agriculture

MDPI

Article

Multimodal Deep Learning for Pest and Disease Recognition and Crop Growth Assessment in Open-Field Agricultural Environments

Jiayu Xiang [1,†], Jianxiang Pan [1,†], Hanwen Zhang [1,†], Xuekun Liu [1,2], Boxiu Liu [2], Jieling Tian [2] and Shuo Yan [1,*]

[1] China Agricultural University, Beijing 100083, China
[2] National School of Development, Peking University, Beijing 100871, China
* Correspondence: yanshuo@cau.edu.cn
† These authors contributed equally to this work.

Abstract

Against the backdrop of the rapid development of smart agriculture, pest and disease monitoring and crop growth assessment for large-scale farmlands are of substantial importance for precision management and risk early warning. However, traditional unimodal visual methods are highly susceptible to illumination variation, canopy occlusion, scale differences, and background interference in real field environments, and thus fail to make full use of environmental sensing information and spatial priors. To address these issues, a multimodal target perception framework for intelligent farmland inspection is proposed in this study. By jointly integrating UAV imagery, time-series data from ground Internet of Things sensors, and spatial positional information, joint modeling of pest and disease recognition and crop growth assessment is achieved through cross-modal alignment and collaborative encoding, multi-scale target perception, and dynamic multimodal fusion and decision-making. Experimental results demonstrate that, in the pest and disease recognition task, the proposed method achieved a Precision of 91.63%, a Recall of 90.27%, an F1-score of 90.94%, and an mAP of 93.15%, significantly outperforming comparison models such as Faster R-CNN with ResNet50 backbone, YOLOv8-m, Swin Transformer-Tiny, and Multimodal Transformer. In the crop growth assessment task, an Accuracy of 89.96%, a Precision of 89.11%, a Recall of 88.74%, and a Macro-F1 of 88.92% were achieved, again clearly exceeding those of ResNet50, EfficientNet-B3, ViT-B/16, and conventional multimodal fusion models. The ablation study further verified the effectiveness of the cross-modal alignment module, the multi-scale target perception module, and the dynamic fusion module, with the complete model reaching 90.94%, 93.15%, and 88.92% in Pest F1, Pest mAP, and Growth Macro-F1, respectively. Furthermore, the net economic return regression experiment at the unit-area level further demonstrates that the proposed method can effectively connect state information with economic outcomes, showing strong application potential in return prediction, performance evaluation, and resource allocation optimization. These findings indicate that the proposed method can effectively improve perception accuracy and robustness in complex farmland environments, thereby providing reliable technical support for intelligent inspection, pest and disease early warning, and precision management in agricultural scenarios.

Keywords: smart agriculture; Internet of Things sensors; pest and disease recognition; precision agriculture; farmland monitoring

Academic Editor: Francesco Marinello

Received: 20 April 2026
Revised: 20 June 2026
Accepted: 25 June 2026
Published: 29 June 2026

1. Introduction

With the intensification of global climate change and the growing pressure on food security, an unprecedented demand has emerged for fine-grained management and intelligent inspection of large-scale farmlands in the context of Smart Agriculture. Among the core tasks of farmland management, early pest and disease monitoring and precise crop growth assessment are of decisive importance for safeguarding the stability of global agricultural productivity and establishing risk early-warning mechanisms [1]. At present, mainstream intelligent inspection paradigms mainly rely on unimodal visual perception methods based on unmanned aerial vehicles (UAVs) or ground cameras. However, in open and highly complex real-world agricultural ecosystems, visual signals are highly susceptible to environmental noise arising from drastic weather fluctuations, complicated illumination conditions, limited viewpoint coverage, and severe crop canopy occlusion. Such inherent environmental sensitivity has imposed severe physical limitations on the detection stability and robustness of unimodal visual models in practical deployment [2]. Zhou et al. [3] proposed a multimodal deep learning framework based on a multi-parameter environmental sensor array, in which high-resolution imagery was integrated with environmental signals such as temperature, humidity, and gas concentration to achieve pest and disease recognition and early warning, and superior performance over unimodal and conventional fusion methods was reported in terms of Accuracy, Precision, Recall, F1-score, and AUC. Huang et al. [4] proposed a multimodal weed infestation prediction framework for efficient farmland management tasks, through which prediction and assessment of weed occurrence severity were achieved, enabling finer characterization of field weed infestation levels and better support for precision management and agricultural decision-making. Liu et al. [5] proposed a multimodal fusion and attention-guided lightweight network framework that integrated RGB imagery, thermal infrared imagery, and environmental sensor information, achieving 91.5% Precision, 89.2% Recall, 90.3% F1-score, and 88.0% mAP@50, while also reaching an inference speed of 25.7 FPS on Jetson Xavier, thereby balancing high accuracy and lightweight deployment capability. Zhang et al. [6] proposed the AgroNVILA framework, in which the AgroOmni multi-view agricultural training set was constructed, and perception–reasoning decoupling was achieved through VCMN and ARPO in the PRD architecture, resulting in a 15.18% improvement over existing advanced models on multi-altitude agricultural reasoning tasks and demonstrating stronger cross-scale spatial understanding and overall agricultural planning capability.

From the perspective of agricultural economics, the occurrence of pests and diseases and fluctuations in crop growth are not merely field production issues, but are more directly related to output value per unit area, net return levels, and input–output efficiency [7]. For operating entities, disease spread, localized yield reduction, and uneven growth often imply declining marginal returns on inputs such as pesticides, water and fertilizer, and labor, and further affect plot-level operational performance and return stability [8]. Especially under large-scale management conditions, the ability to identify high-risk areas in a timely manner, anticipate potential output-value losses, and optimize resource allocation has become an important factor influencing operational decision-making and economic performance [9]. Therefore, intelligent farmland inspection should not only serve crop state recognition itself, but should also provide quantifiable economic support for output-value prediction, loss estimation, and precision management [10]. Although multi-source heterogeneous data have shown substantial potential in agricultural perception, existing studies remain limited by deficiencies in multimodal fusion. First, shallow concatenation ignores the fundamental spatiotemporal mismatch between UAV imagery and ground Internet of Things sensor signals. UAV images are dense spatial observations, whereas temperature, humidity, and soil moisture are sparse temporal measurements. Simple feature- or decision-

level fusion therefore neglects spatiotemporal asynchrony, introduces noise, and weakens the physical correlations among modalities [11]. Second, the semantic gap between visual phenotypes and environmental drivers can readily cause modality collapse. Visual features describe instantaneous crop appearance, whereas sensor data reflect long-term physical conditions. Without deep cross-modal alignment, heterogeneous information is difficult to map into a unified semantic space, and models may become overly dependent on a single surviving modality when illumination changes abruptly or sensors fail [3]. Third, existing methods often struggle to balance micro-scale lesions and large-scale growth patterns, while geospatial coordinates are frequently marginalized, resulting in a gap between local pathological evolution and macro-level spatial distribution [12]. Consequently, the semantic gap across modalities and the lack of systematic collaborative modeling remain key barriers to higher-level intelligent farmland perception.

To address the aforementioned systematic deficiencies, a novel Multimodal Target Perception Framework for intelligent farmland inspection is proposed in this study. The proposed framework is designed to break modality barriers by deeply integrating high-resolution UAV imagery, time-series data from ground IoT sensors, and precise spatial prior information, and to achieve joint perception of complex farmland pest and disease distribution and crop growth evolution through an end-to-end multimodal deep learning architecture. Specifically, rather than merely stacking standard fusion techniques, we introduce an agricultural-domain-specific cross-modal alignment mechanism that explicitly bridges the inherent spatiotemporal mismatch between sparse environmental temporal signals and dense spatial visual grids. Furthermore, we design a multi-scale target collaborative modeling strategy featuring a bi-directional context propagation mechanism to uniquely link pathological cues directly with macro-level crop growth gradients, and an adaptive dynamic fusion module that explicitly evaluates physical sensor reliability to prevent modality collapse. By tailoring these modules to the unique constraints of agricultural data, heterogeneous features are effectively mapped, thereby fundamentally alleviating a series of scientific challenges, including visual signal instability; extreme scale variation of field targets, such as the transition from tiny lesions to large-scale crop canopies; and the difficulty of achieving deep collaboration among multi-source sensing data [13].

In summary, the core academic contributions of this study can be summarized as follows:

1. The first comprehensive multimodal dataset for intelligent farmland inspection was constructed, in which high-resolution UAV imagery, continuous ground sensor monitoring data, and geospatial location features were systematically integrated under real agricultural scenarios, thereby providing an important data foundation for multimodal agricultural AI research.
2. A novel multimodal target perception network was proposed. By designing a cross-modal alignment and collaborative perception module, precise alignment of high-level semantic features across heterogeneous physical space and visual space was achieved.
3. Significant performance gains were achieved. Extensive experiments demonstrated that the proposed model significantly outperformed existing unimodal and conventional multimodal baseline methods in multiple key tasks, including fine-grained pest and disease recognition and crop growth grading assessment, thereby fully validating its superior robustness and generalization capability under complex agricultural environmental variations.
4. From an economic perspective, a net economic return regression analysis at the unit-area level was further conducted, verifying that the proposed method can effectively characterize the relationship between state variation and economic outcomes, and provide quantitative support for return prediction and resource allocation optimization.

2. Related Work

2.1. Vision-Based Farmland Inspection and Pest and Disease Recognition

In recent years, benefiting from the rapid development of UAV remote sensing platforms and computer vision technologies, farmland inspection and pest and disease monitoring have been undergoing a paradigm shift from traditional manual experience-based sampling toward high-throughput and automated phenotypic analysis. Mainstream methods based on deep convolutional neural networks (CNNs) and large-scale visual models, such as Vision Transformers, have achieved remarkably high detection accuracy in laboratory-controlled environments or static and balanced datasets [14]. However, when such unimodal visual models are deployed in real-world and highly unstructured agricultural ecosystems, their perceptual robustness encounters severe physical and algorithmic bottlenecks [15]. First, drastic natural illumination fluctuation and severe crop canopy occlusion are prevalent in farmland environments, resulting in a sharp decline in the signal-to-noise ratio of visual features. Second, the evolution of pests and diseases exhibits extreme scale variation, ranging from early lesions to meter-level macroscopic growth degradation, while two-dimensional imaging from a single viewpoint is highly prone to losing spatial contextual priors [16]. These limitations indicate that reliance solely on visual morphological features is no longer sufficient to satisfy the practical requirements of large-scale and all-weather precision farmland inspection, and that overcoming the inherent limitations of vision urgently requires the introduction of additional physical-dimensional information.

2.2. Applications of Multi-Sensor Data in Smart Agriculture

The dynamic evolution of crop growth and the outbreak of pests and diseases do not constitute isolated visual phenotypic phenomena, but are instead profoundly driven by complex underlying physical and biochemical micro-environments, such as soil temperature and moisture, photosynthetically active radiation, and ambient gas concentration. Consequently, ground environmental sensors and soil sensor networks based on the IoT have been widely deployed in modern smart agriculture, providing continuous monitoring data with high temporal resolution for fine-grained agricultural management [17]. However, a careful examination of existing studies reveals that, in most cases, these high-value sensor data are still treated as independent telemetry monitoring systems. In practical applications, environmental and soil sensor data are mainly reduced and utilized for statistical trend analysis at the macro-regional level, threshold-based warning, or auxiliary decision-making based on simple rules [18]. This research paradigm, which algorithmically separates physical driving variables, namely sensor-based temporal data, from phenotypic outcome variables, namely visual spatial imagery, severely restricts the system's ability to excavate causal features underlying crop pathological evolution, thereby leaving existing agricultural systems without sufficient depth in predictive capability for complex environmental stress conditions.

2.3. Multimodal Fusion and Cross-Modal Learning Methods

To overcome the performance ceiling of unimodal perception and bridge the gap between visual phenotypes and physical environments, multimodal fusion has become one of the most prominent frontiers in the interdisciplinary fields of Earth observation and agricultural informatics. Existing studies have preliminarily attempted to combine visual signals with environmental priors; however, the vast majority of methods still remain at the stage of shallow data-level concatenation, namely early concatenation, or terminal decision-level fusion, namely late fusion [11]. Such coarse fusion mechanisms face three fundamental theoretical deficiencies. First, spatiotemporal asynchrony exists, in that the dense spatial grids of UAV imagery and the sparse temporal sequences of IoT sensors are

naturally misaligned in terms of physical coordinates and sampling frequency. Second, modality heterogeneity and the semantic gap remain unresolved, as explicit cross-modal semantic alignment mechanisms across heterogeneous physical quantities are lacking. Authoritative studies have demonstrated that the absence of such alignment mechanisms can readily lead to modality collapse or modality dominance during training, wherein the network degenerates and becomes excessively dependent on a single intuitive visual modality, thereby completely losing the fault tolerance and robustness that a multimodal system should possess [19]. Third, insufficient modeling of information complementarity persists. Under dynamically changing farmland environments, existing networks lack adaptive dynamic weight allocation strategies and are therefore unable to accurately capture the dynamic collaborative relationships among multi-source heterogeneous data [20]. Therefore, how to achieve cross-modal semantic alignment, multi-scale collaboration, and dynamic adaptive fusion of heterogeneous data within a unified deep learning framework remains a core scientific problem that must be addressed urgently for large-scale intelligent farmland inspection.

