agriculture

MDPI

Article

An Artificial Intelligence-Driven UAV and Ground Sensor Fusion Framework for Crop Growth Assessment in Smart Agriculture

Puxing Gao [1], Keyue Wang [2,3], Yunuo Li [3], Jiayue Zhang [1], Qingyu Li [3], Wenjie Lu [1] and Yihong Song [1,2,3,*]

[1] China Agricultural University, Beijing 100083, China
[2] Beijing Forestry University, Beijing 100083, China
[3] Peking University, Beijing 100871, China
* Correspondence: songyihong@cau.edu.cn

Abstract

With the rapid development of artificial intelligence, UAV remote sensing, and agricultural Internet of Things technologies, crop growth monitoring is evolving from manual inspection and single-source analysis toward intelligent decision-making based on multisource perception. However, existing methods still suffer from limited robustness under environmental variations, insufficient integration between UAV imagery and sparse ground sensor observations, and weak capability for transforming predictions into practical agricultural management recommendations. This study proposes a UAV–ground sensor collaborative lightweight framework for crop growth assessment and agricultural decision support. The proposed framework integrates UAV RGB and multispectral imagery with ground sensor observations through a region-level aerial–ground alignment mechanism and a sensor-guided attention fusion module, enabling environmental conditions to enhance visual feature interpretation. Furthermore, a fact-constrained decision module is developed to generate management recommendations based on crop status, environmental risks, and field information. Experimental results demonstrate that the proposed method achieves superior performance in crop growth classification and yield-trend prediction, reaching Accuracy, Precision, Recall, and F1-score values of 92.47%, 91.86%, 91.39%, and 91.62%, respectively, with an RMSE of 0.381 and an R^2 of 0.902. The lightweight framework requires only 6.18M parameters and 0.91G FLOPs, achieving 39.56 ms inference latency and 25.28 FPS on edge devices. The proposed framework also improves decision reliability, achieving an expert agreement rate of 89.34% and a risk identification accuracy of 90.18%. Economic analysis indicates that the proposed framework reduces labor cost, water consumption, and fertilizer input by 49.7%, 26.7%, and 23.0%, respectively, while increasing net benefit by 46.1% compared with conventional field management practices. These results demonstrate that the proposed method provides an accurate, interpretable, and deployable AI-driven solution for intelligent crop management in smallholder and medium-sized farming systems.

Keywords: smart agriculture; artificial intelligence in agriculture; precision agriculture; agricultural economics; agricultural input–output efficiency

Check for updates

Received: 28 May 2026
Revised: 27 July 2026
Accepted: 29 July 2026
Published: 31 July 2026

1. Introduction

With global population growth, declining agricultural labor, and increasing extreme climate events, agricultural production is facing challenges such as resource constraints and insufficient field management [1]. Crop growth monitoring is a fundamental component of precision agriculture, directly affecting irrigation, fertilization, pest warning, and yield formation [2]. Traditional manual field inspection methods are subjective, inefficient, and unable to provide continuous monitoring, especially in large-scale farmland [3,4]. Therefore, developing intelligent, automated, and low-cost crop monitoring systems is of great importance. In recent years, the advancement of UAV remote sensing, agricultural IoT, and artificial intelligence has provided new solutions for crop monitoring [5]. UAVs can rapidly acquire high-resolution farmland images to reflect canopy characteristics and crop growth conditions [6], while ground sensors continuously collect environmental data such as soil moisture, temperature, humidity, light intensity, and electrical conductivity [7]. The integration of "aerial remote sensing + ground sensing" enables more comprehensive monitoring of crops and their environments, becoming an important direction in smart agriculture [8]. Existing crop growth monitoring methods still have significant limitations [9]. Vegetation-index-based and traditional machine learning approaches are sensitive to illumination, weather, and environmental variations [10], while their ability to represent complex spatial textures and nonlinear features remains limited [11]. Ground-sensor-based methods can continuously monitor environmental conditions, but they lack effective spatial characterization of field-wide crop variability [12]. As a result, relying on a single data source makes it difficult to achieve stable, accurate, and interpretable crop growth assessment [13].

With the development of deep learning, methods based on convolutional neural networks (CNNs), Transformers, and multimodal fusion have become major research directions in smart agriculture [14]. Models such as ResNet, ViT, and Swin Transformer can automatically extract complex canopy features from UAV imagery and have been widely applied to crop classification, growth assessment, and yield prediction [15,16]. Researchers have also begun integrating UAV remote sensing with environmental sensor data to improve agricultural state recognition performance [17,18]. However, several challenges remain: (1) differences in spatial and temporal scales between multimodal data make collaborative modeling difficult; (2) existing models often lack interpretability regarding the relationship between environmental risks and crop abnormalities; and (3) large deep learning models require substantial computational resources, limiting deployment on edge devices in real agricultural scenarios. Recent studies on 3D crop phenotyping, edge-computing UAV systems, and intelligent autonomous perception have demonstrated the potential of AI-enabled UAV technologies in precision agriculture, providing a technical foundation for future intelligent crop monitoring systems.

To address the above challenges, a UAV–ground sensor collaborative lightweight crop growth assessment and decision support framework is proposed for smart agriculture scenarios. The proposed framework aims to achieve three objectives: (1) establishing effective aerial–ground multimodal integration for robust crop growth assessment; (2) improving the interpretability of growth recognition through environmental-aware feature fusion; and (3) enabling practical agricultural decision support through lightweight and deployable AI models. UAV RGB and multispectral images are used to extract crop visual characteristics, while ground sensors provide environmental state representations. A region-level aerial–ground alignment mechanism and a sensor-guided attention fusion module are designed to integrate spatial crop information with temporal environmental conditions. Furthermore, a lightweight agricultural decision service module combines model predictions, environmental risks, and agricultural knowledge rules to generate management recommendations while supporting edge deployment.

The main contributions of this study can be summarized as follows:

1. A lightweight aerial–ground collaborative multimodal framework is proposed to jointly model UAV remote sensing images and ground sensor time series, improving the robustness of crop growth assessment under complex environmental variations.
2. A region-level aerial–ground alignment mechanism and a sensor-guided attention fusion strategy are developed to incorporate environmental information into visual feature learning, enhancing multimodal interpretation and model explainability.
3. A sensor-fact-constrained agricultural decision service module is constructed to establish a closed-loop process from crop state recognition to management recommendation, with edge deployment experiments validating its practical feasibility.
4. The proposed framework is further evaluated from an agricultural economic perspective, demonstrating its potential to improve resource utilization efficiency and reduce management costs in practical farming scenarios.

2. Related Work

2.1. UAV Remote Sensing for Crop Growth Monitoring

UAV remote sensing technology has been widely applied to agricultural tasks such as crop growth monitoring, pest and disease detection, and yield prediction [19]. Compared with traditional satellite remote sensing, UAVs provide higher spatial resolution, more flexible data acquisition, and lower operational costs, making them more suitable for precision agriculture in complex farmland environments [20]. Using RGB and multispectral imagery, canopy color, texture, coverage, and vegetation indices such as NDVI, GNDVI, and NDRE can be extracted to evaluate crop growth conditions, physiological status, and yield trends [21], while also enabling estimation of indicators such as leaf area index, canopy coverage, and plant height. With the advancement of deep learning, methods based on CNNs and Transformers have significantly improved feature representation in UAV remote sensing [22]. Models such as ResNet, EfficientNet, MobileNet, ViT, and Swin Transformer can automatically learn complex canopy textures, multi-scale structural features, and long-range spatial dependencies, achieving strong performance in crop classification, growth assessment, pest identification, and yield prediction [23]. However, most existing approaches still rely on a single visual modality and are sensitive to illumination changes, shadow occlusion, weather conditions, and phenological variations, which reduces model robustness [24]. In addition, image-only methods lack the ability to explain environmental causes behind crop growth variations [25]. Although abnormal growth regions can be detected, it remains difficult to determine whether the stress is caused by water deficiency, salinity, high temperature, or nutrient shortage, limiting the effectiveness of current UAV remote sensing methods in agricultural decision support [26].

2.2. Ground Sensor-Based Agricultural Monitoring and Multimodal Fusion

Ground sensors and agricultural IoT technologies mainly focus on continuous monitoring of farmland environments [27]. Distributed sensors are used to collect real-time data such as soil moisture, temperature, air humidity, light intensity, EC, pH, rainfall, and wind speed [28]. These environmental variables are closely related to crop growth and can explain growth variations from an environmental perspective [29], thereby supporting irrigation management, fertilization regulation, and environmental stress identification [30]. To obtain more stable and comprehensive agricultural state representations, researchers have increasingly integrated UAV remote sensing imagery with environmental sensor data through multimodal fusion [31]. Existing methods mainly employ feature concatenation, temporal modeling, and attention mechanisms for collaborative analysis of visual and environmental information [32]. For example, combining soil moisture and meteorological data with UAV imagery can improve crop growth recognition accuracy [33], while Transformer-based and cross-modal attention methods enhance the perception of dynamic environmental changes and key regions [34]. However, several challenges remain in current multimodal approaches. UAV remote sensing data are spatially continuous, whereas ground sensor data are discretely distributed at point locations, resulting in significant differences in spatial scale and temporal frequency that make stable association difficult [35]. In addition, limited sensor coverage makes it challenging to map point-based environmental information to region-level image features [36]. Existing studies also focus more on improving recognition accuracy than on interpreting the relationship between environmental risks and visual abnormalities, which limits their practical application in agricultural management decision-making [37].

2.3. Lightweight AI-Based Decision Support Systems in Smart Agriculture

Lightweight intelligent agricultural decision support systems aim to perform agricultural state analysis, risk warning, and management decision-making on low-cost edge devices [38], thereby improving the deployability of smart agriculture systems in real farmland environments [39]. Since agricultural scenarios are often constrained by poor network conditions, limited computing resources, and cost-sensitive devices, traditional large-scale deep learning models are difficult to deploy directly [40]. To address this issue, researchers have introduced lightweight networks, model compression, and edge computing technologies [41]. Models such as MobileNet, ShuffleNet, and EfficientNet-lite reduce computational cost through optimized architectures, while pruning, quantization, and knowledge distillation further decrease storage and inference overhead, enabling deployment on low-power devices such as Jetson Nano and Raspberry Pi [42]. At the same time, agricultural knowledge graphs, rule bases, and language models have been increasingly applied to agricultural knowledge services and intelligent question-answering systems [43], providing recommendations for irrigation, fertilization, and pest management [44]. This reflects the transition of agricultural AI systems from simple recognition toward decision support. However, current systems still face important limitations [45]. Some systems only output recognition results without actionable management recommendations, while others can generate agricultural suggestions but lack constraints from real field data, leading to recommendations that may not match actual environmental conditions [46]. As a result, existing intelligent agricultural systems still lack a complete closed loop of "perception–interpretation–decision" [47].

2.4. Agricultural Economics and Decision Support in Smart Agriculture

From the perspective of agricultural economics, the value of smart agriculture technologies is reflected not only in improved crop recognition accuracy but also in enhanced production-factor allocation efficiency, reduced decision-making costs for farmers, and improved input–output relationships. Traditional farmland management often relies on farmers' experience to arrange irrigation, fertilization, and field inspection, and is easily affected by insufficient information, delayed risk perception, and resource constraints. These factors may further lead to excessive water and fertilizer inputs, labor waste, or delayed management during critical growth stages. In recent years, increasing attention has been paid to the effects of digital agriculture, precision agriculture, and intelligent decision support systems on agricultural production efficiency. It has been suggested that sensor monitoring, remote sensing assessment, and intelligent recommendation can reduce the cost of obtaining field information for farmers and increase the marginal returns of agricultural inputs. However, existing studies have mostly been conducted from the perspectives of technical performance or macroeconomic benefits, while relatively limited attention has been paid to linking crop growth recognition results, environmental risk interpretation, and specific agricultural operation recommendations. Therefore, agricultural economic concepts need to be incorporated into intelligent crop growth assessment models, so that such models can not only determine crop status but also support optimized allocation of water and fertilizer resources, management cost control, and technology adoption by smallholders.

3. Materials and Method

3.1. Data Collection

Data collection was conducted from May 2025 to April 2026 in Wuyuan County, Bayannur City, Inner Mongolia, China, within the Hetao Irrigation District. The study crop was spring maize (*Zea mays* L., cultivar Xianyu 335). The experimental area contained 16 plots, each approximately 30 m $\times$ 20 m, with a total area of about 0.96 ha. Maize was planted at a row spacing of 0.60 m and an average plant spacing of 0.25 m, corresponding to approximately 66,700 plants ha^{-1}. The soil was classified as silty loam, with a mean pH of 8.12, soil electrical conductivity of 1.21 dS m^{-1}, and soil organic matter content of 14.8 g kg^{-1}. Approximately 300 kg N ha^{-1}, 90 kg P_2O_5 ha^{-1}, and 60 kg K_2O ha^{-1} were applied during the growing season. Four main irrigation events were conducted at the six-leaf, tasseling, silking, and grain-filling stages, with a total irrigation amount of approximately 4500 m^3 ha^{-1}. An overview of the acquisition equipment, sampling design, and final dataset scale is provided in Table 1.

Table 1. Composition, acquisition equipment, and data scale of the crop growth assessment dataset.

