. https://doi.org/10.1016/j.jag.2015.02.015

[4] Bronson KF, Conley MM, French AN, et al (2020) Which active optical sensor vegetation index is best for nitrogen assessment in irrigated cotton? Agronomy Journal 112(2):1143–1156. https://doi.org/10.1002/agj2.20146

[5] Chen J, Jonsson P, Tamura M, et al (2004) A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky-Golay filter. Remote Sensing of Environment 91(3-4):332–344. https://doi.org/10.1016/j.rse.2004.03.014

[6] Chollet F (2017) Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1251–1258, https://doi.org/10.1109/CVPR.2017.195

[7] Daughtry CST, Walthall CL, Kim MS, et al (2000) Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. Remote Sensing of Environment 74(2):229–239. https://doi.org/10.1016/S0034-4257(00)00113-9

[8] Delegido J, Verrelst J, Meza CM, et al (2013) A red-edge spectral index for remote sensing estimation of green LAI over agroecosystems. European Journal of Agronomy 46:42–52. https://doi.org/10.1016/j.eja.2012.12.001

[9] Farhad MM, et al (2024) Preliminary results from three years of uas-based gnss-r field campaign over agricultural fields for field-scale soil moisture retrieval. In: IGARSS 2024. IEEE, pp 4477–4481, https://doi.org/10.1109/IGARSS53475.2024.10641343

[10] Gitelson AA, Merzlyak MN (1998) Remote estimation of chlorophyll content in higher plant leaves. International Journal of Remote Sensing 19(14):2699–2711. https://doi.org/10.1080/014311698214929

[11] Guo Q, Han B, Chu P, et al (2025) MF-FusionNet: A lightweight multimodal network for monitoring drought stress in winter wheat based on remote sensing imagery. Agriculture 15(15):Article 1639. https://doi.org/10.3390/agriculture15151639

[12] Haboudane D, Miller JR, Tremblay N, et al (2002) Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. Remote Sensing of Environment 81(2-3):416–426. https://doi.org/10.1016/S0034-4257(02)00018-4

[13] Han X, Wei Z, Chen H, et al (2021) Inversion of winter wheat growth parameters and yield under different water treatments based on UAV multispectral remote sensing. Frontiers in Plant Science 12:Article 609876. https://doi.org/10.3389/fpls.2021.609876

[14] He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778, https://doi.org/10.1109/CVPR.2016.90

[15] Howard A, Sandler M, Chu G, et al (2019) Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),

pp 1314–1324, https://doi.org/10.1109/ICCV.2019.00140

[16] Huete A, Justice C, van Leeuwen W (1999) MODIS vegetation index (MOD 13). Algorithm theoretical basis document (atbd) mod13, University of Virginia

[17] Jacquemoud S, Verhoef W, Baret F, et al (2009) PROSPECT+ SAIL models: a review of use for vegetation characterization. Remote Sensing of Environment 113:S56–S66. https://doi.org/10.1016/j.rse.2008.01.026

[18] Kempeneers P, Claverie M, D'Andrimont R (2023) Using a vegetation index as a proxy for reliability in surface reflectance time series reconstruction. Remote Sensing 15(9):Article 2303. https://doi.org/10.3390/rs15092303

[19] Liu M, He W, Zhang H (2026) CN_Wheat10: a 10 m resolution dataset of spring and winter wheat distribution in China (2018–2024) derived from time-series remote sensing. Earth System Science Data 18(1):465–491. https://doi.org/10.5194/essd-18-465-2026

[20] Liu W, Zhu X, Yang S, et al (2025) Physics-driven machine-learning retrieval and uncertainty quantification of crop leaf area index. Remote Sensing 17(23):Article 3924. https://doi.org/10.3390/rs17233924

[21] Louhaichi M, Borman MM, Johnson DE (2001) Spatially located platform and aerial photography for documentation of grazing impacts on wheat. Geocarto International 16(1):65–70. https://doi.org/10.1080/10106040108542184

[22] Ma N, Zhang X, Zheng HT, et al (2018) ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 116–131, https://doi.org/10.1007/978-3-030-01264-9_8

[23] Ma X, Dai X, Bai Y, et al (2024) Rewrite the stars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 5694–5703, https://doi.org/10.1109/CVPR52733.2024.00539