2.4. Agricultural Economics Perspectives on Farmland Monitoring and Output Analysis

From the perspective of agricultural economics, the core value of farmland monitoring lies not only in identifying variations in crop status, but also in explaining how such variations further affect output value, net return, and input–output efficiency [21]. Existing agricultural economics studies have generally focused on the effects of climatic shocks, pest and disease losses, resource input, and management practices on production outcomes, and statistical regression, panel analysis, or empirical modeling methods have often been employed to interpret differences in returns [22]. However, such studies usually rely on ex post aggregated data and lack continuous characterization of micro-level field conditions, thereby making it difficult to finely reveal the dynamic transmission relationships among pest and disease evolution, growth differentiation, and economic outcomes [23]. At the same time, economic analyses based solely on remote sensing or sensor variables frequently neglect the important information contained in crop phenotypes, local damage severity, and spatial heterogeneity. Therefore, how to effectively connect farmland perception results with economic output analysis, so that crop status recognition can further serve output prediction, loss assessment, and resource allocation optimization, has become an important issue in the interdisciplinary research field spanning intelligent agriculture and agricultural economics [24].

3. Materials and Methods

3.1. Data Collection

Data collection was conducted in the context of real-world intelligent farmland inspection scenarios, as shown in Table 1, and the study area was located in typical large-scale cultivated fields in Linhe District, Bayannur City, Inner Mongolia Autonomous Region. This region constitutes an important agricultural production area within the Hetao Irrigation District and is characterized by concentrated crop cultivation, contiguous field distribution, relatively standardized field management practices, and representative ecological and environmental variations, thereby providing suitable support for the practical requirements of multimodal farmland perception research. The overall data collection period was set from May to September during 2022 to 2024, continuously covering the seedling stage, vegetative growth stage, vigorous growth stage, and pre-maturity stage of crops, while specifically incorporating critical periods with high pest and disease incidence and pronounced crop growth differentiation. The collected data in this study mainly covered representative crop types distributed across typical large-scale cultivated plots in Linhe District, with a particular focus on maize and wheat, which are two common field crops in

the region. In terms of disease categories, as shown in Figure 1, the dataset mainly involved common field symptoms such as northern corn leaf blight, southern corn leaf blight, and rust in maize, as well as stripe rust, powdery mildew, and leaf blight-like diseases in wheat. In terms of pest categories, the data mainly included leaf notching, perforation, curling, and local damage caused by aphids, armyworms, and corn borers. In addition to clearly defined pest and disease targets, the dataset also contained abnormal growth states such as chlorosis, growth suppression, and sparse canopy distribution, so as to support the crop growth assessment task.

Table 1. Statistics of the multimodal farmland intelligent inspection dataset

Data Type	Data Source	Quantity	Sampling Interval/Unit
UAV images	Low-altitude UAV inspection	18,500 images	1–2 flights per week
Image patch samples	Cropped from UAV images	62,400 patches	Sample-level cropping
Sensor time-series records	IoT monitoring nodes	1,296,000 records	10–30 min
Sensor nodes	Fixed field deployment	36 nodes	Continuous fixed-point collection
Geospatial records	UAV GPS and field mapping	18,500 records	One record per image

Figure 1. Representative disease samples in the dataset, illustrating the visual phenotypic characteristics of northern corn leaf blight, southern corn leaf blight, rust in maize, wheat stripe rust, wheat powdery mildew, and wheat leaf blight-like disease.

UAV image data were primarily acquired through periodic low-altitude remote sensing inspections conducted over the target fields during the study period. The UAV platform was equipped with a 20-megapixel visible-light imaging sensor. Flight missions were executed at a designated altitude of 30 m, utilizing an 80 percent forward overlap and a 70 percent side overlap, which resulted in a ground sampling distance of approximately 0.8 cm per pixel. Repeated multi-flight coverage of the experimental area was performed according to pre-planned flight paths. Each flight mission simultaneously generated orthographic-view images and raw multi-view images, containing information on crop canopy texture, leaf morphological structure, color variation, lesion distribution characteristics, and the overall spatial pattern of the farmland. Image acquisition was mainly arranged during periods of relatively stable illumination each day, while a portion of data collected under cloudy conditions, low-light conditions, and mild wind disturbance was also retained to enhance the adaptability of the subsequent model to complex environmental variations. Under normal conditions, UAV inspections were conducted 1 to 2 times per week; during periods of high pest and disease occurrence and rapid crop growth variation, the sampling frequency was appropriately increased to form a continuous observation sequence spanning the entire growing season.

Ground sensor data were mainly collected from IoT monitoring nodes deployed within the interior and boundary regions of the experimental fields. The acquired data types mainly included air temperature, relative air humidity, soil temperature, volumetric soil moisture content, light intensity, and, in certain areas, environmental variables such as rainfall and wind speed. All sensors operated in a continuous automatic sampling mode, with the sampling interval set from 10 min to 30 min, thereby generating environmental time-series data with high temporal resolution and parallel multivariate records.

Geospatial information was mainly derived from the UAV onboard positioning system, boundary mapping results of the experimental fields, and on-site positioning records of ground sensor nodes. For each UAV image and its subsequently generated cropped samples, the corresponding longitude and latitude coordinates, relative position indices, and plot identifiers were retained; for each sensor node, its fixed deployment location was recorded, and a spatial mapping relationship between the image sampling region and the ground sensor observation point was further established. Meanwhile, pest and disease category labels and crop growth status labels were obtained through manual annotation. Pest and disease labels were jointly determined based on field survey records and image-detail interpretation, whereas crop growth status levels were comprehensively assigned with reference to canopy coverage conditions, leaf color performance, population uniformity, and expert field records.

The entire labeling process was independently conducted by three agricultural experts. To ensure the reliability and consistency of the dataset, inter-annotator agreement was evaluated using the Fleiss kappa metric, achieving an average score of 0.86, which indicates strong consensus among the experts. Furthermore, the final dataset exhibits a comprehensive class distribution to ensure balanced model training. To explicitly address potential class imbalance across the diverse taxonomy of agricultural anomalies, we carefully monitored the sample acquisition across the six diseases and various pest categories. The disease categories comprise 4200 instances of northern corn leaf blight, 3800 instances of southern corn leaf blight, 3500 instances of rust in maize, 4100 instances of wheat stripe rust, 3900 instances of wheat powdery mildew, and 3600 instances of wheat leaf blight-like disease. The pest categories consist of 5500 samples demonstrating aphid damage, 4800 with armyworm damage, and 4500 with corn borer damage. For the crop growth assessment task, the data are distributed into 8000 normal growth samples, 6500 growth suppression samples, 5800 chlorosis samples, and 4200 sparse canopy samples. By maintaining these sample sizes within a relatively tight numerical range, we mitigated the risk of the model disproportionately biasing toward majority classes during optimization. Ultimately, a balanced multimodal farmland intelligent inspection dataset was constructed, simultaneously containing visual images, environmental temporal signals, spatial coordinate priors, and task label information.

3.2. Data Augmentation

In multimodal perception tasks for intelligent farmland inspection, data preprocessing and augmentation are essential not only for improving training stability and generalization capability, but also for enabling effective collaborative modeling of UAV imagery, ground sensor data, and spatial positional information. Since raw multimodal data are often affected by sensor noise, scale inconsistency, temporal asynchrony, spatial mismatch, and environmental disturbances, direct input of unprocessed data can readily cause feature disorder, semantic misalignment across modalities, and unstable optimization. Therefore, systematic processing is required to convert heterogeneous observations into comparable and learnable inputs under a unified framework. For UAV imagery, large-scale remote sensing images are first divided into local samples through sliding-window cropping or region-based tiling so that deep networks can focus on dense crop textures, lesion patterns, and canopy structures while reducing computational burden. If the original image is denoted as I with size $H \times W$, the cropped sub-image can be written as:

$$x_{i,j} = I[is_h : is_h + h,\ js_w : js_w + w]. \tag{1}$$

To reduce truncation effects, overlap between adjacent windows is usually retained. After cropping, scale normalization is performed to remove feature shifts caused by different

flight heights and acquisition resolutions. Let the cropped image be $x \in \mathbb{R}^{h \times w \times c}$ and the normalized output be $\hat{x} \in \mathbb{R}^{H_0 \times W_0 \times c}$; then, the resizing process is expressed as:

$$\hat{x}(u, v) = x\left(\frac{u}{H_0}h, \frac{v}{0}w\right). \tag{2}$$

In practice, bilinear interpolation is adopted:

$$\hat{x}(u, v) = \sum_{m=0}^{1}\sum_{n=0}^{1} \omega_{mn} x(p_m, q_n). \tag{3}$$

Moreover, geometric distortion introduced by lens effects and UAV attitude variation is corrected through camera calibration. Let (x, y) denote the ideal point and (x_d, y_d) denote the distorted point. The radial distortion model is written as:

$$x_d = x\left(1 + k_1 r^2 + k_2 r^4 + k_3 r^6\right), \tag{4}$$

$$y_d = y\left(1 + k_1 r^2 + k_2 r^4 + k_3 r^6\right), \tag{5}$$

where $r^2 = x^2 + y^2$. Such correction improves the structural consistency of crop rows, leaf boundaries, and lesion regions. For ground sensor sequences, preprocessing mainly includes outlier detection, missing-value imputation, and smoothing. Given a sensor time series $\{s_t\}_{t=1}^{T}$, the Z-score can be used for abnormality detection:

$$z_t = \frac{s_t - \mu_s}{\sigma_s}. \tag{6}$$

If missing data occur, linear interpolation can be adopted:

$$\tilde{s}_t = s_{t_a} + \frac{t - t_a}{t_b - t_a}(s_{t_b} - s_{t_a}). \tag{7}$$

To suppress high-frequency noise, moving average smoothing is applied:

$$\bar{s}_t = \frac{1}{L}\sum_{i=t-L+1}^{t} s_i. \tag{8}$$

For multimodal fusion, temporal alignment is performed by aggregating sensor observations around the UAV acquisition time t^{img}:

$$g(t^{img}) = \frac{1}{|\Omega(t^{img})|}\sum_{t_k^{sen} \in \Omega(t^{img})} s_{t_k^{sen}}, \tag{9}$$

where $\Omega(t^{img}) = \{t_k^{sen} : |t_k^{sen} - t^{img}| \le \Delta t\}$. Spatial alignment is established by associating image patches with nearby sensor nodes. If the image-patch center is (x_i, y_i) and the j-th sensor position is (x_j^{sen}, y_j^{sen}), the spatial distance is:

$$d_{ij} = \sqrt{(x_i - x_j^{sen})^2 + (y_i - y_j^{sen})^2}. \tag{10}$$

The corresponding weighted sensor representation is defined as:

$$e_i = \frac{\sum_{j=1}^{M} w_{ij} s_j}{\sum_{j=1}^{M} w_{ij}}. \tag{11}$$

$$w_{ij} = \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right). \tag{12}$$

This strategy improves the precision of spatial correspondence across modalities. Data augmentation is used to simulate realistic disturbances and improve robustness. Geometric transformation can be expressed as:

$$x' = \mathcal{T}_{geo}(x; \theta, \gamma). \tag{13}$$

For geometric transformations, random rotation with an angle $\theta \in [-30°, 30°]$ and random scaling with a factor $\gamma \in [0.8, 1.2]$ were applied with an execution probability of 0.5. Brightness and contrast perturbation can be written as:

$$x' = \beta x + \delta. \tag{14}$$

Specifically, the brightness coefficient β was randomly sampled from $[0.8, 1.2]$ and the contrast shift δ from $[-0.1, 0.1]$, both with an execution probability of 0.5. Occlusion augmentation is implemented through random erasing:

$$x'(u, v) = \begin{cases} 0, & (u, v) \in \mathcal{R} \\ x(u, v), & (u, v) \notin \mathcal{R}. \end{cases} \tag{15}$$

In this process, a random region $\mathcal{R}$ covering 2 percent to 30 percent of the image area was erased with a probability of 0.3. In addition, blur perturbation is introduced as:

$$x' = x * K. \tag{16}$$

Gaussian blur was implemented using a kernel K of size 3×3 or 5×5 with an execution probability of 0.2. To ensure rigorous evaluation of the proposed framework, all the aforementioned data augmentation techniques were applied strictly to the training set only, while the validation and testing sets remained in their original state to reflect actual field deployment conditions. Through these preprocessing and augmentation procedures, richer and more stable multimodal representations can be obtained, thereby improving robustness in real inspection scenarios.