Data Type	Acquisition Equipment and Sampling Design	Data Scale
UAV RGB images	DJI Mavic 3 Multispectral UAV; 15 flights at 60 m altitude; 80% forward overlap and 75% side overlap	3840 images; 768 images per growth stage
Multispectral images and vegetation indices	DJI Mavic 3 Multispectral camera; green, red, red-edge, and near-infrared bands; NDVI, GNDVI, and NDRE were calculated	3840 image sets; 768 sets per growth stage
Air temperature and humidity	HOBO MX2301A sensors; 12 monitoring nodes; 30-min sampling interval	14,400 records
Soil moisture, temperature, and EC	TEROS 12 sensors installed at depths of 20 and 40 cm; 12 monitoring nodes	28,800 records
Photosynthetically active radiation	Apogee SQ-500 sensors; 12 monitoring nodes	14,400 records
Rainfall and wind speed	Spectrum WatchDog 2000 Series mini weather station	14,400 records for each variable
Manual agronomic observations	16 plots with 16 fixed quadrats per plot; five growth stages; independent evaluation by two technicians and adjudication by one senior agronomist	1280 samples; 256 samples per growth stage
Field-management records	Irrigation, fertilization, pest control, field inspection, and abnormal-event logs	420 records

UAV images were acquired using a DJI Mavic 3 Multispectral UAV during five major growth stages: seedling, rapid growth, flowering, grain filling, and maturity, as shown in Figure 1. A total of 15 flights were conducted, with three flights per growth stage. The UAV was operated at an altitude of 60 m, with 80% forward overlap and 75% side overlap, producing ground sampling distances of approximately 1.6 cm pixel^{-1} for RGB images and 3.2 cm pixel^{-1} for multispectral images. Flights were conducted between 10:00 and 12:00 under clear or slightly cloudy conditions and low wind speed. After quality control, 3840 RGB images and 3840 corresponding multispectral image sets were retained, including 768 samples from each growth stage. The multispectral bands were used to calculate NDVI, GNDVI, and NDRE.

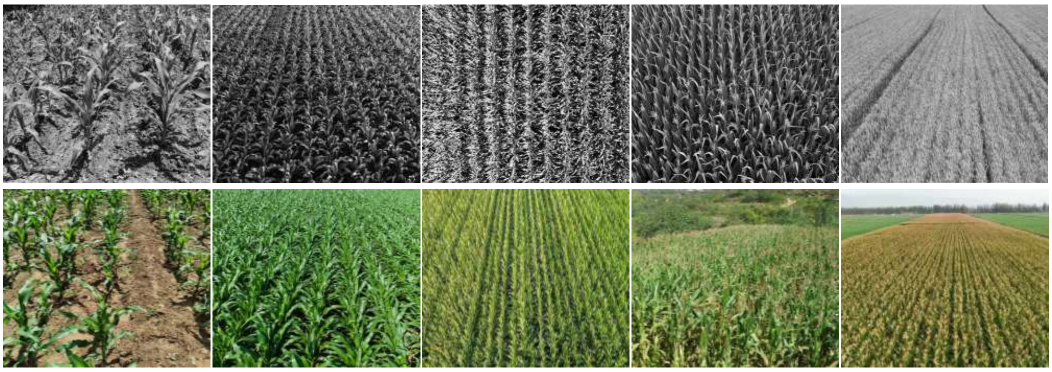

Figure 1. Representative UAV-based multispectral maize canopy images collected at different growth stages.

Twelve ground monitoring nodes were deployed across the experimental area. Their locations covered field centers, plot edges, irrigation inlets, low-lying areas, and regions with evident salinity or growth differences. TEROS 12 sensors were installed at depths of 20 and 40 cm to measure soil moisture, soil temperature, and EC. HOBO MX2301A sensors measured air temperature and humidity above the canopy, while Apogee SQ-500 sensors measured photosynthetically active radiation. Rainfall and wind speed were recorded using a Spectrum WatchDog 2000 Series mini weather station. Sensor data were collected at 30 min intervals, transmitted wirelessly, and stored locally as a backup. After quality control and temporal synchronization, 28,800 soil moisture, temperature, and EC records and 14,400 records for each meteorological variable were retained.

Manual agronomic observations were conducted within 3 h of UAV acquisition. Sixteen fixed 1 m × 1 m quadrats were established in each plot, resulting in 256 quadrats per growth stage and 1280 observations in total. Plant height was calculated as the mean of ten representative plants in each quadrat. Canopy coverage was estimated from nadir quadrat photographs, while leaf area index (LAI), above-ground biomass, leaf color, growth uniformity, and pest or disease symptoms were also recorded. Harvest-stage yield was measured in the corresponding sampling areas and corrected to a standard grain moisture content of 14%.

To obtain reproducible crop-growth labels, the four growth grades were assigned using a stage-specific composite evaluation protocol rather than visual judgement alone. Plant height, canopy coverage, LAI, and above-ground biomass were first normalized within each growth stage relative to the stage-specific reference values obtained from normally growing quadrats. Leaf color was rated on a five-point scale, where 1 represented severe chlorosis and 5 represented a uniformly dark-green canopy. Growth uniformity was also rated on a five-point scale. Pest, disease, lodging, and other visible stress symptoms were recorded as the percentage of affected plants within each quadrat.

A composite crop-growth score, S_g, was calculated as

$$S_g = 0.20S_{\mathrm{PH}} + 0.25S_{\mathrm{CC}} + 0.20S_{\mathrm{LAI}} + 0.20S_{\mathrm{BM}} + 0.10S_{\mathrm{LC}} + 0.05S_{\mathrm{GU}} - P_{\mathrm{stress}}, \quad (1)$$

where S_{PH}, S_{CC}, S_{LAI}, S_{BM}, S_{LC}, and S_{GU} denote the normalized scores for plant height, canopy coverage, LAI, biomass, leaf color, and growth uniformity, respectively. All component scores were converted to a 0–100 scale. The stress penalty P_{stress} ranged from 0 to 15 points and was determined from the incidence and severity of pest, disease, lodging, chlorosis, wilting, and other visible stress symptoms. Yield or yield components were used as supplementary evidence at the maturity stage but were not used to label samples from earlier growth stages.

The final growth grades were defined as follows:

- *Weak*: $S_g < 50$, generally characterized by substantial reductions in plant height, canopy coverage, LAI, or biomass, together with evident chlorosis, wilting, lodging, or pest and disease symptoms;
- *Moderate*: $50 \leq S_g < 65$, representing below-average growth with sparse or uneven canopy development and mild-to-moderate stress symptoms;
- *Good*: $65 \leq S_g < 80$, representing normal crop development, relatively uniform canopy structure, and no severe stress symptoms;
- *Excellent*: $S_g \geq 80$, representing vigorous and highly uniform growth, high canopy coverage and biomass, healthy leaf color, and negligible visible stress.

Two agricultural technicians independently evaluated all 1280 samples without access to each other's labels or to the model predictions. Their initial agreement was 91.3%, and the weighted Cohen's κ coefficient was 0.87, indicating strong inter-rater agreement. Samples with inconsistent labels were re-examined by a senior agronomist using the original field measurements and quadrat photographs, and the final label was determined by consensus. The resulting class distribution is reported in Table 2.

Table 2. Distribution of the manually labeled crop-growth grades.

Growth Grade	Number of Samples	Percentage (%)
Weak	246	19.22
Moderate	331	25.86
Good	427	33.36
Excellent	276	21.56
Total	1280	100.00

Field-management records included irrigation, fertilization, pest control, field inspection, and abnormal events. A total of 420 validated records were retained. UAV images, sensor windows, manual observations, and management records were matched using plot identifiers, region identifiers, geographic coordinates, and acquisition time. The composition and scale of all data sources are summarized in Table 1.

3.2. Data Preprocessing and Augmentation

UAV imagery and ground-sensor time series were processed separately before region-level multimodal sample construction. All preprocessing parameters and normalization statistics were determined using the training set and were applied unchanged to the validation and test sets. Data augmentation was used only during model training.

3.2.1. UAV Image Preprocessing and Calibration

Raw UAV images were processed using DJI Terra 3.9.4. The DJI Mavic 3 Multispectral UAV was operated with the RTK positioning module connected to a DJI D-RTK 2 mobile base station. When a stable RTK fixed solution was available, the geotagged image positions had centimeter-level horizontal accuracy. Image records without a fixed RTK solution were retained only when sufficient ground control was available for geometric correction.

Before the first flight, ten ground targets were distributed across the experimental area, including the four corners, field edges, and interior locations. Their coordinates were measured using an RTK-GNSS receiver. Six targets were used as ground control points for aerial triangulation, whereas the remaining four were used as independent checkpoints and were not included in model fitting. The same target locations were maintained throughout the acquisition period whenever field conditions permitted.

Image geotags, RTK positioning information, camera calibration parameters, and ground control points were jointly used for aerial triangulation, orthorectification, and mosaicking. A dense point cloud and digital surface model were first reconstructed, after which RGB and multispectral orthomosaics were generated in the WGS 84/UTM Zone 49N coordinate system. The mean checkpoint root-mean-square errors were approximately 2.8 cm in the horizontal direction and 4.1 cm in the vertical direction. Orthomosaics with a horizontal checkpoint error greater than 5 cm were reprocessed or excluded.

Radiometric calibration was performed for every multispectral flight. A manufacturer-compatible Lambertian reflectance panel with calibrated reflectance values for the green, red, red-edge, and near-infrared bands was imaged immediately before and after each flight under the same exposure settings used for field acquisition. The panel occupied at least 200×200 pixels in the calibration images, and saturated or shadowed panel images were discarded. The multispectral camera's onboard sunlight sensor simultaneously recorded incident irradiance during flight.

Raw digital numbers were first corrected using the exposure time, sensor gain, and band-specific calibration coefficients stored in the image metadata. Band-wise conversion coefficients were then derived from the measured digital values of the reflectance panel and its certified reflectance values. The pre-flight and post-flight coefficients were averaged when their difference was below 5%. When the difference exceeded 5%, time-dependent linear interpolation was used to compensate for illumination changes during the flight. The irradiance measurements from the onboard sunlight sensor were subsequently used to correct short-term variation in incoming solar radiation. This procedure converted the multispectral images to unitless surface reflectance values in the range of 0–1.

The RGB and calibrated multispectral orthomosaics were co-registered using RTK coordinates and ground control points and were resampled to a common spatial resolution of 3.2 cm pixel^{-1}. Because the four multispectral cameras have separate optical centers, residual inter-band displacement was corrected using band-to-band feature matching, with the red band used as the reference. Registration accuracy was evaluated at 20 manually selected control locations distributed across the field. Image sets with a residual displacement greater than two pixels were corrected through local affine registration; sets that still exceeded this threshold were excluded.

NDVI, GNDVI, and NDRE were calculated from the calibrated reflectance orthomosaics rather than from raw digital numbers:

$$\mathrm{NDVI} = \frac{\rho_{\mathrm{NIR}} - \rho_{\mathrm{Red}}}{\rho_{\mathrm{NIR}} + \rho_{\mathrm{Red}}}, \qquad \mathrm{GNDVI} = \frac{\rho_{\mathrm{NIR}} - \rho_{\mathrm{Green}}}{\rho_{\mathrm{NIR}} + \rho_{\mathrm{Green}}}, \tag{2}$$

$$\mathrm{NDRE} = \frac{\rho_{\mathrm{NIR}} - \rho_{\mathrm{RedEdge}}}{\rho_{\mathrm{NIR}} + \rho_{\mathrm{RedEdge}}}, \tag{3}$$

where ρ denotes the radiometrically calibrated reflectance of the corresponding band. Pixels for which the denominator was below 10^{-6} were marked as invalid. Index values outside the physical range of $[-1, 1]$ were also treated as invalid.

Plot boundaries surveyed by RTK-GNSS were used to remove roads, irrigation canals, field margins, and other non-crop regions. The valid plot areas were divided into 224×224-pixel image patches with a stride of 112 pixels. Patches in which vegetation occupied less than 60% of the area were discarded. Vegetation pixels were identified using an NDVI threshold of 0.20 for multispectral data. For RGB-only samples, the excess-green index was used, and pixels with ExG > 0.05 were treated as vegetation. Patches near plot boundaries or containing strong shadow, motion blur, saturation, panel targets, or substantial non-crop objects were manually reviewed.

The model input included the RGB channels, calibrated green, red, red-edge, and near-infrared reflectance bands, and the derived NDVI, GNDVI, and NDRE maps. Each channel was normalized using the mean and standard deviation calculated exclusively from the training set. The same normalization statistics were applied to the validation and independent test sets.

3.2.2. Sensor Quality Control and Temporal Alignment

Ground sensors recorded environmental variables at 30 min intervals. Sensor values outside the physical operating ranges specified by the manufacturers were first removed. Remaining outliers were detected independently for each sensor variable using a Hampel filter with a centered window of seven observations and a threshold of three median absolute deviations.

Missing sequences of no more than six consecutive observations, corresponding to a maximum duration of 3 h, were completed by linear interpolation. Longer missing intervals were not imputed; instead, the corresponding values were marked using a binary missing-value mask. A sample was excluded when more than 20% of the observations for any required sensor variable were missing within its temporal window.

For each UAV acquisition at time t_u, sensor observations from the preceding 72 h were selected, resulting in 144 time steps per sensor variable. The 72 h interval was selected based on preliminary experiments because it retained short-term changes in soil moisture, temperature, salinity, rainfall, and radiation while avoiding excessive temporal redundancy. Each temporal window included soil moisture, soil temperature, EC, air temperature, relative humidity, photosynthetically active radiation, rainfall, and wind speed.

Sensor variables were standardized using the training-set mean and standard deviation. In addition to the complete temporal sequence, four summary statistics were calculated for each variable: mean, minimum, maximum, and linear trend over the 72 h period. Sensor observations were associated with each image region using the recorded sensor coordinates, UAV acquisition time, plot identifier, and regional identifier.