[24] Oguntunde PG, Olukunle OJ, Fasinmirin JT, et al (2012) Performance of the SunScan canopy analysis system in estimating leaf area index of maize. Agricultural Engineering International: CIGR Journal 14(3):1–7

[25] Peng Z, Huang Y, Xu Z, et al (2025) Star with bilinear mapping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 25292–25302

[26] Rondeaux G, Steven M, Baret F (1996) Optimization of soil-adjusted vegetation indices. Remote Sensing of Environment 55(2):95–107. https://doi.org/10.1016/0034-4257(95)00186-7

[27] Rouse JW, Haas RH, Schell JA, et al (1974) Monitoring vegetation systems in the Great Plains with ERTS. NASA Special Publication 351:309

[28] Sone C, Saito K, Futakuchi K (2009) Comparison of three methods for estimating leaf area index of upland rice cultivars. Crop Science 49(4):1438–1443. https://doi.org/10.2135/cropsci2008.07.0425

[29] Tan M, Le QV (2019) EfficientNet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning (ICML), pp 6105–6114, https://doi.org/10.48550/arXiv.1905.11946

[30] Tang J, Xu Y, Liu Y, et al (2025) Shaping initial state prevents modality competition in multi-modal fusion: A two-stage scheduling framework via fast partial information decomposition. arXiv preprint arXiv:250920840 https://doi.org/10.48550/arXiv.2509.20840

[31] Verma V, Lamb A, Beckham C, et al (2019) Manifold mixup: Better representations by interpolating hidden states. In: International Conference on Machine Learning (ICML), pp 6438–6447, https://doi.org/10.48550/arXiv.1806.05236

[32] Wu S, Deng L, Guo L, et al (2022) Wheat leaf area index prediction using data fusion based on high-resolution unmanned aerial vehicle imagery. Plant Methods 18(1):68. https://doi.org/10.1186/s13007-022-00896-8

[33] Wu Y, Sicard B, Gadsden SA (2024) Physics-informed machine learning: A comprehensive review on applications in anomaly detection and condition monitoring. Expert Systems with Applications 255:Article 124678. https://doi.org/10.1016/j.eswa.2024.124678

[34] Yao H, Wang Y, Li S, et al (2022) C-mixup: Improving generalization in regression. In: Advances in Neural Information Processing Systems (NeurIPS), pp 33645–33658, https://doi.org/10.48550/arXiv.2210.05775

[35] Yin Q, Yu X, Li Z, et al (2024) Estimating summer maize biomass by integrating UAV multispectral imagery with crop physiological parameters. Plants 13(21):Article 3070. https://doi.org/10.3390/plants13213070

[36] Yu P, Teng F, Zhu W, et al (2025) Cloud–edge–device collaborative computing in smart agriculture: Architectures, applications, and future perspectives. Frontiers in Plant Science 16:Article 1668545. https://doi.org/10.3389/fpls.2025.1668545

[37] Zhang H, Cisse M, Dauphin YN, et al (2018) mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (ICLR), https://doi.org/10.48550/arXiv.1710.09412

[38] Zhang L, Zhang B, Zhang H, et al (2025) Multi-source feature fusion network for LAI estimation from UAV multispectral imagery. Agronomy 15(4):Article 988.

https://doi.org/10.3390/agronomy15040988

[39] Zhao D, Yang G, Xu T, et al (2025) Dynamic maize true leaf area index retrieval with KGCNN and TL and integrated 3D radiative transfer modeling for crop phenotyping. Plant Phenomics 7(1):Article 100004. https://doi.org/10.1016/j.plaphe.2025.100004

[40] Zou M, Liu Y, Fang M, et al (2023) Combining spectral and texture feature of UAV image with plant height to improve LAI estimation of winter wheat at jointing stage. Frontiers in Plant Science 14:1272049. https://doi.org/10.3389/fpls.2023.1272049

[41] Zu J, Yang H, Wang J, et al (2024) Inversion of winter wheat leaf area index from UAV multispectral images: Classical vs. deep learning approaches. Frontiers in Plant Science 15:Article 1367828. https://doi.org/10.3389/fpls.2024.1367828