3.3. Proposed Method

3.3.1. Overall

An end-to-end joint perception framework was constructed in this study, consisting of a visual branch, a sensor branch, a positional information encoding branch, a cross-modal alignment and collaborative encoding module, a multi-scale target perception module, and a dynamic multimodal fusion and decision module. The preprocessed UAV images were first fed into the visual backbone network, where hierarchical convolutional operations or multi-level feature extraction generated multi-layer visual representations containing local textures, lesion boundaries, leaf structures, canopy distributions, and regional spatial patterns. Among these representations, shallow features primarily preserved small-target lesions and fine-grained appearance anomalies, whereas deep features focused on expressing crop population growth status and the overall semantic structure of the farmland. In parallel, the aligned environmental sensor time-series data were input into the sensor encoding branch, where temporal modeling and feature compression were performed to obtain environmental semantic representations capable of characterizing the variation patterns of micro-environmental factors such as temperature, humidity, soil moisture, and illumination. Spatial coordinates, plot indices, and node positional relationships were then

introduced into the positional encoding module and mapped into spatial prior features that could participate in deep learning computation. Subsequently, the three modalities of features were projected into a unified semantic space within the cross-modal alignment and collaborative encoding module. Unlike conventional attention-based alignment, our interaction across modalities is strictly anchored by spatial positional priors to establish physically consistent associations among visual phenotypes, environmental drivers, and spatial distributions. In this way, local anomalies in images could be reinterpreted under the constraints of environmental conditions and geographical location, while sensor and positional features were simultaneously supported by visual contextual information. The aligned fusion representations were then delivered to the multi-scale target perception module, where a specialized cross-scale transmission mechanism from local to global and from detail to semantics was established across different feature levels. This specific architectural choice enables the model to simultaneously capture lesion cues and large-scale canopy growth variations while explicitly preserving the structural continuity between microscopic anomalies and macroscopic spatial backgrounds. Finally, the enhanced multimodal features were fed into the dynamic multimodal fusion and decision module. Instead of utilizing conventional static gating or simple concatenation, the network adaptively adjusted the contribution weights of each modality within the joint representation according to a dedicated quality assessment subnetwork that dynamically evaluates the image clarity, sensor reliability, and spatial discriminative value of the current sample. A unified high-level representation for pest and disease recognition and crop growth assessment was thereby generated, and final predictions were produced through task-related output layers. Through this progressively organized structure, from unimodal encoding to customized unified alignment, multi-scale enhancement, and dynamically regularized decision-making, deep collaborative perception of multi-source heterogeneous information under complex farmland scenarios was achieved.

3.3.2. Cross-Modal Alignment and Collaborative Encoding Module

In the proposed cross-modal alignment and collaborative encoding module, the processed visual features, environmental sensor temporal features, and spatial positional information were not directly combined through simple concatenation. Instead, they were first introduced into their respective modality-specific encoding branches and then mapped into a common latent semantic space through a unified projection layer, thereby establishing the foundation for subsequent cross-modal interaction. As shown in Figure 2, let the multi-scale image representation output by the visual branch be denoted as $F^v \in \mathbb{R}^{N_v \times d_v}$, where N_v represents the number of visual tokens and d_v represents the dimensionality of visual features. Let the temporal environmental representation output by the sensor branch be denoted as $F^s \in \mathbb{R}^{N_s \times d_s}$, and let the spatial prior representation output by the positional encoding module be denoted as $F^p \in \mathbb{R}^{N_p \times d_p}$. Owing to the significant differences among the three modalities in terms of original statistical distribution, sampling density, and semantic hierarchy, independent linear projection matrices W_v, W_s, and W_p were first assigned to each modality so that they could be mapped into a common representation space with a unified dimensionality d, yielding $\hat{F}^v = F^v W_v$, $\hat{F}^s = F^s W_s$, and $\hat{F}^p = F^p W_p$. This process essentially corresponds to the projection operation in the structural diagram. Its role is not merely to complete dimensional transformation, but rather to compress visual phenotypes, physical environmental drivers, and spatial priors into high-level semantic representations that are comparable and interactive through learnable mappings, thereby providing the prerequisite for subsequent consistency modeling.

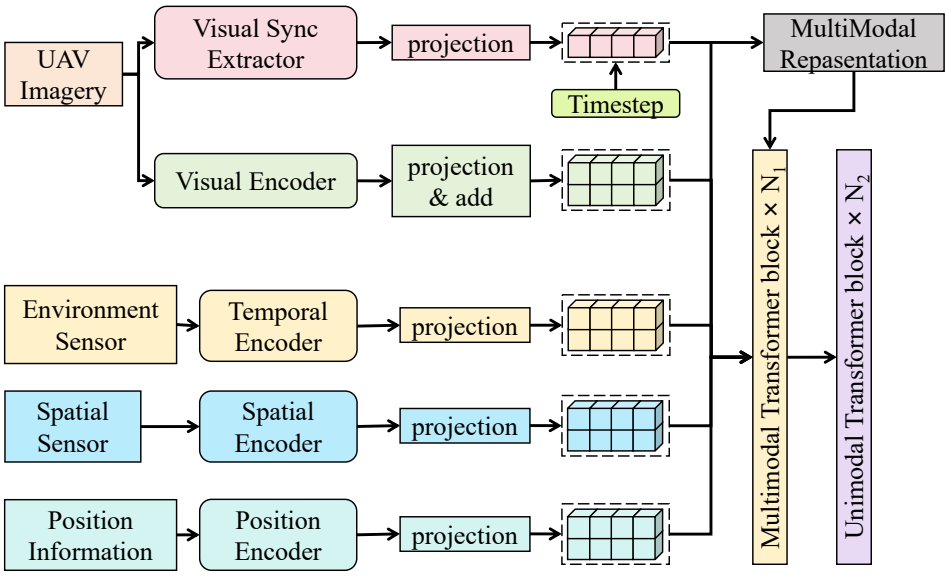

Figure 2. Illustration of the Cross-Modal Alignment and Collaborative Encoding Module.

After unified projection, a cross-modal collaborative encoder was further introduced. Its overall form is consistent with the concept of a multimodal Transformer block shown in the diagram, in which explicit dependencies among modality tokens are established through a shared multi-head attention mechanism. To prevent the visual modality from suppressing the other modalities during training, unidirectional static fusion was not adopted. Instead, a conditional interaction path was constructed in which visual tokens served as the primary queries, while sensor and positional tokens served as keys and values, and the reverse modulation of visual representations by the sensor and positional modalities was simultaneously preserved. Let the query, key, and value be denoted as $Q = \hat{F}^v W_Q$, $K = [\hat{F}^s; \hat{F}^p] W_K$, and $V = [\hat{F}^s; \hat{F}^p] W_V$, respectively. Then, the cross-modal attention can be written as $A_{v \leftarrow sp} = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V$. This term reflects the semantic recalibration capability of environmental conditions and spatial location with respect to local visual regions. Specifically, when certain lesion regions in the image are not visually salient, but the corresponding temporal window exhibits increasing humidity, fluctuating soil moisture, or a location within a high-risk propagation area, the response of that region is enhanced after interaction. Correspondingly, $A_{s \leftarrow vp}$ and $A_{p \leftarrow vs}$ were also constructed to form bidirectional coupling, enabling temporal sensor representations to be interpreted under visual context and allowing spatial priors to acquire task-related spatial semantics under the joint action of vision and environment. The fused output representation is denoted as $H^m = \text{FFN}(\text{LN}([\hat{F}^v + A_{v \leftarrow sp}; \hat{F}^s + A_{s \leftarrow vp}; \hat{F}^p + A_{p \leftarrow vs}]))$, where LN denotes layer normalization and FFN denotes the feed-forward mapping. The residual structure is employed to preserve the original discriminative information of each modality and prevent excessive disruption of unimodal features during cross-modal interaction.

To further suppress semantic drift caused by modality heterogeneity, an explicit alignment constraint was introduced during collaborative encoding, so that the representations of different modalities corresponding to the same spatiotemporal sample remain consistent in the latent space, while representations of different spatiotemporal samples remain distinguishable. Let the global representations of the visual, sensor, and positional modalities in the shared space be denoted as z^v, z^s, and z^p, respectively. Then, a consistency constraint can be imposed by minimizing $\mathcal{L}_{align} = \sum\left(|z^v - z^s|_2^2 + |z^v - z^p|_2^2 + |z^s - z^p|_2^2\right)$. A contrastive constraint based on a similarity matrix may also be further introduced so that the similarity of positive sample pairs is higher than that of negative sample pairs. Such a design has clear mathematical significance, in that the model is forced to learn the joint distribution among

lesion phenotypes, environmental drivers, and spatial locations, rather than relying solely on visual textures for shallow fitting. In other words, if a certain visual anomaly cannot be supported by environmental and positional features, its consistency score in the shared space becomes relatively low and is therefore suppressed during training. Conversely, if a local region simultaneously satisfies visual abnormality, environmental abnormality, and high-risk spatial distribution, its joint representation becomes more concentrated and is thus more favorable for subsequent classification and assessment.

From the perspective of network structure parameters, the key hyperparameters of this module mainly consist of the shared embedding dimension d, the number of attention heads h, the number of cross-modal encoding layers L, and the hidden dimension of the feed-forward network d_f. Among them, d determines the semantic carrying capacity after different modalities enter the unified space, h determines the model capacity to learn visual–environment–location relationships from multiple subspaces, L controls the depth of modality interaction, and d_f determines the upper limit of the nonlinear transformation for expressing complex coupling patterns. Since the tasks addressed in this study simultaneously include fine-grained pest and disease recognition and crop growth assessment, with the former requiring preservation of local fine-grained anomalies and the latter requiring integration of macro-regional semantics, the hierarchical structure composed of unified projection, bidirectional attention, residual updating, and consistency constraints is especially necessary. Its advantages lie in the fact that, on the one hand, instability of the visual modality under complex illumination, occlusion, and scale variation can be alleviated, thereby allowing sensor and positional priors to become important supports for visual discrimination. On the other hand, sensor temporal data are prevented from degenerating into static auxiliary variables and are enabled to participate directly in visual semantic reconstruction and spatial risk modeling. Ultimately, the cross-modal alignment and collaborative encoding module provides a semantically consistent, structurally stable, and interpretable joint representation foundation for subsequent multi-scale target perception and dynamic fusion decision-making, thereby improving the robustness and generalization capability of the model in real farmland environments.

3.3.3. Multi-Scale Target Perception Module

In the proposed multi-scale target perception module, the intermediate features output by the visual backbone network are not directly delivered to the classification or regression head. Instead, they are organized into a hierarchical representation system with top-down and cross-level bidirectional interaction so that local lesions, small-scale texture anomalies, regional growth degradation, and large-scale canopy distribution changes can be simultaneously characterized. The visual features are divided into four levels, denoted as X_1, X_2, X_3, X_4, whose spatial sizes are $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$, and $\frac{H}{32} \times \frac{W}{32}$, respectively, and whose channel dimensions are C_1, C_2, C_3, C_4. Among them, shallow features possess higher spatial resolution and are therefore suitable for preserving lesion boundaries, abrupt leaf color changes, and texture details, whereas deep features exhibit stronger semantic abstraction capability and are thus more suitable for representing population growth status, regional spatial patterns, and background context. To prevent multi-scale information from gradually attenuating only along a single backbone path, two core substructures, namely cross-scale contextual attention and intra-scale content attention, are constructed within each column-wise interaction unit, and this interaction process is repeatedly stacked for m columns so that features from different levels can gradually accomplish sufficient coupling from local to global and from detail to semantics after multiple iterations.