3.3. Proposed Method

3.3.1. Overall

A UAV–ground sensor collaborative lightweight crop growth assessment and agricultural decision-making framework is proposed. The overall process starts from the processed region-level multimodal sample input, and sequentially performs visual feature extraction, sensor temporal encoding, region-level aerial–ground alignment, sensor-guided remote sensing attention fusion, and fact-constrained agricultural decision generation. Specifically, the preprocessed UAV image patches, vegetation-index maps, ground sensor temporal-window features, and plot management records are first organized into unified region-level input units. The UAV image patches and vegetation-index maps are fed into a lightweight visual encoding branch to extract crop canopy color, texture, coverage, vegetation-index response, and spatial heterogeneity features. Meanwhile, sensor sequences within the corresponding temporal window, including soil moisture, air temperature and humidity, light intensity, EC, and rainfall, are fed into a temporal encoding branch to obtain sensor representations that reflect local environmental states and potential stress risks. Subsequently, the region-level aerial–ground alignment module maps the originally sparse point-based sensor information into environmental contexts corresponding to image regions according to the position of each image region, the distribution of sensor points, and the acquisition-time relationship. In this way, each regional sample simultaneously contains remote sensing visual information and spatiotemporally matched environmental information. On this basis, the sensor-guided remote sensing attention fusion module further uses environmental risk representations to dynamically modulate visual features, enabling the model to adaptively enhance visual channels and spatial regions associated with growth degradation under different environmental states, such as water deficit, high temperature, salinity stress, or abnormal illumination, rather than simply concatenating image features and sensor features. The fused multimodal regional representations are then fed into the growth classification head and the trend regression head to output regional growth grades, risk confidence scores, and yield or biomass variation trends. Finally, the sensor-fact-constrained lightweight agricultural decision service module receives model prediction results, key risk factors, and management records, and combines them with agronomic knowledge rules to generate irrigation, fertilization, field inspection, and risk-warning recommendations. Candidate recommendations are subjected to fact-consistency verification in this module, ensuring that recommendations are constrained by current sensor states and model judgments, thereby avoiding generalized suggestions that deviate from actual field conditions. Through the above process, an end-to-end closed loop from regional perception and multimodal fusion to growth assessment and agricultural decision-making is formed.

3.3.2. Region-Level Aerial–Ground Alignment Module

The objective of the region-level aerial–ground alignment module is to unify the spatially continuous regions in UAV remote sensing images and the point-based temporal observations from ground sensors into the same modeling unit, thereby providing reliable regional environmental contexts for subsequent sensor-guided fusion, as shown in Figure 2. Let the processed UAV image be divided into N regional image patches. The visual input corresponding to the i-th region is denoted as P_i, and the center coordinate of this region is denoted as c_i. A total of M ground sensor points are deployed in the farmland, and the coordinate of the j-th sensor is denoted as l_j. Its multivariate sequence within the temporal window around the UAV acquisition time is denoted as $S_j \in \mathbb{R}^{T \times C_s}$, where T denotes the temporal-window length and C_s denotes the dimensionality of sensor variables. A lightweight temporal encoder is first used to encode the time series of each sensor point. This encoder can be composed of a one-dimensional convolutional layer, a gated temporal unit, and a linear projection layer. The one-dimensional convolution is used to capture short-term environmental fluctuations, the gated unit is used to model the continuous variation trends of variables such as soil moisture, air temperature and humidity, light intensity, and EC, and the linear projection layer maps temporal representations from different sensor points into a unified dimension. This process can be expressed as $h_j = f_s(S_j; \theta_s)$, where $h_j \in \mathbb{R}^{d_s}$ denotes the environmental state representation of the j-th sensor point, $f_s(\cdot)$ denotes the shared-parameter sensor temporal encoder, and θ_s denotes its learnable parameters. The shared encoder design avoids constructing separate models for different points, thereby reducing the number of parameters and ensuring the comparability of output representations from different sensor points.

After point-level sensor representations are obtained, soft alignment weights from regions to sensors are further constructed. Since the environmental state of a region is not only related to the nearest sensor but may also be jointly affected by multiple surrounding sensors, hard matching is not adopted. Instead, normalized weights are jointly calculated according to spatial distance, temporal proximity, and sensor reliability.

For the i-th region and the j-th sensor, the unnormalized correlation can be expressed as $e_{ij} = \psi_d(c_i, l_j)\psi_t(\Delta t_j)\psi_q(q_j)$, where $\psi_d(c_i, l_j) = \exp(-d(c_i, l_j)/\tau_d)$ denotes the spatial distance decay function, $\psi_t(\Delta t_j) = \exp(-\Delta t_j/\tau_t)$ denotes the temporal-offset decay function, $\psi_q(q_j)$ denotes the sensor quality or availability mapping function, $d(c_i, l_j)$ denotes the spatial distance, Δt_j denotes the temporal deviation between the sensor window and the UAV acquisition time, and τ_d and τ_t are scale control parameters for spatial and temporal distances, respectively. The region-level normalized weight $\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{m=1}^{M}\exp(e_{im})}$ is then obtained through Softmax, and the sensor representations from multiple points are aggregated into the environmental context of the i-th region as $E_i = \sum_{j=1}^{M}\alpha_{ij}h_j$. The mathematical significance of this soft alignment strategy is that discrete point-based observations are transformed into a weighted estimation of regional environmental states, enabling sensor information to be smoothly propagated in space without losing local differences.

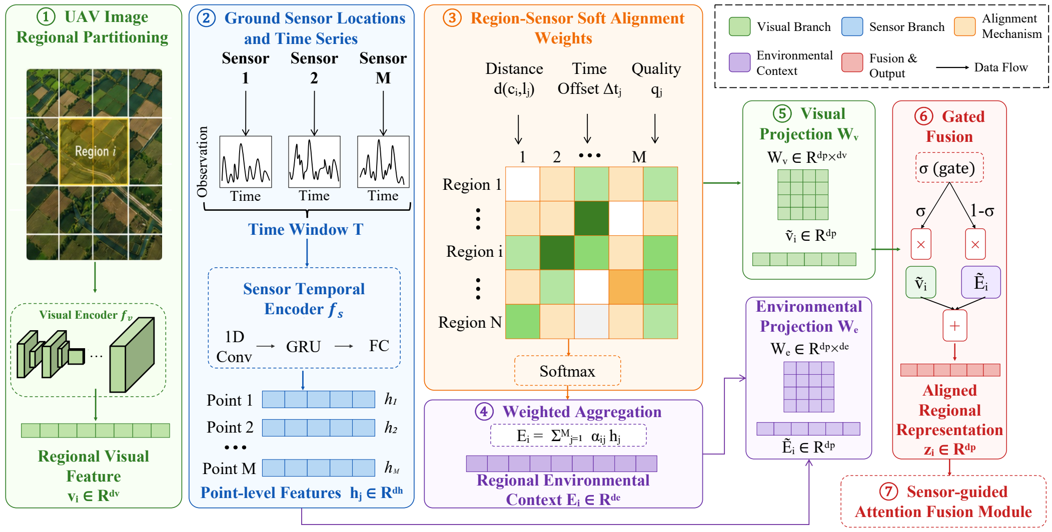

Figure 2. The region-level aerial–ground alignment module generates aligned regional representations for subsequent sensor-guided attention fusion through UAV image partitioning, ground sensor temporal encoding, region–sensor soft alignment weighting, environmental context aggregation, and visual–environmental gated fusion.

To further adapt the regional environmental context to the visual feature space, the regional feature $v_i = f_v(P_i; \theta_v)$ output by the visual encoder and the environmental context E_i are fed into a lightweight alignment projection network to obtain region-level multimodal representations with a unified dimension. Specifically, this process can be expressed as $\tilde{E}_i = W_e E_i + b_e$ and $\tilde{v}_i = W_v v_i + b_v$, where W_e and W_v are learnable projection matrices used to eliminate differences in dimensionality, scale, and distribution between visual features and sensor features. Subsequently, gated fusion can be adopted to obtain the aligned regional representation $z_i = \sigma(W_g[\tilde{v}_i, \tilde{E}_i] + b_g) \odot \tilde{v}_i + (1 - \sigma(W_g[\tilde{v}_i, \tilde{E}_i] + b_g)) \odot \tilde{E}_i$. The gate coefficient is adaptively generated according to the visual state and environmental state of the current region, enabling more remote sensing representations to be retained when visual information is reliable, while allowing greater reliance on sensor contexts under illumination variation, occlusion, or unstable vegetation-index responses. This design avoids modal redundancy and noise propagation caused by simple concatenation, and also ensures that each regional sample already has a spatially consistent, temporally matched, and semantically unified multimodal representation before entering the subsequent attention fusion module. From the perspective of the task considered in this study, this module provides three main advantages. First, the scale difference between spatially dense UAV images and sparsely distributed sensor points is addressed, so that crop growth assessment no longer depends on a global environmental average over the whole field but can obtain region-level environmental explanations. Second, the spatial–temporal–quality joint weighting strategy reduces the interference caused by distant sensors, temporally misaligned data, and abnormal sensor readings, thereby improving the reliability of multi-modal sample construction. Finally, the gated projection mechanism aligns sensor contexts with visual features in a unified representation space, providing stable inputs for subsequent sensor-guided attention. As a result, growth abnormalities caused by water deficit, salinity stress, high temperature, or local management differences can be more accurately distinguished, thereby improving the overall credibility of growth classification, yield trend prediction, and agricultural recommendation generation.

3.3.3. Sensor-Guided Remote Sensing Attention Fusion Module

The sensor-guided remote sensing attention fusion module is located after the region-level aerial–ground alignment module. Its inputs are the remote sensing visual feature $F_i^v \in \mathbb{R}^{H\times W\times C_v}$ and the regional environmental context feature $E_i \in \mathbb{R}^{T\times C_s}$ of the i-th region. Here, H and W denote the spatial height and width of the UAV regional feature map, C_v denotes the number of visual channels, T denotes the length of the sensor temporal window, and C_s denotes the dimensionality of sensor variables. Structurally, this module consists of a sensor risk encoding branch, a visual feature projection branch, a cross-modal attention guidance branch, and a lightweight fusion output layer. The visual branch first maps the input features to $\hat{F}_i^v \in \mathbb{R}^{H\times W\times C}$ through L_v depthwise separable convolutional layers or lightweight Transformer blocks, each of which contains convolutional kernel mapping, normalization, and nonlinear activation. The sensor branch encodes E_i into an environmental risk embedding $z_i^s \in \mathbb{R}^C$ through L_s one-dimensional convolutional layers, gated temporal units, and fully connected projection, as shown in Figure 3. Different from conventional self-attention, in which Q, K, and V all originate from the same visual feature space and mainly model spatial dependencies within image regions, the query information in the proposed module is generated from sensor risk embeddings, while the keys and values are derived from UAV remote sensing features. Therefore, the attention weights are not determined by the self-similarity of visual regions, but are conditionally guided by the importance of environmental states to remote sensing regions. The difference can be expressed as:

$$\mathrm{SelfAttn}(F_i^v) = \mathrm{Softmax}\left(\frac{Q(F_i^v)K(F_i^v)^T}{\sqrt{d}}\right)V(F_i^v), \tag{4}$$

$$\mathrm{SGA}(F_i^v, z_i^s) = \mathrm{Softmax}\left(\frac{Q(z_i^s)K(F_i^v)^T}{\sqrt{d}}\right)V(F_i^v). \tag{5}$$

In the specific implementation, the visual feature $\hat{F}_i^v$ is first flattened into $X_i^v \in \mathbb{R}^{HW\times C}$, representing the HW visual tokens at spatial positions within the region. The sensor risk embedding z_i^s is mapped into the query vector $Q_i^s \in \mathbb{R}^{r\times d}$ through a linear layer, where r denotes the number of risk queries generated by the sensor state and d denotes the single-head attention dimension. Visual tokens are mapped into $K_i^v \in \mathbb{R}^{HW\times d}$ and $V_i^v \in \mathbb{R}^{HW\times d}$ through different linear layers. If A attention heads are adopted, each head independently calculates sensor-conditioned spatial attention, and the outputs are concatenated along the channel dimension before being restored to the C-dimensional space through an output projection matrix. This process can be written as:

$$Q_{i,a}^s = z_i^s W_a^Q, \quad K_{i,a}^v = X_i^v W_a^K, \quad V_{i,a}^v = X_i^v W_a^V, \tag{6}$$

$$A_{i,a} = \mathrm{Softmax}\left(\frac{Q^s_{i,a}(K^v_{i,a})^T}{\sqrt{d}}\right), \tag{7}$$

$$O_{i,a} = A_{i,a}V^v_{i,a}, \tag{8}$$

$$O_i = \mathrm{Concat}(O_{i,1}, O_{i,2}, \cdots, O_{i,A})W^O. \tag{9}$$

Meanwhile, to avoid sensor guidance acting only on spatial positions while ignoring the importance of different visual channels, a channel modulation branch is further introduced. The scaling factor γ_i and bias factor β_i are generated from z^s_i to conditionally recalibrate visual features:

$$\gamma_i = \sigma(W_\gamma z^s_i + b_\gamma), \quad \beta_i = W_\beta z^s_i + b_\beta, \tag{10}$$

$$\tilde{X}^v_i = \gamma_i \odot X^v_i + \beta_i. \tag{11}$$

Finally, the spatial attention output O_i, the channel-modulated feature $\tilde{X}^v_i$, and the original visual residual feature are jointly fed into the lightweight fusion layer to obtain the sensor-guided regional fusion representation H_i:

$$H_i = \mathrm{LN}\big(X^v_i + \phi\big([O_i, \tilde{X}^v_i, z^s_i]\big)\big), \tag{12}$$

where $\phi(\cdot)$ denotes a lightweight fusion network composed of linear projection, nonlinear activation, and Dropout, and $\mathrm{LN}(\cdot)$ denotes layer normalization. From a mathematical perspective, the advantage of this module over ordinary self-attention lies in the fact that its attention distribution is controlled by the conditional variable z^s_i. Therefore, conditional visual responses under environmental states are learned instead of pure visual similarity. If the growth prediction target is denoted as y_i, ordinary self-attention approximately models $p(y_i|F^v_i)$, whereas the proposed module models $p(y_i|F^v_i, z^s_i)$. According to the property that conditional information does not reduce discriminative information, when sensor features are correlated with growth states, the following relationship holds:

$$I(y_i; F^v_i, z^s_i) = I(y_i; F^v_i) + I(y_i; z^s_i|F^v_i) \geq I(y_i; F^v_i), \tag{13}$$

where $I(\cdot;\cdot)$ denotes mutual information. Therefore, when sensor variables such as soil moisture, EC, temperature, and light intensity can explain part of the visual phenotypic variation, the introduction of z^s_i can increase the effective discriminative information for growth status. On the other hand, the attention weight $A_{i,a}$ is determined by the similarity between sensor risk queries and visual keys, enabling environmental states such as water deficit, high temperature, or salinity stress to actively select relevant canopy regions and visual channels. As a result, interference from illumination variation, background noise, and irrelevant textures can be reduced during fusion. Thus, the module not only improves the stability of growth-grade classification and yield trend prediction, but also gives attention heatmaps explicit environmental interpretability. In other words, the regions attended by the model can be traced back to specific sensor facts rather than being determined solely by internal texture relationships within the image.