As shown in Figure 3, in the t-th column, the input features from different levels are denoted as $X_l^{(t)}$. First, the channels of all levels are unified into a common dimension $\tilde{C}_l$

through linear mapping or convolutional projection, thereby obtaining the scale-aligned representations $U_l^{(t)}$. For multi-scale contextual interaction, the module not only receives features from the current level but also aggregates semantic information from adjacent and distant scales. Therefore, the context-enhanced representation at the l-th level is defined as

$$\widetilde{U}_l^{(t)} = U_l^{(t)} + \sum_{k \neq l} \alpha_{lk}^{(t)} , \Phi_{k \to l}! \left(U_k^{(t)} \right), \tag{17}$$

where $\Phi_{k \to l}(\cdot)$ denotes the resolution transformation and channel mapping operator from the k-th level to the l-th level. If $k < l$, downsampling aggregation is adopted; if $k > l$, upsampling reconstruction is employed. The coefficient $\alpha_{lk}^{(t)}$ denotes the cross-scale contextual weight, which is generated adaptively from the input and is used to characterize the contribution of different scales to the current level. This structure corresponds to the multi-scale contextual attention module in the diagram. Its essence lies in explicitly connecting microscopic pathological cues with macroscopic spatial backgrounds so that anomalies at any scale do not become isolated from the scene semantics to which they belong. After the cross-scale context-enhanced features are obtained, intra-level content-aware attention is further introduced at each level to strengthen spatial regions and channel responses that are genuinely related to the target. For the feature at the l-th level, the spatial–channel joint recalibration process is defined as

$$Y_l^{(t)} = \Psi_l! \left(\widetilde{U}_l^{(t)} \right) \odot \widetilde{U}_l^{(t)} + \widetilde{U}_l^{(t)}, \tag{18}$$

where $\Psi_l(\cdot)$ denotes the intra-scale content attention generation function and $\odot$ denotes element-wise multiplication. This equation implies that the model does not treat all positions and channels equally within the same level, but instead adaptively emphasizes lesion regions, sparse canopy areas, edge-degraded areas, and similar target-related patterns according to content saliency, while residual connections are used to maintain stable transmission of the original information. Subsequently, $Y_l^{(t)}$ is taken as the input to the next column interaction module, satisfying

$$X_l^{(t+1)} = \Gamma_l! \left(Y_l^{(t)} \right), \tag{19}$$

where $\Gamma_l(\cdot)$ denotes the scale updating unit composed of convolution, normalization, and nonlinear mapping. After m columns of interaction, the final four-level outputs $X_1^{(m)}, X_2^{(m)}, X_3^{(m)}, X_4^{(m)}$ jointly constitute the multi-scale target perception representation and are further delivered to the subsequent dynamic multimodal fusion module. From a theoretical perspective, the effectiveness of this module arises from two aspects. First, let the shallow small-target representation be denoted as δ_s and the deep global representation be denoted as δ_g. Conventional single-scale networks usually optimize only one of them approximately, and thus their risk functions are closer to the single-term minimization of either $\mathcal{R}(\delta_s)$ or $\mathcal{R}(\delta_g)$. By contrast, through Equation (1), cross-scale context is explicitly introduced at each level in this study, so that the final representation satisfies joint constraints from both local detail terms and global structural terms. Accordingly, the optimization objective can be expressed as

$$\mathcal{J} = \mathcal{L}_{task} + \lambda_1 \mathcal{L}_{local} + \lambda_2 \mathcal{L}_{global} + \lambda_3 \mathcal{L}_{cons}, \tag{20}$$

where $\mathcal{L}_{task}$ corresponds to the main-task loss for pest and disease recognition and growth assessment, $\mathcal{L}_{local}$ is used to constrain local small-target discriminative capability, $\mathcal{L}_{global}$ is used to constrain large-scale regional perception capability, and $\mathcal{L} * cons$ is used to

guarantee consistency among different scale representations. It can therefore be seen that, when λ_1, λ_2, and λ_3 are positive, the optimal solution of the model is no longer biased toward a single scale, but instead approaches a joint optimum that simultaneously accommodates local and global information under multiple constraints. Second, since the content attention explicitly amplifies highly responsive regions while the residual term guarantees stability of gradient propagation, shallow details corresponding to early lesions, which are spatially small but semantically critical in farmland scenarios, are prevented from being rapidly overwhelmed during cross-level transmission. Meanwhile, large-scale canopy differences and regional growth gradients can be continuously supplemented back into shallow representations through deep contextual information, thereby preventing the model from misclassifying local anomalies as random noise.

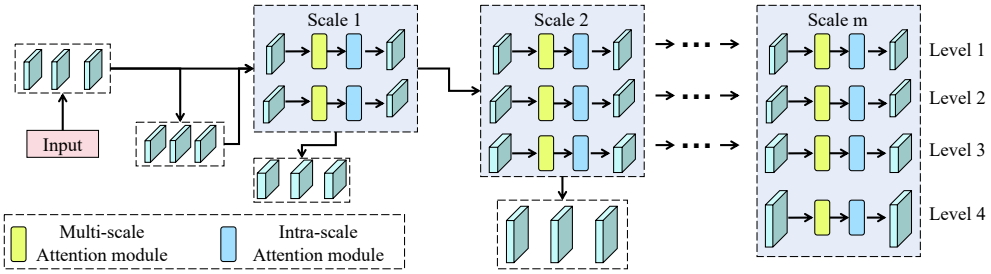

Figure 3. Illustration of the Multi-Scale Target Perception Module.

When applied to the tasks of this study, this design exhibits clear advantages. For pest and disease recognition, the shallow high-resolution branch is capable of more sensitively perceiving tiny lesions, insect-feeding gaps, and local discoloration, while the cross-scale context can further determine, by means of higher-level regional semantics, whether these anomalies belong to genuine pathological scenarios, thereby reducing false detections caused by shadows, reflections, and leaf overlap. For crop growth assessment, deep features are more capable of modeling population canopy structure, coverage uniformity, and spatial gradients within plots, while shallow details provide low-level evidence such as leaf color, edge damage, and texture degradation. Consequently, the final representation possesses not only stronger scale adaptability, but also better consistency with the actual formation mechanism of farmland targets from microscopic pathological evolution to macroscopic growth variation.

3.3.4. Dynamic Multimodal Fusion and Decision Module

In the proposed dynamic multimodal fusion and decision module, the visual representation, environmental sensor representation, and spatial positional representation obtained after cross-modal alignment and multi-scale target perception are no longer combined by means of static concatenation with fixed proportions. Instead, final inference is completed through a dynamic joint mechanism composed of a quality assessment subnetwork, a class-adaptive fusion subnetwork, and task-related decision heads. As shown in Figure 4, let the visual token set formed after pooling and mapping the multi-scale visual output be denoted as $V = v_{i}{}_{i=1}^{N_v}$, let the environment representation output by the sensor branch be denoted as $S = s_{j}{}_{j=1}^{N_s}$, and let the spatial representation output by positional encoding be denoted as $P = p_{k}{}_{k=1}^{N_p}$. The three types of features are first introduced into modality-specific compression layers, and global representations g_v, g_s, and g_p under the unified dimension d are obtained through linear transformation and normalization. Considering that the imaging clarity, sensor integrity, and effectiveness of spatial priors are not constant across different samples in real farmland scenarios, a quality assessment network is first employed to generate a reliability score for each modality. This network is composed of a two-layer

perceptron and gating activation. Its input includes not only the modality feature itself, but also auxiliary statistics composed of attention responses, temporal fluctuations, and spatial dispersion. Thus, the modality score can be expressed as

$$q_m = \sigma!\left(W_m^{(2)}, \phi!\left(W_m^{(1)}[g_m; r_m] + b_m^{(1)}\right) + b_m^{(2)}\right), \quad m \in v, s, p, \tag{21}$$

where r_m denotes the quality descriptor of modality m, $\phi(\cdot)$ denotes a nonlinear mapping, and $\sigma(\cdot)$ denotes the gating function. This design corresponds to the adaptive selection idea driven by semantic conditions in the structural diagram, and its core purpose is to allow the network to first determine which modality is more reliable at the current moment and then decide which modality should be more strongly utilized.

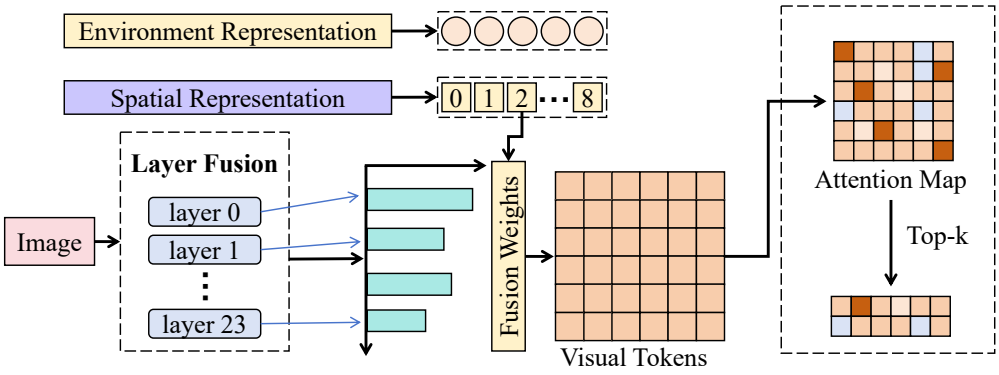

Figure 4. Illustration of the Dynamic Multimodal Fusion and Decision Module.

After modality quality estimation is obtained, a class-adaptive fusion mechanism is further constructed. Since the dependencies of pest and disease recognition and crop growth assessment on modalities are not entirely identical, local lesion recognition usually relies more heavily on shallow visual details and local environmental perturbations, whereas growth assessment relies more on deep visual semantics, continuous environmental trends, and spatial distribution relationships. Therefore, a class-conditional vector c is introduced as a task-driven signal, generated by a lightweight classifier from the shared semantic representation and then fed back to modulate modality weights. The fusion coefficient is defined as

$$\alpha_m = \frac{\exp!\left(\omega_m^\top [q_m; c]\right)}{\sum_{u \in v,s,p} \exp!\left(\omega_u^\top [q_u; c]\right)}, \tag{22}$$

where α_m denotes the final contribution weight of modality m under the current sample and current task semantics. Accordingly, the unified fusion representation is written as

$$z = \alpha_v, \Pi_v(V) + \alpha_s, \Pi_s(S) + \alpha_p, \Pi_p(P), \tag{23}$$

where $\Pi_v(\cdot)$, $\Pi_s(\cdot)$, and $\Pi_p(\cdot)$ denote the modality-specific aggregation operators. For the visual modality, all tokens are not directly summed with equal weights. Instead, inspired by the class-adaptive selection idea shown in the structural diagram, a token importance map is first generated according to task semantics, after which high-response regions are preserved and redundant responses are compressed by clustering. Therefore, the visual aggregation can be written as

$$\Pi_v(V) = \sum_{i=1}^{N_v} \beta_i v_i, \qquad \beta_i = \frac{\exp(\tau_i)}{\sum_{t=1}^{N_v} \exp(\tau_t)}, \tag{24}$$

where τ_i is jointly determined by the visual token and the class-conditional vector c. The mathematical implication of this design is that the fusion process no longer responds equally to all visual regions, but instead prioritizes lesion regions, canopy anomaly areas, or edge gradient areas that are most relevant to the current task, thereby reducing the interference of complex background noise in multimodal fusion.

At the decision stage, the unified representation z is introduced into the shared backbone and is then split into two task heads, producing pest and disease category probabilities and crop growth grade distributions, respectively. If the prediction functions of the two tasks are denoted as $f_d(\cdot)$ and $f_g(\cdot)$, respectively, then

$$\hat{y}_d = f_d(z), \qquad \hat{y}_g = f_g(z). \tag{25}$$

To ensure that dynamic fusion learns genuine complementary relationships rather than incidental biases, the training objective jointly optimizes the task loss and a modality-balance constraint, so that weight allocation both follows sample differences and avoids collapsing excessively toward a single modality. The objective function can be expressed as

$$\mathcal{L} = \mathcal{L}_d + \lambda \mathcal{L}_g + \mu \sum_m \alpha_m \log \alpha_m, \tag{26}$$

where the last term is used to suppress excessively concentrated modality allocation and enhance the robustness of the system in open environments. It can therefore be seen that, when the quality of a certain modality decreases, its q_m also decreases, which further leads to a reduction in α_m, and thus its disturbance to the final representation is naturally weakened. By contrast, when visual evidence is consistent with environmental and spatial priors, the corresponding modality weights increase collaboratively, thereby enhancing the confidence of correct discrimination. For the tasks investigated in this study, the advantages of such a design are highly evident. Under conditions such as shadow occlusion, weak-light imaging, short-term sensor drift, or inconspicuous local spatial priors, the model can automatically reallocate information sources according to the current sample state, thereby avoiding the accumulation of misclassification caused by static fusion. Meanwhile, the class-adaptive mechanism further enables pest and disease recognition and growth assessment to emphasize their most critical information sources on the basis of a shared multimodal representation, thereby simultaneously improving fine-grained anomaly detection capability and large-scale regional state assessment capability.

4. Results and Discussion

4.1. Experimental Configuration

4.1.1. Hardware and Software Platform

From a hardware perspective, the experiments were conducted on a high-performance deep learning workstation. The server-side processor adopted an Intel Xeon or an equivalent multi-core CPU, with a clock-frequency configuration sufficient to support large-scale data preprocessing and parallel loading. The graphics computing unit employed an NVIDIA GeForce RTX 4090 GPU with 24 GB of video memory, which was capable of supporting the training and inference of the multimodal target perception network under relatively large input resolutions and relatively complex fusion structures. The system memory was configured as 128 GB to ensure stable UAV image cropping, multi-source sensor data buffering, and batch data reading during training. High-speed solid-state drives were used for dataset reading and writing as well as model parameter storage, thereby reducing I/O bottlenecks during the joint loading of large-scale farmland imagery and sensor sequences. The overall hardware environment was able to satisfy the computational demands of

multimodal feature extraction, cross-modal fusion training, and multiple rounds of ablation experiments, comparative experiments, and cross-validation, while also providing reliable support for model convergence efficiency and experimental reproducibility.