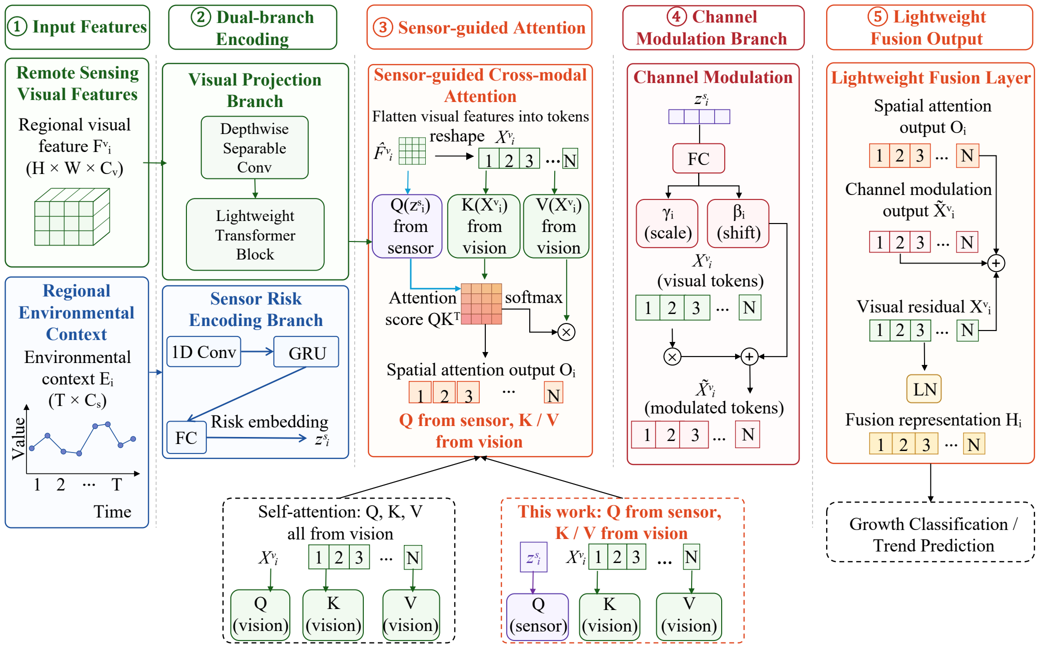

Figure 3. The sensor-guided remote sensing attention fusion module incorporates environmental risk information into remote sensing visual feature modeling through dual-branch encoding, sensor-risk-query-driven cross-modal attention, channel modulation, and lightweight fusion output, thereby enhancing environmental awareness for growth classification and trend prediction.

3.3.4. Sensor-Fact-Constrained Lightweight Agricultural Decision Service Module

The sensor-fact-constrained lightweight agricultural decision service module is located at the output end of the growth assessment network. Its core objective is to further transform region-level multimodal representations into management recommendations that conform to field facts, agronomic logic, and the application requirements of smallholders.

As shown in Figure 4, let the regional fusion representation output by the preceding module be $H_i \in \mathbb{R}^{H_d \times W_d \times C_d}$, where H_d, W_d, and C_d denote the height, width, and number of channels of the decision input feature map, respectively. This representation is first processed through global average pooling and a linear compression layer to obtain the regional state vector $u_i \in \mathbb{R}^{C_u}$. Meanwhile, the growth-grade probability $p_i^g \in \mathbb{R}^{C_g}$, yield trend vector $p_i^r \in \mathbb{R}^{C_r}$, sensor risk factor vector $r_i \in \mathbb{R}^{C_f}$, and historical management record encoding $m_i \in \mathbb{R}^{C_m}$ predicted by the model are jointly fed into the decision module. To ensure the lightweight nature of the module, large-scale generative models are not adopted. Instead, a decision network composed of L_d lightweight perceptron layers, a rule-constraint mapping layer, and a fact-consistency correction layer is designed. The feature compression layer maps information from different sources into a unified dimension C_o, the width of the intermediate layer is set to the symbolic dimension C_h, and the output layer corresponds to C_a types of agricultural recommendations, such as irrigation, fertilization, field inspection, salinity checking, pest and disease rechecking, and risk warning. The overall decision input can be expressed as:

$$x_i^d = \mathrm{Proj}\Big(\mathrm{GAP}(H_i), p_i^g, p_i^r, r_i, m_i\Big), \tag{14}$$

where $\mathrm{GAP}(\cdot)$ denotes global average pooling, and $\mathrm{Proj}(\cdot)$ denotes a lightweight projection network composed of a linear layer, normalization layer, and nonlinear activation function.

Subsequently, the decision network generates a candidate recommendation distribution through multi-layer nonlinear transformations:

$$h_i^{(l)} = \delta\left(W_d^{(l)} h_i^{(l-1)} + b_d^{(l)}\right), \quad l = 1, \cdots, L_d, \tag{15}$$

$$\pi_i = \text{Softmax}\left(W_a h_i^{(L_d)} + b_a\right), \tag{16}$$

where $h_i^{(0)} = x_i^d$, $\delta(\cdot)$ denotes the nonlinear activation function, and $\pi_i \in \mathbb{R}^{C_a}$ denotes the candidate agricultural recommendation probability before fact constraints are applied. To prevent candidate recommendations from deviating from real sensor facts, a fact-constraint matrix $C_i \in [0,1]^{C_a}$ is further constructed to characterize the support degree of the current regional environmental state for different agricultural operations. This matrix is not a fixed empirical label but is jointly generated from sensor risk factors, agronomic threshold intervals, and historical management states. Let the fact-support function corresponding to the k-th type of agricultural recommendation be denoted as $\Omega_k(\cdot)$. Its constraint weight can be expressed as:

$$C_{i,k} = \sigma(\Omega_k(r_i, m_i, \eta_i)), \tag{17}$$

where η_i denotes the encoding of crop growth stage, regional management unit, and agronomic knowledge rules, and $\sigma(\cdot)$ is used to normalize the fact-support strength into a comparable range. For example, when soil moisture risk is high, growth confidence decreases, and no recent irrigation record exists, the fact-support weight for irrigation recommendations will increase. When the EC risk is significant while the soil moisture state does not support conventional irrigation, the weight of a simple irrigation recommendation will be suppressed, whereas the weights for salinity checking or salt-leaching management recommendations will be increased. The final recommendation distribution is jointly determined by the candidate recommendation probability and the fact-constraint weight:

$$\hat{\pi}_{i,k} = \frac{\pi_{i,k} C_{i,k}}{\sum_{q=1}^{C_a} \pi_{i,q} C_{i,q} + \epsilon}, \tag{18}$$

where ϵ is a stability term used to avoid an excessively small denominator. This design ensures that the decision result depends not only on the implicit representation learned by the deep model but also on fact-consistency verification based on current sensor states and agronomic rules. From the perspective of the optimization objective, this module simultaneously constrains recommendation accuracy, fact consistency, and lightweight deployment requirements. Let the expert recommendation label be a_i, the fact-conflict penalty function be $\mathcal{R}_{con}(\hat{\pi}_i, r_i, m_i)$, and the model parameter-complexity constraint be $\mathcal{R}_{size}(\Theta_d)$. The training objective of the decision module can then be expressed as:

$$\mathcal{L}_d = -\sum_{k=1}^{C_a} \mathbb{I}(a_i = k) \log \hat{\pi}_{i,k} + \lambda_c \mathcal{R}_{con}(\hat{\pi}_i, r_i, m_i) + \lambda_s \mathcal{R}_{size}(\Theta_d). \tag{19}$$

The first term is used to learn the expert recommendation distribution, the second term penalizes recommendations that contradict sensor facts, and the third term controls the parameter scale of the module, making it suitable for edge-device deployment. Its mathematical rationale is that the unconstrained recommendation distribution π_i only reflects the statistical preference learned from historical samples, whereas the constrained distribution $\hat{\pi}_i$ introduces a fact-feasible domain into the candidate recommendation space.

If the feasible recommendation set is defined as $\mathcal{A}_i^+ = \{k|C_{i,k} > 0\}$ and the infeasible recommendation set is defined as $\mathcal{A}_i^- = \{k|C_{i,k} = 0\}$, then:

$$\sum_{k \in \mathcal{A}_i^-} \hat{\pi}_{i,k} = 0, \qquad \sum_{k \in \mathcal{A}_i^+} \hat{\pi}_{i,k} = 1. \tag{20}$$

Therefore, fact constraints can directly exclude recommendations that clearly conflict with the current field state from the probability space, reducing the probability of incorrect management operations. Furthermore, when the expert recommendation a_i belongs to the fact-feasible domain, the constrained distribution enhances the relative discriminability among feasible recommendations, and its cross-entropy risk is no longer affected by the probabilities of infeasible recommendations, thereby improving the stability of the recommendation learning process. When applied to the task considered in this study, this module can transform growth grades, yield trends, and environmental risk factors into executable management recommendations. As a result, the system not only determines whether crops are growing well, but also explains why poor growth may occur and which measures should be prioritized. Meanwhile, the combination of lightweight multi-layer perceptrons and rule-constraint mapping avoids dependence on large cloud-based models, making the framework more suitable for smallholder scenarios with weak network conditions and low-computing-power agricultural terminals.

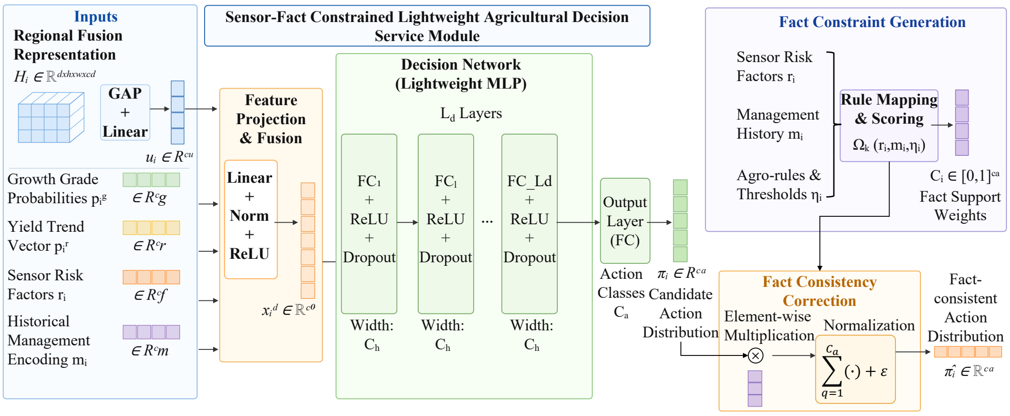

Figure 4. The sensor-fact-constrained lightweight agricultural decision service module integrates regional representations, growth predictions, environmental risks, management records, and agronomic rules to generate candidate recommendations, which are further corrected by fact-constraint weights to produce interpretable management decisions consistent with field conditions.

4. Results and Discussion

4.1. Experimental Configuration

4.1.1. Implementation Details and Experimental Platform

The experiments were conducted on Ubuntu 22.04.3 LTS using Python 3.10.13, PyTorch 2.1.2, Torchvision 0.16.2, CUDA 12.1, and cuDNN 8.9.7. The training server was equipped with an Intel Xeon Gold 6330 CPU, 256 GB RAM, and one NVIDIA RTX 4090 GPU with 24 GB memory. UAV photogrammetric processing was performed using DJI Terra 3.9.4, while OpenCV 4.8.1, GDAL 3.6.4, Rasterio 1.3.9, NumPy 1.26.3, SciPy 1.11.4, Pandas 2.1.4, and Scikit-learn 1.3.2 were used for image and sensor-data processing. Each image sample was resized to 224 × 224 pixels and contained 10 channels: RGB, green, red, red-edge, near-infrared, NDVI, GNDVI, and NDRE. Sensor measurements were recorded every 30 min, and the preceding 72 h were used for each UAV observation, producing a 144 × 8 sequence containing soil moisture, soil temperature, EC, air temperature, relative humidity, photosynthetically active radiation, rainfall, and wind speed. Missing intervals not exceeding 3 h were filled by linear interpolation, whereas samples with more than 20% missing observations were excluded. The visual encoder was a modified MobileNetV3-Small whose first convolutional layer accepted 10 channels. Its output was projected from 576 to 256 dimensions. The sensor encoder consisted of a one-dimensional convolution with 64 channels and kernel size 5, followed by a single-layer bidirectional GRU with 64 hidden units in each direction and a 256-dimensional projection. In the aerial–ground alignment module, the spatial and temporal decay parameters were set to $\tau_d = 30$ m and $\tau_t = 2$ h, respectively, and sensors within an 80 m radius were considered. The sensor-guided attention module used four attention heads, a head dimension of 64, a total embedding dimension of 256, and a Dropout rate of 0.3. The classification, regression, and decision heads produced four growth grades, one continuous yield-trend value, and six agricultural recommendation probabilities, respectively. The classification loss was weighted cross-entropy, the regression loss was Huber loss with $\delta = 1.0$, and the recommendation loss was binary cross-entropy. Their weights were set to $\lambda_{\text{cls}} = 1.0$, $\lambda_{\text{reg}} = 0.5$, and $\lambda_{\text{dec}} = 0.3$.