From a software perspective, the experimental environment was established on a 64-bit Ubuntu 22.04 LTS operating system. PyTorch 2.2 was employed as the deep learning framework, Python 3.10 was adopted as the programming language, CUDA 12.1 was used, and the cuDNN version was maintained to be compatible with CUDA so that GPU parallel computing capability could be fully exploited. Data processing and scientific computing were mainly completed by using NumPy, Pandas, and SciPy. Image preprocessing and augmentation were implemented by combining OpenCV and Pillow, while UAV image cropping, distortion correction, and scale normalization were conducted within a unified data pipeline. The model training procedure was implemented by integrating Torchvision and customized modules. The multimodal temporal and spatial alignment procedures were completed through Python scripts for sensor data parsing, coordinate mapping, and sample organization. Experimental result statistics and visualization were mainly performed with Matplotlib 3.10.3 and Scikit-learn. The entire software environment was executed under fixed random seeds in order to reduce experimental fluctuation and improve result reproducibility.

In terms of hyperparameter configuration, the constructed multimodal farmland inspection dataset was divided into a training set, a validation set, and a test set according to a ratio of 7:2:1, where the training set was used for model parameter learning, the validation set was used for model selection and early stopping, and the test set was used for final performance evaluation. To further verify the stability and generalization capability of the model, a 5-fold cross-validation strategy was additionally adopted beyond the main experiments [25]. Specifically, all samples were divided into 5 mutually exclusive subsets, and one subset was selected in each round as the validation or test portion, while the remaining subsets were used for training. The average of the 5 rounds was then reported as a supplementary evaluation. During training, AdamW was adopted as the optimizer, the initial learning rate was set to 1×10^{-4}, the weight decay coefficient was set to 1×10^{-2}, the batch size was set to 16, and the number of training epochs was set to 100. An early stopping mechanism was employed when the validation metric no longer improved for several consecutive epochs, so as to prevent overfitting. A cosine annealing strategy was used for learning-rate scheduling, enabling the learning rate to decay smoothly during training. The momentum-related parameters β_1 and β_2 were set to 0.9 and 0.999, respectively. For the input data, UAV images were uniformly resized to 640×640, and sensor features were standardized prior to input so that they followed zero-mean and unit-variance distributions. The hidden-layer dimension in the multimodal fusion module was set to 256, which corresponds to the shared embedding dimension d in the cross-modal alignment and collaborative encoding module. Furthermore, for the collaborative encoding process, the number of attention heads h was set to 8, the number of cross-modal encoding layers L was set to 4, and the hidden dimension of the feed-forward network was configured as 512. The Dropout ratio was set to 0.3 in order to enhance model robustness and suppress overfitting. To ensure experimental fairness, all comparative methods were trained and tested under the same data split, the same number of training epochs, and the same evaluation criteria, thereby ensuring the objectivity and reliability of performance comparisons.

4.1.2. Baseline Models and Evaluation Metrics

Faster R-CNN with ResNet50 backbone [26], YOLOv8-m [27], RetinaNet with ResNet101 backbone [28], DETR with ResNet50 backbone [29], Swin Transformer-Tiny [30],

ViT-B/16 [31], ResNet50 + Sensor Feature Concatenation [32], Swin-T + Late Fusion MLP [30], and Multimodal Transformer (Image+Sensor+Location) [33] were selected to represent different technical routes, ranging from unimodal visual detection to multimodal fusion modeling. To ensure experimental transparency and reproducibility, the detailed hyperparameter settings of all baseline models are systematically summarized in Table 2. Among them, Faster R-CNN with ResNet50 backbone is a two-stage detection method based on a ResNet50 backbone network, in which candidate regions are first generated, and classification and regression are subsequently completed. Its advantage lies in its relatively high detection accuracy and strong adaptability to complex backgrounds. YOLOv8-m is a medium-scale single-stage object detection model that directly outputs target categories and locations in an end-to-end manner, and its advantage lies in a favorable balance between speed and accuracy. RetinaNet with ResNet101 backbone is a single-stage detection method that combines ResNet101-based feature extraction with focal loss, and its advantage lies in its better handling of positive–negative sample imbalance. DETR with ResNet50 backbone uses ResNet50 to extract visual features and models the detection task as a set prediction problem through a Transformer, and its advantage lies in strong global relation modeling capability. Swin Transformer-Tiny is a lightweight hierarchical visual Transformer that extracts multi-scale features through a window attention mechanism, and its advantage lies in high computational efficiency and suitability for vision-dense tasks. ViT-B/16 is a basic visual model in which an image is divided into 16×16 patches and then encoded by a Transformer, and its advantage lies in its remarkable capability for long-range dependency modeling. ResNet50 + Sensor Feature Concatenation is a method in which sensor features are directly concatenated with image features extracted by ResNet50 for joint prediction, and its advantage lies in its simple implementation and preliminary exploitation of multimodal complementary information. Swin-T + Late Fusion MLP is a method in which image features extracted by Swin-Tiny are integrated with sensor information through a late-fusion multilayer perceptron, and its advantage lies in the flexibility of the fusion process and the convenience of independent modeling for different modalities. Multimodal Transformer (Image+Sensor+Location) is a method in which image, sensor, and positional information are jointly introduced into a Transformer framework for collaborative modeling, and its advantage lies in its stronger capability to exploit the associations and complementarity among multiple modalities.

Table 2. Summary of hyperparameter settings for the baseline models.

Baseline Model	Optimizer	Learning Rate	Batch Size	Weight Decay
Faster R-CNN with ResNet50	AdamW	1×10^{-4}	16	1×10^{-2}
YOLOv8-m	AdamW	1×10^{-4}	16	1×10^{-2}
RetinaNet with ResNet101	AdamW	1×10^{-4}	16	1×10^{-2}
DETR with ResNet50	AdamW	1×10^{-4}	16	1×10^{-2}
Swin Transformer-Tiny	AdamW	1×10^{-4}	16	1×10^{-2}
ViT-B/16	AdamW	1×10^{-4}	16	1×10^{-2}
ResNet50 + Sensor Feature Concatenation	AdamW	1×10^{-4}	16	1×10^{-2}
Swin-T + Late Fusion MLP	AdamW	1×10^{-4}	16	1×10^{-2}
Multimodal Transformer	AdamW	1×10^{-4}	16	1×10^{-2}

Precision is used to measure the proportion of true-positive samples among all samples predicted as positive, thereby reflecting the ability to control false detections. Recall is used to measure the proportion of true-positive samples that are correctly identified, thereby reflecting the ability to control missed detections. $F1$ is used to comprehensively evaluate the balance between Precision and Recall. mAP is used to comprehensively measure the

overall detection accuracy of the model across different categories or different thresholds. Their mathematical expressions are given as follows:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{27}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{28}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{29}$$

$$AP = \int_0^1 P(R), dR, \tag{30}$$

$$mAP = \frac{1}{C} \sum_{i=1}^{C} AP_i, \tag{31}$$

where TP denotes true positives, FP denotes false positives, and FN denotes false negatives; Precision denotes precision, and Recall denotes recall; $P(R)$ denotes the precision curve with recall R as the independent variable; AP_i denotes the average precision of the i-th category; C denotes the total number of categories; and mAP denotes the mean of the average precision across all categories.

4.2. Pest and Disease Recognition

The purpose of this experiment was to systematically validate the effectiveness of the proposed method in farmland pest and disease recognition and to further analyze the differences in adaptability among various model categories under complex field scenarios.

As shown in Table 3 and Figure 5, traditional convolution-based detection models exhibited relatively stable overall performance, although their upper performance limits remained constrained. Among them, Faster R-CNN with a ResNet50 backbone achieved Precision, Recall, F1-score, and mAP values of 84.37%, 82.94%, 83.65%, and 85.12%, respectively, indicating that although the two-stage detection framework possessed strong region proposal modeling capability, insufficient recall was still observed in farmland scenarios characterized by large lesion scale variation and strong background interference. The overall performance of YOLOv8-m increased to 86.45%, 84.68%, 85.56%, and 87.03%, indicating that the single-stage detector exhibited stronger advantages in feature aggregation and real-time perception. RetinaNet with a ResNet101 backbone performed slightly better than Faster R-CNN, but remained limited by the local receptive field modeling mechanism inherent to convolution-based architectures. The results of DETR with a ResNet50 backbone and Swin Transformer-Tiny were further improved, among which Swin Transformer-Tiny achieved an F1-score of 87.10% and an mAP of 88.94%, suggesting that models based on self-attention or hierarchical window mechanisms could more effectively exploit long-range dependencies and regional context, and were therefore more sensitive to complex canopy occlusion and irregular lesion distributions. By contrast, ViT-B/16 underperformed relative to Swin Transformer-Tiny, indicating that purely global patch-based modeling tends to lose fine local texture details in agricultural small-target scenarios and is therefore not advantageous for early fine-grained lesion recognition.

From a theoretical perspective, the core trend underlying the performance improvement lies in the progressive transition from unimodal visual perception to multimodal collaborative perception. ResNet50 + Sensor Feature Concatenation improved the F1-score to 87.32%, indicating that even simple feature concatenation allowed environmental sensor information to provide additional prior support for visual discrimination. Swin-T + Late Fusion MLP further achieved an F1-score of 87.90% and an mAP of 90.08%, indicating that

a stronger visual backbone combined with decision-level fusion could alleviate part of the visual noise problem. Multimodal Transformer (Image+Sensor+Location) achieved an F1-score of 88.79% and an mAP of 90.84%, indicating that the cross-modal attention mechanism could establish dependencies among images, environment, and spatial location in a unified representation space, thereby improving the discrimination of actual diseased regions. The proposed method ultimately achieved Precision, Recall, F1-score, and mAP values of 91.63%, 90.27%, 90.94%, and 93.15%, significantly outperforming all comparison models. This superiority can be attributed to the fact that the proposed method does not rely solely on static modality fusion, but instead integrates cross-modal alignment, multi-scale target modeling, and dynamic weight allocation, enabling local visual responses of tiny lesions, environmental driving signals, and spatial propagation priors to form complementary relationships within a unified framework. In terms of mathematical characteristics, traditional convolutional models tend to favor local stationary feature fitting, whereas standard Transformer architectures are more inclined toward global relation modeling. By contrast, the proposed method establishes conditional associations and adaptive reweighting across different scales and modalities, thereby simultaneously enhancing intra-class compactness and inter-class separability in the representation space, which enables higher recognition accuracy and more robust recall performance under complex farmland environments.

Table 3. Performance comparison of different methods on the pest and disease recognition task. Results are presented as mean ± standard deviation. Significance levels compared to the best baseline are denoted by *** ($p < 0.001$).

Method	Precision (%)	Recall (%)	F1-Score (%)	mAP (%)
Faster R-CNN with ResNet50 backbone	84.37 ± 0.52	82.94 ± 0.61	83.65 ± 0.48	85.12 ± 0.55
YOLOv8-m	86.45 ± 0.45	84.68 ± 0.50	85.56 ± 0.41	87.03 ± 0.47
RetinaNet with ResNet101 backbone	85.74 ± 0.48	84.13 ± 0.55	84.93 ± 0.49	86.41 ± 0.51
DETR with ResNet50 backbone	87.21 ± 0.42	85.74 ± 0.47	86.47 ± 0.39	88.26 ± 0.44
Swin Transformer-Tiny	87.86 ± 0.38	86.35 ± 0.41	87.10 ± 0.35	88.94 ± 0.40
ViT-B/16	85.96 ± 0.51	84.11 ± 0.58	85.02 ± 0.46	86.75 ± 0.53
ResNet50 + Sensor Feature Concatenation	88.14 ± 0.36	86.52 ± 0.39	87.32 ± 0.33	89.41 ± 0.38
Swin-T + Late Fusion MLP	88.76 ± 0.31	87.05 ± 0.34	87.90 ± 0.28	90.08 ± 0.32
Multimodal Transformer (Image+Sensor+Location)	89.42 ± 0.28	88.16 ± 0.31	88.79 ± 0.25	90.84 ± 0.29
Proposed Method	91.63 ± 0.18 ***	90.27 ± 0.22 ***	90.94 ± 0.15 ***	93.15 ± 0.20 ***

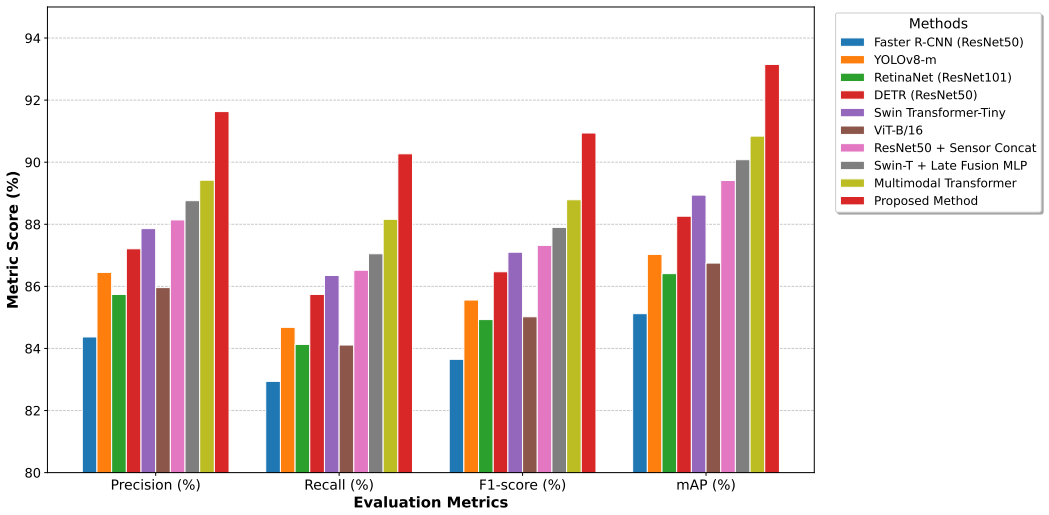

Figure 5. The comparison of different methods on four evaluation metrics, namely Precision, Recall, F1-score, and mAP, for the pest and disease recognition task.