To control for spatiotemporal data leakage, all RGB and multispectral tiles, sensor windows, agronomic labels, and management records associated with a specific plot were treated as a single unit. The 16 plots were divided into 11 for training, 2 for validation, and 3 for independent testing. To further control for data leakage along the temporal dimension, all samples from the same UAV flight date and plot were assigned to the same subset. No 72 h sensor windows were permitted to span the boundaries between the training, validation, and test sets. Normalization statistics, missing value thresholds, class weights, and hyperparameters were determined solely based on data from the training and validation plots, without utilizing information from the independent test plots. In addition to the fixed plot-level testing, a “leave-one-date-out” evaluation was conducted across the 15 UAV data acquisition dates. In each iteration, all samples from a specific date were held out for testing, while samples from the remaining dates were used for training and validation. The average performance metrics from these 15 evaluations were reported as a supplementary measure of temporal generalization capability.

The model was trained using AdamW with an initial learning rate of 1×10^{-4}, weight decay of 1×10^{-5}, batch size 32, and a maximum of 100 epochs. A five-epoch warm-up was followed by cosine annealing to a minimum learning rate of 1×10^{-6}, and early stopping was applied when the validation loss did not improve for 15 consecutive epochs. Training-time augmentation included horizontal and vertical flipping ($p = 0.5$), rotations in increments of 90°, random resized cropping with a scale of 0.85–1.00, RGB brightness, contrast, and saturation perturbations within ±15%, multispectral reflectance noise within ±3%, and MixUp with $\alpha = 0.2$ for 30% of training batches. The 16 plots were divided at the plot level into 11 training plots, two validation plots, and three independent test plots to prevent spatial leakage. Edge inference was evaluated on a Jetson Nano using JetPack 4.6.4 and TensorRT 8.2.1, and on a Raspberry Pi 4B using ONNX Runtime 1.16.3.

4.1.2. Baseline Models and Evaluation Metrics

The comparative methods used in the experiments mainly included four categories: traditional remote sensing index methods, unimodal deep learning methods, multimodal fusion methods, and lightweight agricultural decision-making methods. Among them, NDVI, GNDVI, and vegetation-index thresholding methods [48] mainly evaluate vegetation vigor by exploiting reflectance differences among different spectral bands, and their advantages lie in simple computation and easy agricultural interpretation. Random forest regression [49] completes growth classification and yield prediction by constructing multiple decision trees and has favorable nonlinear modeling capability and stability. Among unimodal deep learning methods, ResNet [50] uses residual structures to alleviate deep network degradation and has strong feature extraction capability. EfficientNet [51] balances network depth, width, and resolution through a compound scaling strategy, achieving a favorable trade-off between accuracy and computational cost. Vision Transformer (ViT) [52] models long-range spatial dependencies through a global self-attention mechanism and is suitable for complex farmland scene analysis. Swin Transformer [53] reduces the computational complexity of Transformers through a shifted-window mechanism while preserving multi-scale spatial modeling capability. MobileNet [54] constructs a lightweight network based on depthwise separable convolution and has a low parameter scale and fast inference speed. Among multimodal fusion methods, Feature Concatenation [55] directly concatenates image features and sensor features to form a joint representation, with a simple and easy-to-implement structure. Late Fusion [56] performs decision-level fusion after independently modeling different modalities, thereby reducing cross-modal interference. Attention Fusion [57] dynamically adjusts the importance of different modalities through an attention mechanism and has favorable cross-modal association modeling capability. Among the lightweight and decision-system comparison methods, the MobileNet-based Classifier [54] was mainly used to verify the deployability of lightweight visual models on edge devices, with the advantage of low computational-resource consumption. The Rule-based Decision System [58] generated irrigation and fertilization recommendations based on agricultural empirical rules and had strong interpretability. The recommendation generation module without sensor-fact constraints [59] mainly relied on model prediction results to directly generate agricultural recommendations. Although its generation process was simple, real environmental states and agricultural constraints could be easily ignored.

To ensure a fair comparison, all trainable baselines were evaluated using the same plot-level training, validation, and test partitions, the same image and sensor preprocessing procedures, the same training-time augmentation policy, and the same evaluation metrics. Hyperparameters for each model were selected using the validation set under the same tuning budget. Visual-only baselines received only the RGB, multispectral, and vegetation-index inputs available to their respective architectures, whereas multimodal baselines received the same visual inputs together with the same temporally aligned sensor variables used by the proposed framework. Therefore, visual-only and multimodal results are reported as separate comparison groups and are interpreted according to their different input-information settings.

To comprehensively evaluate the performance of the proposed UAV–ground sensor collaborative crop growth assessment framework, evaluations were conducted from four aspects: classification performance, regression performance, edge deployment efficiency, and agricultural recommendation effectiveness. Accuracy was used to measure the overall classification correctness. Precision was used to evaluate the correctness of positive predictions. Recall was used to measure the model's ability to identify true positive samples. F1-score comprehensively reflected the balance between Precision and Recall. Cohen's Kappa was used to evaluate the consistency between classification results and ground-truth labels. MAE, $RMSE$, $MAPE$, and R^2 were used to evaluate the regression error of yield or biomass prediction. Params, FLOPs, Latency, FPS, Memory Usage, and Energy Consumption were used to measure edge deployment efficiency. Expert agreement rate, risk identification accuracy, and user usability score were used to evaluate the reliability and practicality of the agricultural recommendation module. The mathematical definitions of the relevant evaluation metrics are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{21}$$

$$Precision = \frac{TP}{TP + FP}, \tag{22}$$

$$Recall = \frac{TP}{TP + FN}, \tag{23}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{24}$$

$$Kappa = \frac{p_o - p_e}{1 - p_e}, \tag{25}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{26}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{27}$$

$$MAPE = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|, \tag{28}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{29}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively; p_o denotes the observed agreement rate, and p_e denotes the expected agreement rate by chance; y_i denotes the ground-truth value, $\hat{y}_i$ denotes the predicted value, $\bar{y}$ denotes the mean of the ground-truth values, and N denotes the total number of samples. For edge deployment metrics, Params denotes the number of model parameters, FLOPs denotes the number of floating-point operations, Latency denotes the inference latency for a single sample, FPS denotes the number of frames processed per second, Memory Usage denotes runtime memory consumption, and Energy Consumption denotes the energy consumed during model execution.

4.2. Performance Comparison of Different Methods

This experiment was designed to systematically compare the overall performance of traditional remote sensing indices, machine learning methods, unimodal deep learning methods, multimodal fusion methods, and the proposed method in terms of growth-grade classification and yield-trend prediction.

As shown in Table 3 and Figure 5, the Accuracy values of NDVI Thresholding and GNDVI Thresholding are 76.42% and 77.35%, respectively, while their RMSE values are 0.682 and 0.651, respectively. This indicates that although a single vegetation index has favorable agronomic interpretability, its representation mainly depends on spectral ratios and is insufficient for characterizing the complex relationships among canopy texture, spatial structure, and environmental stress. The Accuracy of Random Forest increases to 81.26%, indicating that nonlinear tree-based models can exploit multidimensional feature combinations to a certain extent. However, their feature-partition-based modeling mechanism remains limited in representing high-dimensional spatial patterns in images. The results of ResNet, EfficientNet, ViT, Swin Transformer, and MobileNet are generally superior to those of traditional methods. Among them, Swin Transformer achieves an Accuracy of 87.24% and an RMSE of 0.493, demonstrating that deep models can learn more complex spatial features from UAV images. Nevertheless, unimodal visual models still lack environmental state information and therefore have limitations in explaining growth differences caused by water deficit, salinity, or high temperature. The overall performance of multimodal methods is further improved. Feature Concatenation, Late Fusion, and Attention Fusion achieve Accuracy values of 88.16%, 88.74%, and 89.63%, respectively, indicating that the introduction of ground sensor information can effectively complement the environmental context missing from image features. Attention Fusion outperforms simple concatenation and late fusion because the attention mechanism can dynamically adjust modality weights according to feature relevance, rather than combining information from different sources in a fixed manner. The proposed method achieves the best results, with Accuracy, F1-score, and R^2 reaching 92.47%, 91.62%, and 0.902, respectively, while RMSE decreases to 0.381. This advantage is attributed to the synergistic effects of region-level aerial–ground alignment, sensor-guided attention, and fact-constrained decision-making. Region-level alignment reduces scale mismatch between point-based sensors and image regions, sensor-guided attention enables environmental risks to serve as conditional information for visual region selection, and gated fusion adaptively allocates information contributions according to the reliability of visual and sensor features. Therefore, the proposed method improves not only classification performance but also the stability of yield-trend prediction.

Table 3. Performance comparison (mean ± std) and statistical significance (*p*-value) against the proposed method.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	RMSE	R^2	*p*-Value
NDVI Thresholding	76.42 ± 0.85	75.18 ± 0.72	74.96 ± 0.88	75.07 ± 0.79	0.682 ± 0.04	0.713 ± 0.03	0.0012
GNDVI Thresholding	77.35 ± 0.92	76.41 ± 0.65	75.83 ± 0.76	76.12 ± 0.68	0.651 ± 0.03	0.731 ± 0.02	0.0018
Random Forest	81.26 ± 0.78	80.34 ± 0.71	80.12 ± 0.69	80.23 ± 0.72	0.596 ± 0.03	0.768 ± 0.02	0.0035
ResNet	84.73 ± 0.66	83.95 ± 0.58	83.48 ± 0.62	83.71 ± 0.60	0.548 ± 0.02	0.794 ± 0.02	0.0052
EfficientNet	85.91 ± 0.55	85.16 ± 0.52	84.72 ± 0.55	84.94 ± 0.53	0.526 ± 0.02	0.812 ± 0.01	0.0087
ViT	86.38 ± 0.52	85.77 ± 0.48	85.31 ± 0.51	85.54 ± 0.50	0.511 ± 0.02	0.821 ± 0.01	0.0124
Swin Transformer	87.24 ± 0.48	86.62 ± 0.45	86.11 ± 0.47	86.36 ± 0.46	0.493 ± 0.02	0.834 ± 0.01	0.0156
MobileNet	83.69 ± 0.62	82.85 ± 0.60	82.41 ± 0.58	82.63 ± 0.59	0.571 ± 0.02	0.781 ± 0.02	0.0041
Feature Concatenation	88.16 ± 0.42	87.43 ± 0.38	87.02 ± 0.41	87.22 ± 0.40	0.472 ± 0.01	0.846 ± 0.01	0.0210
Late Fusion	88.74 ± 0.39	88.05 ± 0.35	87.61 ± 0.38	87.83 ± 0.37	0.459 ± 0.01	0.853 ± 0.01	0.0285
Attention Fusion	89.63 ± 0.35	88.97 ± 0.32	88.54 ± 0.34	88.75 ± 0.33	0.438 ± 0.01	0.866 ± 0.01	0.0342
Proposed Method	92.47 ± 0.28	91.86 ± 0.25	91.39 ± 0.27	91.62 ± 0.26	0.381 ± 0.01	0.902 ± 0.01	-

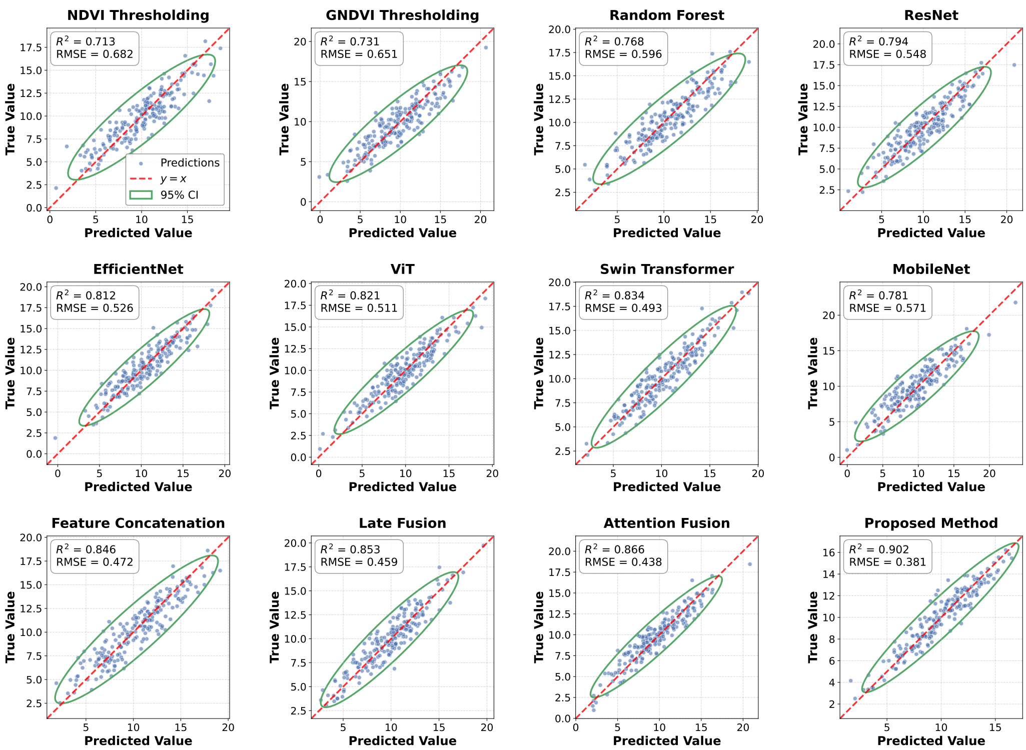

Figure 5. Regression comparison between true and predicted values for yield trend prediction across different methods, showing that the proposed method achieves the highest R^2 and the lowest RMSE, with predictions more closely aligned with the ideal fitting line.