To further delineate the per-class recognition performance, we present the confusion matrix of the proposed method in Figure 6. The model demonstrates highly stable discrimination across all six disease categories and pest variations, with true-positive rates generally exceeding 90 percent. Specifically, northern corn leaf blight and wheat leaf blight-like dis-

ease achieved recognition rates of 92.6 percent and 94.0 percent, respectively, indicating robust feature extraction for distinct macroscopic lesions. However, a detailed analysis of the misclassifications reveals specific failure cases that highlight the remaining challenges in practical field deployments. For instance, the confusion matrix indicates a 2.8 percent misclassification rate between rust in maize and northern corn leaf blight. A review of these specific failure cases shows that during the extremely early stages of infection, before the characteristic pustules of rust fully mature, the preliminary visual discoloration strongly mimics the early symptoms of leaf blight. Additionally, under severe canopy occlusion and overlapping leaves, the model occasionally misses highly localized pest damage, leading to misclassifications where affected areas are treated as background structural shadows. These failure cases suggest that while our multimodal fusion effectively compensates for many environmental disturbances, relying on visible-light imagery still presents inherent physical limitations when pathological symptoms are entirely obscured or morphologically ambiguous in their incipient stages.

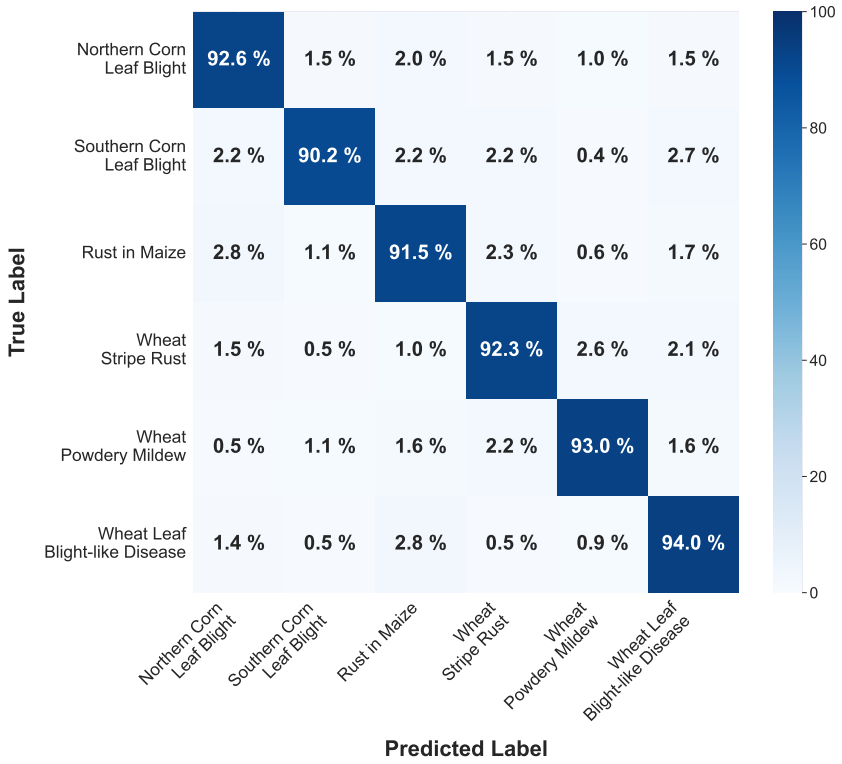

Figure 6. Confusion matrix results of the proposed method.

4.3. Crop Growth Status Assessment

The purpose of this experiment was to validate the effectiveness of the proposed method in the crop growth status assessment task and to analyze the capability differences among different model categories in population growth representation, regional status discrimination, and adaptation to complex field backgrounds.

As shown in Table 4 and Figure 7, unimodal visual models exhibited relatively stable baseline performance overall. Among them, ResNet50 achieved Accuracy, Precision, Recall, and Macro-F1 values of 81.74%, 80.96%, 80.41%, and 80.68%, respectively, indicating that traditional convolutional networks could extract a certain amount of growth-related information from canopy texture, color distribution, and local structure, although their capability for modeling cross-regional scale variation and long-range context remained limited. EfficientNet-B3 and DenseNet121 performed slightly better than ResNet50, suggesting that a more efficient compound scaling strategy and more sufficient feature reuse mechanism

contributed to improved representation of growth differences. ViT-B/16, Swin Transformer-Tiny, and MLP-Mixer-B/16 further improved the overall performance, among which Swin Transformer-Tiny achieved an Accuracy of 84.67% and a Macro-F1 of 83.80%, outperforming other visual models. This indicates that hierarchical window attention can better integrate mid-range and long-range spatial context while preserving local structural perception, and is therefore more suitable for describing canopy coverage, population uniformity, and regional gradient variation involved in growth assessment. In contrast, Sensor-only MLP achieved only 76.34% Accuracy and 75.15% Macro-F1, indicating that although temporal environmental variables such as temperature, humidity, and soil moisture can reflect growth-driving conditions, they are insufficient for directly characterizing the actual phenotypic state, and therefore show obvious limitations in crop growth grade discrimination.

Table 4. Performance comparison of different methods on the crop growth status assessment task. The best results are highlighted in bold.

Method	Accuracy (%)	Precision (%)	Recall (%)	Macro-F1 (%)
ResNet50	81.74	80.96	80.41	80.68
EfficientNet-B3	82.53	81.88	81.17	81.52
DenseNet121	82.16	81.47	80.96	81.21
ViT-B/16	83.96	83.11	82.75	82.93
Swin Transformer-Tiny	84.67	84.02	83.58	83.80
MLP-Mixer-B/16	83.41	82.74	82.19	82.46
Sensor-only MLP	76.34	75.42	74.88	75.15
ResNet50 + Sensor Feature Concatenation	85.52	84.90	84.31	84.60
ViT-B/16 + Late Fusion MLP	86.37	85.74	85.28	85.51
Multimodal Transformer (Image+Sensor+Location)	87.48	86.95	86.37	86.66
Proposed Method	**89.96**	**89.11**	**88.74**	**88.92**

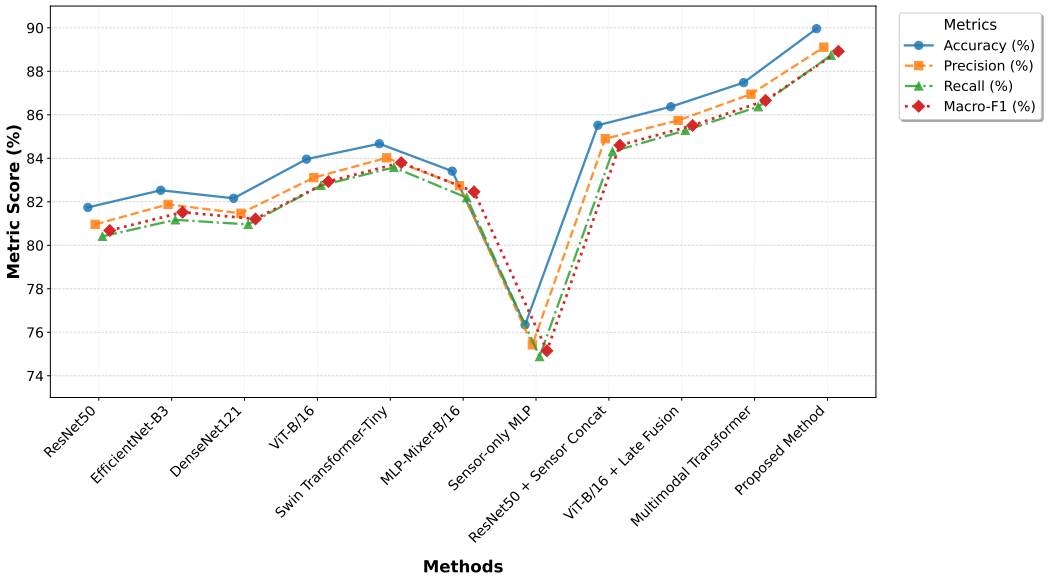

Figure 7. The comparison of different methods on four metrics, namely Accuracy, Precision, Recall, and Macro-F1, for the crop growth status assessment task.

From a theoretical perspective, crop growth status assessment is essentially a comprehensive discrimination problem involving both phenotypic spatial structure and environmentally driven background information. Therefore, the improvement trend in model performance reflects the effectiveness of the transition from single-phenotype recognition to multimodal joint modeling. ResNet50 + Sensor Feature Concatenation improved Macro-F1 to 84.60%, indicating that even a relatively simple modality concatenation strategy allowed environmental sensor information to provide auxiliary support for the interpretation of visual phenotypes. ViT-B/16 + Late Fusion MLP further achieved 85.51%, indicating

that stronger global visual modeling capability combined with decision-level fusion enabled more effective integration of multi-source information. Multimodal Transformer (Image+Sensor+Location) achieved an Accuracy of 87.48% and a Macro-F1 of 86.66%, indicating that the cross-modal attention mechanism could establish deeper semantic associations among images, environment, and spatial location, thereby allowing more accurate discrimination of different growth grades. The proposed method ultimately achieved 89.96%, 89.11%, 88.74%, and 88.92%, significantly outperforming all comparison models. This superiority fundamentally arises from the fact that the proposed method not only integrates multimodal information, but also enables local leaf color variation, overall canopy structure, environmental fluctuation trends, and spatial distribution priors to form stable collaboration within a unified representation space through cross-modal alignment, multi-scale target perception, and dynamic weight adjustment. In terms of mathematical properties, traditional convolutional models are more inclined toward local stationary pattern fitting, whereas standard Transformer architectures are more adept at global dependency modeling. The proposed method further strengthens the conditional coupling and adaptive weighting capabilities among different modalities and across different scales, thereby simultaneously improving intra-class consistency and inter-class separability, and ultimately achieving higher assessment accuracy and more robust overall performance in complex farmland scenarios.

4.4. Ablation Study

The purpose of this ablation experiment was to validate the independent contributions and collaborative effects of the three core modules proposed in this study within the overall framework, and to further analyze which category of structural improvement was responsible for the observed performance gains.

As shown in Table 5 and Figure 8, after Multimodal Transformer (Image+Sensor+Location) was adopted as the unified baseline, it can be clearly observed that all three modules brought stable gains, indicating that the improvement achieved by the proposed method did not arise from accidental parameter stacking, but rather from the structural design itself. The baseline model already possessed a certain degree of cross-modal modeling capability, achieving a Pest F1 of 88.79% and a Pest mAP of 90.84% on the pest and disease recognition task, and a Growth Macro-F1 of 86.66% on the crop growth assessment task. After CMA was introduced, all metrics increased to 89.46%, 91.52%, and 87.31%, indicating that cross-modal alignment and collaborative encoding can effectively enhance semantic consistency among images, sensors, and spatial location. When only MSP was introduced, Pest F1 and Pest mAP reached 89.38% and 91.41%, respectively, while Growth Macro-F1 increased to 87.44%, indicating that multi-scale target perception provides particularly strong benefits for growth assessment, because this task relies more heavily on the joint modeling of local texture and global canopy structure. Stable improvements were also obtained when only DMF was introduced, indicating that the dynamic fusion strategy can adjust modality weights according to sample conditions, thereby reducing interference from unreliable modalities. Furthermore, all pairwise combinations, including CMA + MSP, CMA + DMF, and MSP + DMF, produced results significantly higher than those of the single-module variants. Among them, CMA + MSP performed most prominently in pest and disease recognition, whereas MSP + DMF maintained particularly strong performance in growth assessment. Ultimately, the complete model integrating all three modules achieved 90.94%, 93.15%, and 88.92%, obtaining the best results across all metrics.

From a theoretical perspective, these results reflect the complementarity of the three modules in terms of their mathematical functions. The core role of CMA lies in reducing representation offsets among heterogeneous modalities, thereby enabling visual pheno-

types, environmental drivers, and spatial priors to form stronger intra-class aggregation in a shared representation space. Therefore, this module provides foundational gains for both pest and disease recognition and growth assessment. The role of MSP is to enhance feature responsiveness to targets of different scales, so that tiny lesions are prevented from being submerged during deep propagation, while large-scale canopy structures and regional growth gradients can in turn constrain local judgments. Accordingly, this module is particularly critical for the tasks addressed in this study, which simultaneously involve small-target recognition and macro-level state assessment. DMF, by contrast, suppresses noise propagation through sample-adaptive modality weighting, allowing the model to maintain stable outputs under conditions such as weak illumination, occlusion, and sensor fluctuation. In terms of mathematical properties, the use of any single module alone optimizes only one aspect of representation learning, whereas pairwise combinations allow different modules to begin interacting across representation space, scale space, and modality weight space, thereby producing more substantial improvements. The complete model is optimal because it simultaneously accomplishes cross-modal semantic correction, multi-scale structural enhancement, and dynamic decision reallocation, allowing feature discriminability, stability, and robustness to be jointly improved within a unified framework, and thereby yielding the best experimental performance.

Table 5. Ablation study of the proposed modules based on the Multimodal Transformer (Image+Sensor+Location). CMA denotes the cross-modal alignment and collaborative encoding module, MSP denotes the multi-scale target perception module, and DMF denotes the dynamic multimodal fusion and decision module. The best results are highlighted in bold.

Method	CMA	MSP	DMF	Pest F1	Pest mAP	Growth Macro-F1
Multimodal Transformer				88.79	90.84	86.66
+Cross-modal Alignment	✓			89.46	91.52	87.31
+Multi-scale Target Perception		✓		89.38	91.41	87.44
+Dynamic Multimodal Fusion and Decision			✓	89.21	91.16	87.18
+CMA + MSP	✓	✓		90.07	92.18	88.03
+CMA + DMF	✓		✓	89.94	91.97	87.88
+MSP + DMF		✓	✓	89.86	91.84	87.95
Proposed Method (CMA + MSP + DMF)	✓	✓	✓	**90.94**	**93.15**	**88.92**

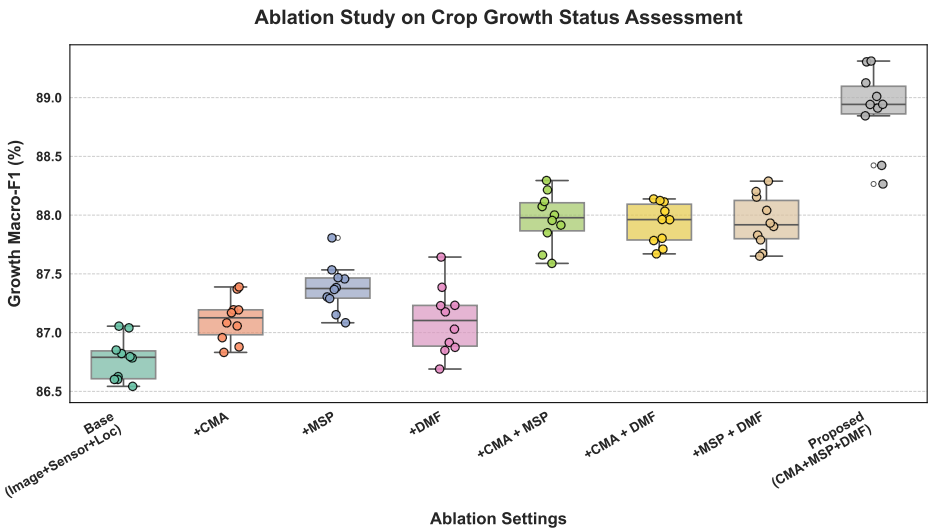

Figure 8. The ablation results on the crop growth status assessment task.

4.5. Agricultural Economic Regression Analysis

The purpose of this experiment was to further verify the application value of the proposed method in economic analysis scenarios and to examine whether variation in economic return per unit area can be effectively explained and predicted. To address the

fundamental economic aspects of this study and provide transparency in our modeling, the target variable, net economic return per unit area, is strictly defined as the difference between total agricultural revenue and total variable input costs, measured in units of thousands of Chinese Yuan per hectare. The economic data were rigorously compiled from three primary sources including detailed field management logs provided by local agricultural cooperatives in Linhe District, formal financial audit records from regional cooperative organizations, and the historical price database of the Inner Mongolia Agricultural Market.

To ensure a standardized and objective assessment, the total input costs were calculated as a sum of several dynamic components directly related to field operations. These include seed procurement expenses, chemical fertilizers, targeted pesticides used for disease control, irrigation water fees, machinery fuel and maintenance costs, and manual labor wages. Fixed costs, such as land rental fees and depreciation of long-term infrastructure, were treated as constant factors across the study area to isolate the direct impact of crop health and growth status on operational performance. The total revenue for each plot was calculated based on the actual measured yield, which was obtained through a combination of standardized manual sampling and verified machine-harvested weight records. This yield was then multiplied by the weighted average local purchasing price during the peak harvest season to minimize the interference of short-term market price fluctuations. Furthermore, to verify the reliability of the model for actual decision support applications, the regression models were independently validated using a completely separate dataset collected from different plots, preventing data leakage and overfitting.

As shown in Table 6, clear differences were observed among different methods in terms of their explanatory power and predictive capability for economic outcomes in the net economic return per unit area regression task. Linear Regression and Ridge Regression achieved MAE values of 0.284 and 0.276, respectively, and R^2 values of 0.681 and 0.697, indicating that traditional linear regression methods can capture part of the average linear relationship between explanatory variables and net return and are suitable for characterizing relatively stable marginal effects. However, their capability remains insufficient for modeling the nonlinear return response, interaction effects among variables, and heterogeneous shock transmission that are commonly observed in real economic processes. The overall results of Random Forest Regressor, XGBoost Regressor, and MLP Regressor were superior to those of linear models, among which XGBoost Regressor achieved an R^2 of 0.768, indicating that more complex conditional relationships in the formation process of net economic return per unit area can be more fully identified when stronger nonlinear fitting capability is provided by the model. Meanwhile, the performance of Sensor-only Regression Network was inferior to that of Image-only Deep Regression Network, suggesting that although a single type of explanatory variable can provide partial economic information, it remains insufficient for fully covering the key dimensions affecting variation in net return, whereas features containing richer state information can significantly enhance the explanatory power for economic outcomes.

Further observation indicates that multimodal methods consistently outperformed unimodal methods, suggesting that net economic return per unit area is not determined by a single factor, but rather is formed through the combined influence of multiple categories of economically relevant information. Image + Sensor Feature Concatenation Regression improved the R^2 value to 0.804, and Multimodal Transformer Regressor further increased it to 0.829, indicating that the fitting capability of the model with respect to the formation mechanism of net return was significantly enhanced when information from different sources was jointly incorporated into a unified modeling framework. The proposed method ultimately achieved the lowest MAE, RMSE, and MAPE, together with the highest R^2, namely 0.176, 0.243, 10.98%, and 0.861, respectively, significantly outperforming all com-

parison models. These results indicate that economically relevant key information can be extracted more effectively by the proposed method, and that a more stable and more accurate return mapping relationship can be established in a unified representation space. Beyond predictive accuracy, this economic analysis bridges the gap between physical agricultural monitoring and practical farm management. By effectively mapping multimodal perceptual features to final net returns, the framework provides a quantitative foundation for proactive intervention. Farm managers can utilize these early-stage economic forecasts to reallocate resources dynamically, shifting focus from uniform plot management to highly targeted, risk-aware operational strategies. From an economic perspective, this implies that the comprehensive response of net return to multidimensional state variables can be more effectively characterized, the identification of differences in operational performance can be improved, and more reliable quantitative support can be provided for return prediction, loss estimation, risk assessment, and resource allocation optimization.

Table 6. Performance comparison of different methods on the agricultural economic regression task. The target variable is net economic return per unit area, measured in thousands of Chinese Yuan per hectare. The best results are achieved by the Proposed Method.

Method	MAE ↓	RMSE ↓	MAPE (%) ↓	R^2 ↑
Linear Regression	0.284	0.362	18.74	0.681
Ridge Regression	0.276	0.351	18.12	0.697
Random Forest Regressor	0.241	0.316	15.67	0.742
XGBoost Regressor	0.228	0.301	14.95	0.768
MLP Regressor	0.235	0.309	15.28	0.756
Sensor-only Regression Network	0.263	0.337	17.46	0.718
Image-only Deep Regression Network	0.221	0.295	14.53	0.779
Image + Sensor Feature Concatenation Regression	0.207	0.281	13.42	0.804
Multimodal Transformer Regressor	0.194	0.266	12.36	0.829
Proposed Method	0.176	0.243	10.98	0.861

4.6. Robustness Analysis Under Environmental Disturbances

In real-world agricultural deployments, multimodal perception systems inevitably face severe environmental disturbances and hardware malfunctions. To systematically evaluate the robustness of the proposed framework, this section conducts simulated perturbation experiments on the test set. Specifically, five typical harsh conditions were designed: (1) Illumination Variation, simulating extreme lighting changes through random brightness and contrast adjustments; (2) Occlusion, simulating canopy overlap and shadows by randomly masking 20 percent to 30 percent of the image areas; (3) Sensor Noise and Drift, introducing Gaussian noise and continuous value drift to the IoT time-series data; (4) Sensor Data Missing, randomly dropping 20 percent of the temporal sequence points to simulate network transmission packet loss; and (5) Spatial Registration Error, adding random positional offsets to the spatial coordinates to simulate GPS inaccuracies. The proposed method is evaluated alongside the best-performing Multimodal Transformer baseline to verify its structural stability under these challenging scenarios.

As shown in Table 7, environmental disturbances and system faults inevitably lead to performance degradation across all models; however, the proposed method demonstrates significantly stronger resilience. Under visual degradation scenarios such as severe illumination variation and canopy occlusion, the performance of the conventional Multimodal Transformer drops sharply, suffering losses of over 6 percent. In contrast, the proposed method only experiences a moderate decline, confirming that the cross-modal collaborative encoding mechanism effectively leverages environmental and spatial priors to compensate for compromised visual cues. Furthermore, when facing IoT sensor faults or spatial misalignments, the proposed dynamic multimodal fusion module plays a critical

role. By adaptively evaluating modality quality and redistributing contribution weights, the network autonomously suppresses the interference from unreliable sensor or positional inputs, thereby preventing the catastrophic failure commonly observed in static fusion networks. These results strongly validate the high reliability and deployment feasibility of the proposed framework in highly unpredictable, real-world smart agriculture applications.

Table 7. Robustness evaluation under different simulated environmental disturbances. Performance is measured by Pest F1-score and Growth Macro-F1.

Disturbance Scenario	Method	Pest F1-Score (%)	Growth Macro-F1 (%)
None (Standard Testing)	Multimodal Transformer	88.79	86.66
	Proposed Method	90.94	88.92
Illumination Variation	Multimodal Transformer	82.15	80.34
	Proposed Method	86.52	84.81
Occlusion	Multimodal Transformer	81.63	79.58
	Proposed Method	85.87	84.15
Sensor Noise and Drift	Multimodal Transformer	85.02	82.47
	Proposed Method	88.16	86.53
Sensor Data Missing	Multimodal Transformer	83.28	81.12
	Proposed Method	87.41	85.94
Spatial Error	Multimodal Transformer	84.55	82.90
	Proposed Method	87.93	86.18

4.7. Discussion

The results of this study indicate that the significance of the proposed method lies not only in improving recognition accuracy, but more importantly in enhancing the interpretability of actual economic outcomes. In real production and management activities, what concerns managers most is often not whether a particular leaf is abnormal, nor whether a single environmental indicator exhibits short-term fluctuation, but rather how such state changes will ultimately affect net economic return per unit area, total output value, and input–output efficiency. Traditional experience-based judgment usually relies on manual inspection and ex post accounting, and problems are often confirmed only after yield reduction, quality deterioration, or declining sales revenue has already occurred. From an economic perspective, such a mode constitutes a typical form of passive response. The proposed method is capable of transforming local state variation, environmental disturbance, and spatial distribution differences into quantitative signals of return variation at an earlier stage, and is therefore more closely aligned with the transmission chain of "state–output–return" emphasized in economic analysis. A critical strength of our approach is the grounding of these economic predictions in verified field data. As detailed in the regression analysis section, the net return per unit area figures were synthesized from primary field management logs, cooperative financial records, and official market databases. By explicitly accounting for dynamic costs such as fertilizer, pesticide, and labor against the weighted average market price of the harvested yield, the model moves beyond theoretical estimation to reflect real-world operational performance. This empirical foundation ensures that the variations captured in the regression results are not merely statistical artifacts but represent the tangible impact of crop health on the profitability of the farming entity.