4.3. Deployment Efficiency Comparison of Different Methods

This experiment was designed to verify the practical deployability of different models on low-power edge devices. The number of parameters, computational complexity, inference latency, frame rate, memory usage, and energy consumption were compared to determine whether the models are suitable for low-cost application scenarios involving smallholders and medium-sized farms.

To ensure the reproducibility of the edge deployment evaluation, all inference measurements were conducted on two representative low-power devices: the NVIDIA Jetson Nano and the Raspberry Pi 4B. The Jetson Nano was configured with JetPack 4.6.4, CUDA 10.2, cuDNN 8.2.1, and TensorRT 8.2.1, while the Raspberry Pi 4B ran the 64-bit Raspberry Pi OS and utilized ONNX Runtime 1.16.3. All models were exported to the ONNX format prior to deployment and optimized using TensorRT on the Jetson Nano. Inference was performed with a batch size of 1 to simulate real-time field monitoring scenarios. Experiments on the Jetson Nano utilized FP16 precision, whereas experiments on the Raspberry Pi 4B employed FP32 precision due to hardware limitations. Prior to formal measurement, 50 warm-up inference runs were executed to eliminate initialization overhead. Subsequently, each model underwent 500 inference runs, with the average value recorded as the final latency result. Memory usage was assessed based on the maximum allocated GPU memory reported by TensorRT during inference on the Jetson Nano and the peak RAM consumption monitored via system-level tools on the Raspberry Pi. Energy consumption was measured using a USB power meter for the Jetson Nano and an inline power monitoring module for the Raspberry Pi 4B, recording the average power over the 500 inference runs. All deployment metrics in Table 4 are presented as the "mean $\pm$ standard deviation" derived from multiple measurements.

As shown in Table 4 and Figure 6, the Params and FLOPs of ViT reach 86.43 M and 16.85 G, respectively. Its Latency is as high as 214.76 ms, while its FPS is only 4.65, and both Memory and Energy are the highest among all methods. This indicates that although global self-attention has strong spatial modeling capability, its computational complexity increases rapidly with the number of tokens, making it unsuitable for real-time edge inference. ResNet and Swin Transformer also have relatively high parameter scales and computational costs, with Latency values of 86.37 ms and 103.29 ms, respectively, indicating that residual convolutional networks and window-based Transformers, although more stable or efficient than ViT, still introduce considerable computational overhead. EfficientNet reduces the model scale through compound scaling, decreasing Latency to 63.42 ms and increasing FPS to 15.77, thereby showing a favorable accuracy–efficiency trade-off. MobileNet and the MobileNet-based Classifier exhibit the highest deployment efficiency. In particular, MobileNet contains only 4.27M Params and 0.62G FLOPs, with a Latency of 31.84 ms and an FPS of 31.41. This demonstrates that depthwise separable convolution can significantly reduce the channel-mixing cost in convolutional computation, thereby providing strong advantages for edge deployment. The proposed lightweight framework has 6.18M Params, 0.91G FLOPs, a Latency of 39.56 ms, and an FPS of 25.28. Although its computational cost is slightly higher than that of pure MobileNet, it is substantially lower than those of ResNet, EfficientNet, ViT, and Swin Transformer. This is because the proposed method retains a lightweight visual backbone while introducing only low-dimensional sensor encoding, region-level alignment, and lightweight fusion layers, without adopting large-scale global attention computation. From the perspective of mathematical structure, high-dimensional image computation is restricted within regional feature spaces, sensor information is compressed into low-dimensional environmental vectors, and fusion is completed through gated and lightweight attention operations, thereby avoiding large-scale fully connected interactions among high-dimensional features. Therefore, the framework maintains multimodal perception capability while preserving favorable edge inference efficiency and practical deployment value.

Table 4. Deployment efficiency comparison of different lightweight models on edge devices (mean ± std) and statistical significance (*p*-value).

Method	Params (M)	FLOPs (G)	Latency (ms)	FPS	Memory (MB)	*p*-Value
ResNet	23.51 ± 0.12	4.12 ± 0.05	86.37 ± 1.24	11.58 ± 0.15	842 ± 4.2	0.0012
EfficientNet	11.24 ± 0.08	2.36 ± 0.03	63.42 ± 0.95	15.77 ± 0.12	624 ± 3.5	0.0034
ViT	86.43 ± 0.45	16.85 ± 0.18	214.76 ± 2.45	4.65 ± 0.06	1786 ± 8.6	0.0001
Swin Transformer	28.31 ± 0.15	4.58 ± 0.06	103.29 ± 1.38	9.68 ± 0.11	936 ± 5.1	0.0008
MobileNet	4.27 ± 0.05	0.62 ± 0.02	31.84 ± 0.62	31.41 ± 0.22	286 ± 2.1	0.0215
MobileNet-based	4.86 ± 0.06	0.74 ± 0.02	35.27 ± 0.75	28.35 ± 0.18	314 ± 2.4	0.0342
Proposed Framework	6.18 ± 0.04	0.91 ± 0.01	39.56 ± 0.58	25.28 ± 0.14	352 ± 1.8	–

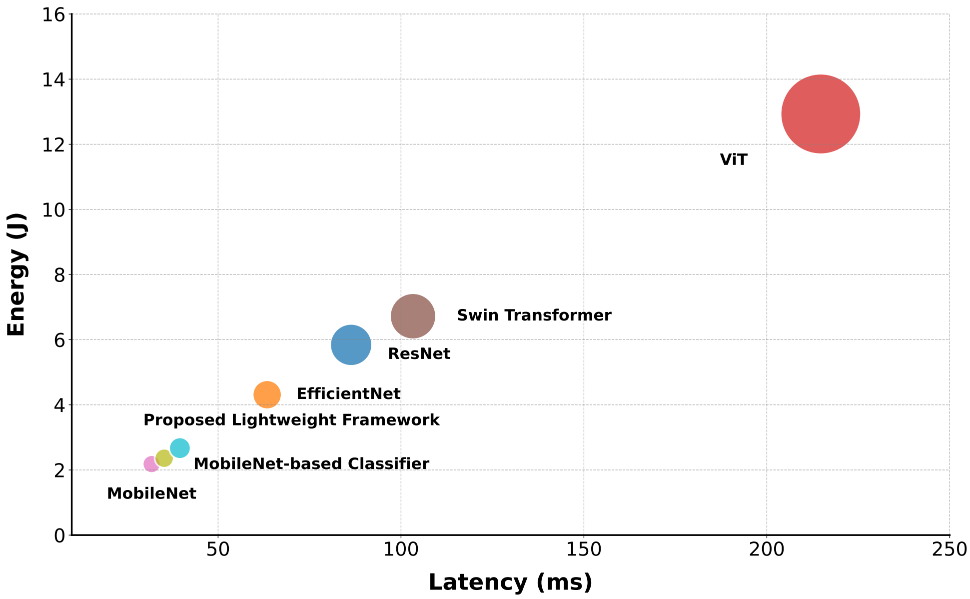

Figure 6. Comparison of latency–energy distribution among different lightweight models on edge devices, showing that the proposed framework achieves favorable deployment efficiency with relatively low inference latency and energy consumption.

4.4. Ablation Study of Different Modules

This experiment was designed to verify the actual contribution of each core module in the proposed framework to growth recognition, yield-trend prediction, and agricultural recommendation generation. It was further used to determine whether the performance improvement was derived from the complete structural design rather than from a single modality or an incidental component.

As shown in Table 5 and Figure 7, the full model achieves the best results across all metrics, with Accuracy, F1-score, R^2, expert agreement, and risk Acc. reaching 92.47%, 91.62%, 0.902, 89.34%, and 90.18%, respectively, while RMSE decreases to 0.381. These results indicate a clear synergistic effect among regional alignment, sensor-guided fusion, and fact-constrained decision-making. After removing region-level alignment, Accuracy decreases to 89.26% and RMSE increases to 0.446, suggesting that without region-level aerial–ground correspondence, point-based sensor information cannot be accurately matched with image regions, leading to mismatches between environmental contexts and visual regions. After removing sensor-guided attention, Accuracy and F1-score decrease to 90.13% and 89.25%, respectively, indicating that ordinary fusion cannot dynamically select key visual regions according to environmental states such as water deficit, high temperature, or salinity stress. The removal of channel modulation and gated fusion also leads to performance degradation, demonstrating that channel recalibration and gating mechanisms can effectively control the contribution ratio of different modal information and reduce redundant features and noise propagation. From a deeper perspective, image-only branch and sensor-only branch perform substantially worse than the complete model. The Accuracy of image-only branch is 86.91%, while that of sensor-only branch is only 82.37%, indicating that visual information is more suitable for capturing canopy spatial differences, whereas sensor information is more suitable for explaining environmental variations. A single modality cannot fully describe the formation process of crop growth status. After sensor-fact constraint is removed, the classification and regression performance decreases only slightly, but expert agreement drops markedly from 89.34% to 80.27%, and Risk Acc. decreases to 82.64%. This indicates that fact constraints have a limited impact on model prediction itself but play a critical role in the reliability of agricultural recommendations. After management records are removed, expert agreement and risk Acc. also decrease significantly, suggesting that historical irrigation, fertilization, and field inspection records provide important management context for recommendation generation. From a mathematical modeling perspective, the complete model reduces spatial errors between modalities through regional alignment, enhances task-relevant information responses through sensor-conditioned attention, adaptively allocates visual and environmental feature weights through gated fusion, and compresses the output space of unreasonable recommendations through fact constraints. Therefore, recognition accuracy, regression stability, and decision credibility can be simultaneously improved.

Table 5. Ablation study of different modules in the proposed method (mean ± std, p-value calculated against the Full Model).

Variant	Accuracy (%)	F1-Score (%)	RMSE	R^2	Expert Agr. (%)	p-Value
Full Model	92.47 ± 0.28	91.62 ± 0.26	0.381 ± 0.01	0.902 ± 0.01	89.34 ± 0.31	–
w/o Reg. Alignment	89.26 ± 0.42	88.41 ± 0.40	0.446 ± 0.02	0.861 ± 0.02	84.72 ± 0.45	0.0014
w/o Sensor Attn.	90.13 ± 0.38	89.25 ± 0.35	0.427 ± 0.02	0.874 ± 0.02	85.96 ± 0.38	0.0028
w/o Ch. Modulation	90.84 ± 0.35	89.93 ± 0.32	0.415 ± 0.02	0.881 ± 0.01	86.71 ± 0.36	0.0051
w/o Gated Fusion	89.78 ± 0.40	88.94 ± 0.38	0.433 ± 0.02	0.869 ± 0.02	85.48 ± 0.41	0.0022
w/o Sensor Const.	91.06 ± 0.32	90.21 ± 0.30	0.406 ± 0.01	0.886 ± 0.01	80.27 ± 0.42	0.0039
w/o Mgmt. Records	91.35 ± 0.30	90.47 ± 0.28	0.397 ± 0.01	0.891 ± 0.01	83.19 ± 0.39	0.0064
Image-only Branch	86.91 ± 0.55	86.05 ± 0.52	0.507 ± 0.03	0.824 ± 0.03	75.62 ± 0.65	0.0002
Sensor-only Branch	82.37 ± 0.62	81.44 ± 0.58	0.581 ± 0.04	0.776 ± 0.04	78.35 ± 0.58	0.0001

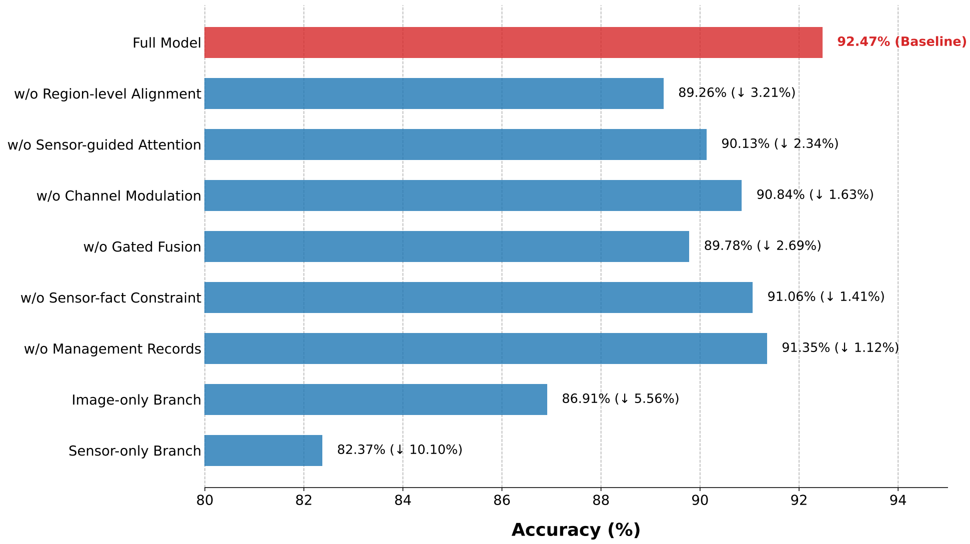

Figure 7. Accuracy comparison in the ablation study of different modules, demonstrating that region-level aerial–ground alignment, sensor-guided attention, gated fusion, and fact constraints all contribute to model performance improvement.

4.5. Comparison of Different Farmland Management Strategies

This experiment was designed to evaluate the differences among different farmland management strategies in terms of labor cost, water use, fertilizer input, yield benefit, and input–output efficiency from the perspective of agricultural economics. All economic values originally recorded in CNY were converted into USD using an exchange rate of 1 USD = 7.2 CNY. This experiment was used to verify whether the proposed method not only improves model prediction performance but can also be transformed into resource savings and economic benefits in practical production.