In specific scenarios, this capability is first reflected in the forward shift of operational decision-making. For example, among multiple plots managed by the same operating entity, even when input levels are broadly similar, substantial differences in output value may still arise across regions due to local damage, growth divergence, or environmental heterogeneity. If accounting can only be conducted on the basis of final harvest outcomes, the manager can merely accept such differences passively after returns have already been realized. By contrast, if regions in which future net return is likely to decline can be identified at an earlier stage during the growth process, labor allocation, irrigation frequency, chemical

input intensity, and inspection effort can be adjusted in advance, thereby reducing marginal losses. For large-scale operating entities, this capability essentially improves the efficiency of resource allocation, in the sense that limited capital, labor, and managerial attention can be prioritized toward regions with the highest return elasticity and the greatest risk exposure, rather than being distributed evenly. Furthermore, this method also has direct value for return assessment and risk management. In real economic activity, net return is not determined by a single variable, but is instead formed through the joint effects of state conditions, environmental shocks, and spatial heterogeneity. The regression analysis conducted in this study demonstrates that the fitting capacity for net economic return per unit area is significantly enhanced after multi-source information is modeled jointly. This implies that differences in operational performance no longer need to rely solely on ex post statistical interpretation, but can instead be continuously characterized and predicted during the process itself. For operating entities, cooperative organizations, and relevant administrative bodies, such capability can be used to identify the scale of potential loss earlier, detect high-risk regions, evaluate the economic consequences of alternative management strategies, and provide a more quantitative basis for input optimization, return forecasting, and performance evaluation. In other words, the core value of the proposed method lies in transforming originally fragmented, dispersed, and economically intractable field information into effective variables that can directly serve output value prediction and net return analysis, thereby improving both the explanatory power for actual economic outcomes and the capability for decision support.

4.8. Limitation and Future Work

Although favorable experimental results were achieved by the proposed method in pest and disease recognition and crop growth assessment tasks, several limitations still remain. First, while the studied fields represent intensive contiguous farming operations, the present study was exclusively conducted within a single geographic region under typical farmland scenarios in Linhe District, Bayannur City. Although the collection period covered multiple growth stages and included complex environmental conditions, the data originates from one specific site, meaning certain regional characteristics still remain in terms of crop categories, planting systems, climate types, and pest and disease spectra. Therefore, validating the generalization capability of the model across geographically diverse regions, varying climate zones, and more crop species remains a critical necessity. Second, although multimodal data integrating UAV imagery, ground sensors, and spatial positional information were employed in this study, time synchronization errors among different modalities, short-term sensor drift, and incomplete local spatial mapping may still exist, and these factors may affect model stability during real deployment. Meanwhile, the current method emphasizes offline modeling and accuracy improvement, whereas real-time inference efficiency, edge-device deployment cost, and long-term continuous operational reliability still require further optimization.

Future work will be further carried out from three aspects, namely model generalization capability, system practicality, and multi-source sensing extension. On the one hand, cross-regional validation will be conducted under geographically diverse regions, more crop species, and more complex farmland conditions in order to enhance the adaptability of the model to different agricultural ecological scenarios. On the other hand, the model structure and inference pipeline will be further optimized to improve lightweight deployment capability on UAV inspection platforms and field edge devices, thereby making the method more suitable for high-frequency inspection requirements in actual agricultural production. In addition, more sensing modalities, such as multispectral data, thermal infrared data, or longer-span continuous environmental sequences, may be introduced to strengthen the

characterization of early latent pest and disease signals and the evolution process of crop growth status, thereby further improving the reliability and application value of intelligent farmland inspection systems.

5. Conclusions

To address the demand for fine-grained inspection of large-scale farmlands in the context of smart agriculture, a multimodal target perception framework for intelligent farmland inspection is proposed in this study. The proposed framework is designed to overcome the limitations of traditional unimodal visual methods, which are highly susceptible to illumination fluctuation, canopy occlusion, scale variation, and background interference in complex field environments, while also failing to fully exploit environmental sensing information and spatial priors. Based on UAV imagery, ground sensor time-series data, and spatial positional information, joint perception of pest and disease recognition and crop growth assessment is achieved through the construction of a cross-modal alignment and collaborative encoding module, a multi-scale target perception module, and a dynamic multimodal fusion and decision module. The main contribution of this study lies in establishing semantic connections among visual phenotypes, environmental drivers, and spatial distributions, thereby enhancing the model's comprehensive representation capability for tiny lesions, regional growth differences, and complex farmland scenarios, while also providing a new perspective for multimodal deep learning research in agricultural applications. The experimental results demonstrate that the proposed method achieves performance significantly superior to existing methods on both core tasks. In the pest and disease recognition task, Precision, Recall, F1-score, and mAP values of 91.63%, 90.27%, 90.94%, and 93.15% were achieved, respectively, clearly outperforming comparison models such as Faster R-CNN, YOLOv8-m, Swin Transformer-Tiny, and Multimodal Transformer. In the crop growth assessment task, Accuracy, Precision, Recall, and Macro-F1 values of 89.96%, 89.11%, 88.74%, and 88.92% were obtained, respectively, again exceeding those of various unimodal visual models and conventional multimodal fusion models. The ablation study further verifies the effectiveness and complementarity of the three core modules, indicating that cross-modal alignment enhances the consistency of heterogeneous information, multi-scale modeling improves joint perception of local and global patterns, and the dynamic fusion mechanism effectively improves the robustness of the model under complex farmland conditions. Overall, the proposed method not only achieves favorable experimental performance, but also demonstrates practical potential for agricultural pest and disease monitoring, crop growth assessment, and precision management, thereby exhibiting strong applicability in agricultural scenarios.

Author Contributions: Conceptualization, J.X., J.P., H.Z. and S.Y.; Methodology, J.X., J.P. and H.Z.; Software, J.X., J.P. and H.Z.; Validation, X.L.; Formal analysis, X.L.; Investigation, X.L.; Resources, B.L. and J.T.; Data curation, B.L. and J.T.; Writing – original draft, J.X., J.P., H.Z., X.L., B.L., J.T. and S.Y.; Visualization, B.L. and J.T.; Supervision, S.Y.; Project administration, S.Y.; Funding acquisition, S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (grant number: 32372631) and the 2115 Talent Development Program of China Agricultural University.

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

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

References

1. Xiong, P.; Zhang, C.; He, L.; Zhan, X.; Han, Y. Deep learning-based rice pest detection research. *PLoS ONE* **2024**, *19*, e0313387. [CrossRef] [PubMed]
2. Zhu, H.; Lin, C.; Liu, G.; Wang, D.; Qin, S.; Li, A.; Xu, J.L.; He, Y. Intelligent agriculture: Deep learning in UAV-based remote sensing imagery for crop diseases and pests detection. *Front. Plant Sci.* **2024**, *15*, 1435016. [CrossRef] [PubMed]
3. Zhou, C.; Cao, Y.; Ming, B.; Luo, J.; Xu, F.; Zhang, J.; Dong, M. A Multimodal Deep Learning Framework for Intelligent Pest and Disease Monitoring in Smart Horticultural Production Systems. *Horticulturae* **2025**, *12*, 8. [CrossRef]
4. Huang, Y.; Wu, X.; Liu, Z.; Wang, Q.; Jin, S.; Xie, C.; Hao, G. Multimodal weed infestation rate prediction framework for efficient farmland management. *Comput. Electron. Agric.* **2025**, *235*, 110294. [CrossRef]
5. Liu, Z.; Li, S.; Yang, Y.; Jiang, X.; Wang, M.; Chen, D.; Jiang, T.; Dong, M. High-Precision Pest Management Based on Multimodal Fusion and Attention-Guided Lightweight Networks. *Insects* **2025**, *16*, 850. [PubMed]
6. Zhang, J.; Hu, J.; Mai, Z.; Chen, Y.; Lou, S.; Huang, H.; Zhao, L.; Huang, J.; Lu, Y.; Fu, H.; et al. AgroNVILA: Perception-Reasoning Decoupling for Multi-view Agricultural Multimodal Large Language Models. *arXiv* **2026**, arXiv:2603.14342.
7. Wuepper, D.; Bukchin-Peles, S.; Just, D.; Zilberman, D. Behavioral agricultural economics. *Appl. Econ. Perspect. Policy* **2023**, *45*, 2094–2105. [CrossRef]
8. Wesseler, J. The EU's farm-to-fork strategy: An assessment from the perspective of agricultural economics. *Appl. Econ. Perspect. Policy* **2022**, *44*, 1826–1843. [CrossRef]
9. Finger, R. Digital innovations for sustainable and resilient agricultural systems. *Eur. Rev. Agric. Econ.* **2023**, *50*, 1277–1309. [CrossRef]
10. Yao, H.; Zhang, Y. Design of intelligent agricultural landscape planning and ecological balance control system based on remote sensing images. *Int. Food Agribus. Manag. Rev.* **2024**, *28*, 667–686. [CrossRef]
11. 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]
12. Guo, Q.; Han, B.; Chu, P.; Wan, Y.; Zhang, J. MF-FusionNet: A Lightweight Multimodal Network for Monitoring Drought Stress in Winter Wheat Based on Remote Sensing Imagery. *Agriculture* **2025**, *15*, 1639. [CrossRef]
13. Zhao, Z.; Xu, S.; Wang, W.; Liang, X.; Lu, H.; Wu, P. Multimodal Fusion Perception and Posture Regulation Integration for Intelligent Detection of Litchi Fruit Borer. *LWT* **2026**, *241*, 119091. [CrossRef]
14. Radočaj, D.; Radočaj, P.; Plaščak, I.; Jurišić, M. Evolution of Deep Learning Approaches in UAV-Based Crop Leaf Disease Detection: A Web of Science Review. *Appl. Sci.* **2025**, *15*, 10778. [CrossRef]
15. Rana, S.; Hensel, O.; Nasirahmadi, A. From Vineyard to Vision: Multi-Domain Analysis and Mitigation of Grape Cluster Detection Failures in Complex Viticultural Environments. *Results Eng.* **2025**, *29*, 108833. [CrossRef]
16. Ahmed, W.A.; Abiola, O.A.; Yang, D.; Olatoyinbo, S.F.; Jing, G. Integrating UAVs and Deep Learning for Plant Disease Detection: A Review of Techniques, Datasets, and Field Challenges with Examples from Cassava. *Horticulturae* **2026**, *12*, 87. [CrossRef]
17. Song, W.; Li, X.; Zhang, J.; Huang, J.; Wang, J.; Feng, W.; Guo, Y.; Ren, L. Soil Sensors in Smart Agriculture: Multi-Type Monitoring Technologies and Ecological Development Pathways. *Agriculture* **2026**, *16*, 359. [CrossRef]
18. Mohammed Aarif, K.O.; Alam, A.; Hotak, Y. Smart sensor technologies shaping the future of precision agriculture: Recent advances and future outlooks. *J. Sens.* **2025**, *2025*, 2460098. [CrossRef]
19. Li, J.; Hong, D.; Gao, L.; Yao, J.; Zheng, K.; Zhang, B.; Chanussot, J. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102926. [CrossRef]
20. Wang, X.; Yan, F.; Li, B.; Yu, B.; Zhou, X.; Tang, X.; Jia, T.; Lv, C. A multimodal data fusion and embedding attention mechanism-based method for eggplant disease detection. *Plants* **2025**, *14*, 786. [CrossRef] [PubMed]
21. Padhiary, M.; Kumar, K.; Hussain, N.; Roy, D.; Barbhuiya, J.A.; Roy, P. Artificial intelligence in farm management: Integrating smart systems for optimal agricultural practices. *Int. J. Smart Agric.* **2025**, *3*, 1–11. [CrossRef]
22. Guo, K.; Li, Y.; Zhang, Y.; Ji, Q.; Zhao, W. How are climate risk shocks connected to agricultural markets? *J. Commod. Mark.* **2023**, *32*, 100367. [CrossRef]
23. Möhring, N.; Huber, R.; Finger, R. Combining ex-ante and ex-post assessments to support the sustainable transformation of agriculture: The case of Swiss pesticide-free wheat production. *Q. Open* **2023**, *3*, qoac022.
24. Ingrao, C.; Strippoli, R.; Lagioia, G.; Huisingh, D. Water scarcity in agriculture: An overview of causes, impacts and approaches for reducing the risks. *Heliyon* **2023**, *9*, e18507. [CrossRef] [PubMed]
25. Ashtiani, S.H.M.; Manafiazar, G.; Do, D.N.; Hu, G.; Davoudi, P.; Miar, Y. Fuzzy-based machine learning for the prediction of feed efficiency and disease status in livestock: A mink case study. *Smart Agric. Technol.* **2026**, *14*, 102104. [CrossRef]
26. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. pp. 1233–1246. [CrossRef]
27. Gökalp, Ç.; Mazhar, Ç.M. An improved pistachio detection approach using YOLO-v8 Deep Learning Models. *Proc. BIO Web Conf. EDP Sci.* **2024**, *85*, 01013. [CrossRef]

28. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
29. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Virtual, 23–28 August 2020; pp. 213–229.
30. 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, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
31. 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 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
32. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. pp. 381–393.