This economic assessment is based on field-measured resource consumption, production records, and management outputs associated with specific strategies, rather than arbitrary assumptions. The formulas for net benefit (NB) and the input-output ratio (I/O) are $NB = P_y Y - C_l - C_w - C_f - C_e$ and $I/O = P_y Y/(C_l + C_w + C_f + C_e)$, respectively, where P_y is the maize selling price, Y is the yield, and C_l, C_w, C_f, and C_e represent costs related to labor, irrigation, fertilizer, and equipment. Cost parameters were determined based on local agricultural conditions in the Hetao Irrigation District, including local maize market prices, agricultural labor costs, irrigation fees, fertilizer prices, and the operating and depreciation costs of drones and sensors. For the proposed framework, resource inputs and yield variations were estimated based on management measures generated by the decision-support module and corresponding field records. Sensitivity analysis was conducted by adjusting key economic parameters within a $\pm 10\%$ range; the results demonstrated that the framework consistently maintained the highest input-output efficiency.

As shown in Table 6 and Figure 8, Traditional Manual Inspection has the highest labor cost, reaching 175 USD/ha, while water use and fertilizer use reach 4860 m^3/ha and 348 kg/ha, respectively. This indicates that management strategies relying on human experience are prone to low field-inspection efficiency and coarse irrigation and fertilization decisions. Compared with manual management, the Rule-based Decision System shows certain improvements, with labor cost reduced to 145 USD/ha and net benefit increased to 1692 USD/ha. However, because rule-based systems mainly rely on fixed thresholds, they are difficult to adapt to complex field heterogeneity. UAV-only Monitoring and Sensor-only Monitoring rely on spatial remote sensing and environmental temporal monitoring, respectively, and both reduce management costs and improve yield. Sensor-only Monitoring performs slightly better in water and fertilizer savings, whereas UAV-only Monitoring is more direct in spatial growth recognition. However, both methods are limited by a single information source and cannot simultaneously account for spatial differences and environmental causal interpretation.

Simple multimodal fusion simultaneously uses UAV images and sensor data, reducing labor cost to 110 USD/ha, water use to 3980 m^3/ha, and fertilizer use to 296 kg/ha, while increasing yield to 8525 kg/ha. The proposed framework achieves the best economic benefits, with labor cost, water use, and fertilizer use reduced to 88 USD/ha, 3560 m^3/ha, and 268 kg/ha, respectively. Meanwhile, yield reaches 8790 kg/ha, and net benefit and input–output ratio increase to 2281 USD/ha and 2.71, respectively. Theoretically, the advantage of the proposed method lies in the fact that region-level aerial–ground alignment reduces information mismatch, enabling resource inputs to be localized to specific plot regions; sensor-guided attention identifies environmental stress sources, allowing irrigation and fertilization decisions to no longer depend solely on apparent growth status; and the fact-constrained decision module excludes management recommendations that conflict with current field conditions, thereby reducing ineffective inputs.

Table 6. Agricultural economic benefit comparison under different farmland management strategies (mean ± std, *p*-value calculated against the Proposed Framework).

Management Strategy	Labor Cost (USD/ha)	Water Use (m^3/ha)	Yield (kg/ha)	Net Benefit (USD/ha)	I/O Ratio	*p*-Value
Traditional Manual	175 ± 8.2	4860 ± 115	7860 ± 95	1561 ± 45	1.84 ± 0.06	0.0012
Rule-based System	145 ± 6.5	4520 ± 98	8050 ± 88	1692 ± 40	1.96 ± 0.05	0.0028
UAV-only Monitoring	128 ± 5.8	4385 ± 92	8235 ± 82	1816 ± 38	2.08 ± 0.05	0.0051
Sensor-only Monitoring	122 ± 5.2	4210 ± 85	8310 ± 78	1881 ± 35	2.17 ± 0.04	0.0084
Simple Multimodal	110 ± 4.5	3980 ± 75	8525 ± 70	2039 ± 32	2.34 ± 0.04	0.0156
Proposed Framework	88 ± 3.8	3560 ± 62	8790 ± 65	2281 ± 28	2.71 ± 0.03	–

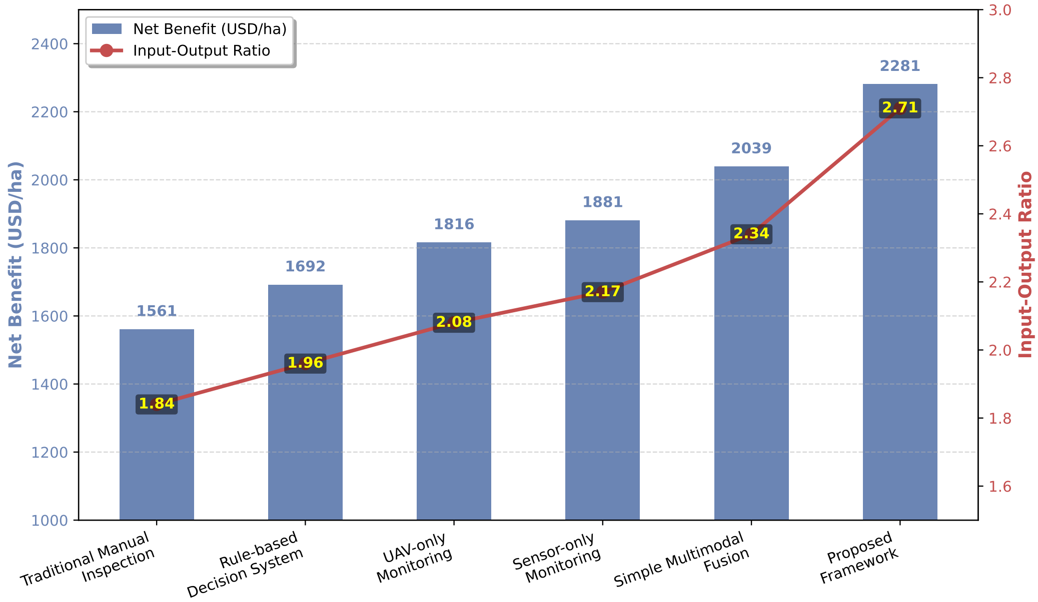

Figure 8. Agricultural economic benefit comparison of net benefit and input–output ratio under different farmland management strategies.

4.6. Discussion

4.6.1. Agricultural Implications

From the perspective of agricultural application, the value of the proposed method is mainly reflected in region-level identification of crop growth differences and causal interpretation under complex field environments. Taking typical irrigated farmland in the Hetao Irrigation District as an example, higher soil moisture near irrigation inlets, stronger evaporation at field edges, salt accumulation in low-lying areas, and uneven local management often coexist within the same plot. If only manual field inspection is relied upon, farmers can usually observe that a certain area appears “lighter in color” or “uneven in growth,” but it is difficult to determine the dominant underlying cause in a timely manner. The proposed framework captures canopy color, coverage, and spatial distribution differences through UAV images, while soil moisture, temperature, EC, light intensity, and meteorological data are used to supplement environmental state information. In this way, the system can associate visually observed growth degradation with specific field environmental changes. For example, when a UAV image shows reduced canopy coverage in a certain region and the corresponding sensor temporal window indicates persistently low soil moisture, the system is more likely to identify water stress in that region. When weak visual growth is accompanied by persistently high EC levels, the system can indicate the risk of salinity stress or insufficient salt-leaching management. This aerial–ground collaborative perception strategy can provide more specific regional evidence for irrigation scheduling, saline–alkali land management, local fertilization supplementation, and field rechecking, rather than merely providing an average growth assessment for the entire field.

In practical agricultural production, the proposed method can also improve the timeliness and targeting of field operations. For field crops such as wheat, maize, and sunflower, water and nutrient supply during key growth stages directly affects yield formation. If farmers can only take action after obvious wilting or large-scale color changes occur, the optimal intervention window has often already been missed. The proposed framework establishes a connection between periodic UAV field inspection and continuous sensor monitoring, enabling growth abnormalities to be detected at an early stage. Environmental data can then be used to determine whether irrigation, fertilization supplementation, salt-leaching management, or enhanced pest and disease inspection is required. For medium-sized and small farms with limited infrastructure or insufficient agricultural technical services, the system can serve as an auxiliary tool for grassroots agricultural technicians, helping them quickly locate problematic regions, reduce blind field-inspection areas, and improve field management efficiency. Meanwhile, the lightweight design enables the model to be deployed on edge devices, allowing basic inference and recommendation generation to be completed in farmland environments with unstable network conditions, thereby enhancing the usability of the system in real production scenarios.

4.6.2. Agricultural Economic Implications

From the perspective of agricultural economics, the significance of the proposed method lies not only in improving growth recognition accuracy but also in reducing the information acquisition cost of farmers and optimizing the allocation of production factors. Traditional farmland management heavily depends on human experience, and farmers must repeatedly inspect fields to determine whether irrigation, fertilization, or control measures should be adopted. This process not only consumes labor but may also leads to excessive or delayed inputs due to insufficient information. For example, continued irrigation under adequate soil moisture conditions increases water consumption and drainage pressure. Blind fertilization before the cause of growth decline is clarified may increase fertilizer costs and introduce the risk of non-point source pollution. By integrating UAV remote sensing, ground sensors, and management records, the proposed framework enables farmers to obtain more fine-grained field status information at a lower cost, thereby reducing ineffective inputs caused by reliance on experience-based judgment. In the comparison of different farmland management strategies, the proposed method shows lower labor costs, lower water and fertilizer inputs, and higher net benefits than traditional manual inspection, indicating that intelligent perception and decision services can improve the input–output relationship in agricultural production. The economic logic is that region-level growth assessment improves information precision, sensor-fact constraints improve decision reliability, and lightweight deployment reduces technology usage costs. These three factors jointly enhance the adoptability of precision agriculture technologies in smallholder scenarios. For farmers, the system does not simply replace human experience; instead, dispersed field information is transformed into executable production recommendations, allowing limited water, fertilizer, labor, and agricultural technical service resources to be preferentially allocated to regions with higher risks and greater potential for benefit improvement. For cooperatives and large-scale farming entities, the framework can also support plot-level zoning management, input budget planning, and production risk warning, thereby improving overall operational efficiency. Therefore, the proposed method establishes a connection between technical performance and economic feasibility, supporting the transition of smart agriculture from being merely recognizable to being applicable, affordable, and scalable.

4.7. Limitations and Future Work

Although the proposed UAV–ground sensor collaborative framework achieved promising results in crop growth assessment, yield-trend prediction, edge deployment, and decision-support evaluation under the experimental conditions investigated in this study, several limitations should be acknowledged. First, the dataset was collected from spring maize (*Zea mays* L., cultivar Xianyu 335) grown in irrigated farmland in Wuyuan County, Bayannur City, Inner Mongolia, China. Although the experimental area represents an important agricultural production region, differences in climate conditions, soil properties, crop varieties, cultivation practices, and sensor deployment strategies may influence model performance in other regions. Therefore, the current results should be interpreted within the specific experimental conditions, and broader validation through cross-region, cross-year, and cross-crop experiments is still required. Second, although the proposed region-level aerial–ground alignment mechanism alleviates the impact of sparse sensor deployment, the number and spatial distribution of ground sensors may still limit environmental context estimation in large-scale farmland or highly heterogeneous fields. Future studies should investigate adaptive sensor deployment strategies and incorporate additional environmental data sources, such as soil sampling, pest and disease observations, weather forecasts, and agricultural machinery operation records, to improve the robustness and explanatory capability of the framework. Third, the agricultural economic analysis in this study mainly focuses on operational benefits, including changes in labor demand and agricultural resource utilization under the investigated management scenario. A complete evaluation of the total cost of ownership (TCO), including capital expenditure associated with UAV platforms, sensors, edge devices, communication infrastructure, maintenance, and equipment depreciation, was not fully considered. In addition, long-term economic factors, such as market fluctuations, policy changes, and farmer adoption behavior, may affect the practical value of intelligent agricultural systems. Therefore, future research should combine technical performance evaluation with comprehensive cost–benefit analysis and technology adoption models.

Future work will focus on validating the framework across multiple regions, seasons, crop types, and management conditions to evaluate its generalization capability. Furthermore, integrating richer agricultural data sources and long-term deployment observations will be explored to improve environmental interpretation and decision reliability. These efforts will help establish a more comprehensive evaluation framework for intelligent agricultural systems and support their practical application under diverse farming scenarios.

5. Conclusions

With the development of precision agriculture, the agricultural Internet of Things, and UAV remote sensing technologies, this study proposes a lightweight UAV–ground sensor collaborative framework for crop growth assessment and agricultural decision support. The framework is designed to address challenges associated with multimodal data integration, environment-aware crop growth assessment, and evidence-based agricultural decision generation. UAV imagery and ground sensor observations are integrated through region-level aerial–ground alignment, sensor-guided attention fusion, and a fact-constrained decision module to support crop growth assessment and management recommendation. Experiments were conducted on spring maize (*Zea mays* L., cultivar Xianyu 335) under the specific field and management conditions of the Hetao Irrigation District in Inner Mongolia. Under these experimental conditions, the proposed method achieved Accuracy, Precision, Recall, and F1-score values of 92.47%, 91.86%, 91.39%, and 91.62%, respectively, together with an RMSE of 0.381 and an R^2 of 0.902 for yield-trend prediction. The lightweight implementation achieved an inference latency of 39.56 ms and a processing speed of 25.28 FPS on the evaluated edge device. Ablation experiments further demonstrated the contribution of the proposed modules within the current dataset and experimental setting. These findings suggest that the proposed framework may provide useful technical support for crop growth assessment and agricultural management under conditions similar to those evaluated in this study.

Author Contributions: Conceptualization, P.G., K.W., Y.L. and Y.S.; Methodology, P.G., K.W. and Y.L.; Software, P.G., K.W. and Y.L.; Validation, Q.L. and W.L.; Formal analysis, Q.L. and W.L.; Investigation, Q.L. and W.L.; Resources, J.Z.; Data curation, J.Z.; Writing—original draft, P.G., K.W., Y.L., J.Z., Q.L., W.L. and Y.S.; Visualization, J.Z.; Supervision, Y.S.; Project administration, Y.S.; Funding acquisition, Y.S., P.G., K.W. and Y.L. contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

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

References

1. Zhao, J.; Liu, D.; Huang, R. A review of climate-smart agriculture: Recent advancements, challenges, and future directions. *Sustainability* **2023**, *15*, 3404. [CrossRef]
2. Sharma, K.; Shivandu, S.K. Integrating artificial intelligence and Internet of Things (IoT) for enhanced crop monitoring and management in precision agriculture. *Sens. Int.* **2024**, *5*, 100292. [CrossRef]
3. Schmitz, A.; Badgujar, C.; Mansur, H.; Flippo, D.; McCornack, B.; Sharda, A. Design of a reconfigurable crop scouting vehicle for row crop navigation: A proof-of-concept study. *Sensors* **2022**, *22*, 6203. [CrossRef] [PubMed]
4. Edan, Y.; Adamides, G.; Oberti, R. Agriculture automation. In *Springer Handbook of Automation*; Springer: Cham, Switzerland, 2023; pp. 1055–1078.
5. Omia, E.; Bae, H.; Park, E.; Kim, M.S.; Baek, I.; Kabenge, I.; Cho, B.K. Remote sensing in field crop monitoring: A comprehensive review of sensor systems, data analyses and recent advances. *Remote Sens.* **2023**, *15*, 354. [CrossRef]
6. Nielsen, K.M.; Duddu, H.S.; Bett, K.E.; Shirtliffe, S.J. UAV image-based crop growth analysis of 3D-reconstructed crop canopies. *Plants* **2022**, *11*, 2691. [CrossRef] [PubMed]
7. Li, X.H.; Li, M.Z.; Li, J.Y.; Gao, Y.Y.; Liu, C.R.; Hao, G.F. Wearable sensor supports in-situ and continuous monitoring of plant health in precision agriculture era. *Plant Biotechnol. J.* **2024**, *22*, 1516–1535. [CrossRef] [PubMed]
8. Sun, G.; Zheng, X.; Li, J.; Kang, H.; Liang, S. Collaborative WSN-UAV data collection in smart agriculture: A bi-objective optimization scheme. In *ACM Transactions on Sensor Networks*; Association for Computing Machinery: New York, NY, USA, 2023.
9. Wu, B.; Zhang, M.; Zeng, H.; Tian, F.; Potgieter, A.B.; Qin, X.; Yan, N.; Chang, S.; Zhao, Y.; Dong, Q.; et al. Challenges and opportunities in remote sensing-based crop monitoring: A review. *Natl. Sci. Rev.* **2023**, *10*, nwac290. [PubMed]
10. Vidican, R.; Mălinaș, A.; Ranta, O.; Moldovan, C.; Marian, O.; Ghețe, A.; Ghișe, C.R.; Popovici, F.; Cătunescu, G.M. Using remote sensing vegetation indices for the discrimination and monitoring of agricultural crops: A critical review. *Agronomy* **2023**, *13*, 3040. [CrossRef]
11. Wang, Z.; Chen, T.; Zhu, D.; Jia, K.; Plaza, A. RSEIFE: A new remote sensing ecological index for simulating the land surface eco-environment. *J. Environ. Manag.* **2023**, *326*, 116851. [CrossRef] [PubMed]
12. Araujo, S.O.; Peres, R.S.; Ramalho, J.C.; Lidon, F.; Barata, J. Machine learning applications in agriculture: Current trends, challenges, and future perspectives. *Agronomy* **2023**, *13*, 2976. [CrossRef]
13. Li, J.; Zhao, X.; Xu, H.; Zhang, L.; Xie, B.; Yan, J.; Zhang, L.; Fan, D.; Li, L. An interpretable high-accuracy method for rice disease detection based on multisource data and transfer learning. *Plants* **2023**, *12*, 3273. [CrossRef] [PubMed]
14. Sekhon, S.S.; Kumar, V.; Patel, A.; Parmar, B.S. Technological advances in smart and sustainable agriculture: The role of Internet of Things, artificial intelligence, big data analysis, machine learning and deep learning. In *Food and Industry 5.0: Transforming the Food System for a Sustainable Future*; Springer: Cham, Switzerland, 2025; pp. 61–71.
15. Guo, X.; Feng, Q.; Guo, F. CMTNet: A hybrid CNN-transformer network for UAV-based hyperspectral crop classification in precision agriculture. *Sci. Rep.* **2025**, *15*, 12383. [CrossRef] [PubMed]
16. Yang, Z.X.; Li, Y.; Wang, R.F.; Hu, P.; Su, W.H. Deep learning in multimodal fusion for sustainable plant care: A comprehensive review. *Sustainability* **2025**, *17*, 5255. [CrossRef]

17. Logeshwaran, J.; Srivastava, D.; Kumar, K.S.; Rex, M.J.; Al-Rasheed, A.; Getahun, M.; Soufiene, B.O. Improving crop production using an agro-deep learning framework in precision agriculture. *BMC Bioinform.* **2024**, *25*, 341. [CrossRef] [PubMed]
18. Yin, S.; Xi, Y.; Zhang, X.; Sun, C.; Mao, Q. Foundation models in agriculture: A comprehensive review. *Agriculture* **2025**, *15*, 847. [CrossRef]
19. Wang, D.; Cao, W.; Zhang, F.; Li, Z.; Xu, S.; Wu, X. A review of deep learning in multiscale agricultural sensing. *Remote Sens.* **2022**, *14*, 559. [CrossRef]
20. Toscano, F.; Fiorentino, C.; Capece, N.; Erra, U.; Travascia, D.; Scopa, A.; Drosos, M.; D'Antonio, P. Unmanned aerial vehicle for precision agriculture: A review. *IEEE Access* **2024**, *12*, 69188–69205. [CrossRef]
21. Lu, M.; Yao, Z.; Wang, D.; Xu, D.; Wang, Z.; Gu, X.; Shen, J.; Peng, C.; Lu, T.; Guo, J.; et al. Non-destructive Monitoring and Modeling of Greenhouse Tomato Growth Based on Multi-sources of RGB, Multispectral, and Thermal Infrared images. *Smart Agric. Technol.* **2026**, *14*, 102171. [CrossRef]
22. Liu, Z.; Zou, Y.; Hu, Z.; Xue, H.; Li, M.; Rao, B. Research on Multi-Modal Fusion Detection Method for Low-Slow-Small UAVs Based on Deep Learning. *Drones* **2025**, *9*, 852. [CrossRef]
23. Ma, H.; Li, X.; Ji, J.; Cui, H.; Shi, Y.; Li, N.; Yang, C. Monitoring indicators for comprehensive growth of summer maize based on UAV remote sensing. *Agronomy* **2023**, *13*, 2888. [CrossRef]
24. Wang, Y.; Yang, Z.; Kootstra, G.; Khan, H.A. The impact of variable illumination on vegetation indices and evaluation of illumination correction methods on chlorophyll content estimation using UAV imagery. *Plant Methods* **2023**, *19*, 51. [CrossRef] [PubMed]
25. Triana-Martinez, J.; Álvarez-Meza, A.; Castellanos-Dominguez, G. Enhancing agricultural data interpretability and visualization with TabNet-driven feature extraction and Local Biplots. *Results Eng.* **2025**, *27*, 106672. [CrossRef]
26. Piao, T.; Shi, H. Consequences of Land Degradation Under Climate Stress in Northern China: Linking Agriculture, Water, and Rural Development. *Land Degrad. Dev.* **2026**, *37*, 2901–2912.
27. Ramani, D.R.; Sujitha, B.B.; Tangade, S. Smart environmental monitoring systems: IoT and sensor-based advancements. In *Environmental Monitoring Using Artificial Intelligence*; Scrivener Publishing LLC: Beverly, MA, USA, 2025; pp. 45–60.
28. Yu, L.; Du, Z.; Li, X.; Zheng, J.; Zhao, Q.; Wu, H.; Yang, Y.; Zhang, Q.; Li, X.; Ma, X.; et al. Enhancing global agricultural monitoring system for climate-smart agriculture. *Clim. Smart Agric.* **2025**, *2*, 100037. [CrossRef]
29. Bhardwaj, R. Environmental Factors Affecting the Crops Growth and Development: An Analytical Study. *Int. J. Psychosoc. Rehabil.* **2024**, *23*, 1160–1167.
30. Ammar, E.E.; Sayed, A.A.; Rabee, M.M.; Keshta, A.E.; Daher, M.G.; Ali, G.A. Environmental and agricultural applications of sensors. In *Handbook of Nanosensors: Materials and Technological Applications*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 1427–1458.
31. Fei, S.; Hassan, M.A.; Xiao, Y.; Su, X.; Chen, Z.; Cheng, Q.; Duan, F.; Chen, R.; Ma, Y. UAV-based multi-sensor data fusion and machine learning algorithm for yield prediction in wheat. *Precis. Agric.* **2023**, *24*, 187–212. [CrossRef] [PubMed]
32. Liu, S.; Peng, W.; Liu, Y.; Zhao, J.; Su, Y.; Zhang, Y. AFCANet: An adaptive feature concatenate attention network for multi-focus image fusion. *J. King Saud Univ.-Comput. Inf. Sci.* **2023**, *35*, 101751. [CrossRef]
33. Huang, X.; Wang, H.; Li, X. A multi-scale semantic feature fusion method for remote sensing crop classification. *Comput. Electron. Agric.* **2024**, *224*, 109185. [CrossRef]
34. Ye, Z.; Yin, S.; Cao, Y.; Wang, Y. AI-driven optimization of agricultural water management for enhanced sustainability. *Sci. Rep.* **2024**, *14*, 25721. [CrossRef] [PubMed]
35. Huang, J.; Hao, X.; Wang, Y.; Song, R.; Mu, Z.; Chu, W.; Papadakis, G.; Niu, S.; Guo, X. Entity-level cross-modal fusion for multimodal chinese agricultural diseases and pests named entity recognition. *Smart Agric. Technol.* **2025**, *12*, 101188. [CrossRef]
36. Caroppo, A.; Giovanni, D.; Leone, A. A systematic review of available multispectral UAV image datasets for precision agriculture applications. *Remote Sens.* **2026**, *18*, 659. [CrossRef]
37. Mancipe-Castro, L.; Gutiérrez-Carvajal, R. Prediction of environment variables in precision agriculture using a sparse model as data fusion strategy. *Inf. Process. Agric.* **2022**, *9*, 171–183. [CrossRef]
38. Wang, Q.; Wang, T. Machine learning-driven sensor fusion for precision agriculture: Current applications and future perspectives. *Sens. Rev.* **2026**, 1–26. [CrossRef]
39. Zhu, Y.; Zhang, S.; Tang, S.; Gao, Q. Research progress and applications of artificial intelligence in agricultural equipment. *Agriculture* **2025**, *15*, 1703. [CrossRef]
40. Gong, R.; Zhang, H.; Li, G.; He, J. Edge computing-enabled smart agriculture: Technical architectures, practical evolution, and bottleneck breakthroughs. *Sensors* **2025**, *25*, 5302. [CrossRef] [PubMed]
41. Junaidi, A.; Hashim, S.Z.M.; Bin Othman, M.S.; Mohamad, M.M.; Alhussian, H.; Abdulkadir, S.J.; Nasser, M.; Bena, Y.A. Deep Learning and Edge Computing in Agriculture: A Comprehensive Review of Recent Trends and Innovations. *IEEE Access* **2025**, *13*, 137464–137490. [CrossRef]

42. Shen, X.; Li, S.; Qiu, F.; Yao, L. A lightweight real-time unified detection model for rice and wheat ears in complex agricultural environments. *Smart Agric. Technol.* **2025**, *11*, 101055. [CrossRef]
43. Jin, P.; Du, W.; Jin, W. Efficient Productivity Prediction Model Based on Edge Data Compression in Smart Farms. *Smart Agric. Technol.* **2025**, *12*, 101242. [CrossRef]
44. Sutherland, L.A.; Banks, E.; Boyce, A.; Martinat, S. *Establishing an Agricultural Knowledge and Innovation System*; Technical report; The James Hutton Institute: Dundee, UK, 2023.
45. Zhao, C. Agricultural knowledge intelligent service technology: A review. *Smart Agric.* **2023**, *5*, 126–148.
46. Sharma, A.; Biswas, A. Benefits and Limitations. *Conservation Agriculture in India: A Paradigm Shift for Sustainable Production*; Taylor and Francis: London, UK, 2022.
47. Yuan, Y.; Chen, L.; Wu, H.; Li, L. Advanced agricultural disease image recognition technologies: A review. *Inf. Process. Agric.* **2022**, *9*, 48–59. [CrossRef]
48. Gitelson, A.A.; Kaufman, Y.J.; Merzlyak, M.N. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sens. Environ.* **1996**, *58*, 289–298. [CrossRef]
49. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
50. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2016; pp. 770–778.
51. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. *Proc. Mach. Learn. Res.* **2019**, *97*, 6105–6114.
52. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
53. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2021; pp. 10012–10022.
54. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
55. Atrey, P.K.; Hossain, M.A.; El Saddik, A.; Kankanhalli, M.S. Multimodal fusion for multimedia analysis: A survey. *Multimed. Syst.* **2010**, *16*, 345–379. [CrossRef]
56. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 423–443. [CrossRef] [PubMed]
57. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
58. Swartout, W.R. Rule-Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project: BG Buchanan and EH Shortliffe, (Addison-Wesley, Reading, MA, 1984); 702 pages, $40.50. *Artif. Intell.* **1985**, *26*, 364–366.
59. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.