

Review

Deep Learning in Multimodal Fusion for Sustainable Plant Care: A Comprehensive Review

Zhi-Xiang Yang ^{1,†}, Yusi Li ^{1,†}, Rui-Feng Wang ^{1,†}, Pingfan Hu ^{2,*} and Wen-Hao Su ^{1,*}

¹ China Agricultural University, Qinghua East Road No. 17, Haidian, Beijing 100083, China; yangzhixiang@cau.edu.cn (Z.-X.Y.); liyusi040723@163.com (Y.L.); sweefongreggie Wong@cau.edu.cn (R.-F.W.)

² Artie McFerrin Department of Chemical Engineering, Texas A&M University, College Station, TX 77843-3122, USA

* Correspondence: pingfanhu@tamu.edu (P.H.); wenhao.su@cau.edu.cn (W.-H.S.)

[†] These authors contributed equally to this work.

Abstract: With the advancement of Agriculture 4.0 and the ongoing transition toward sustainable and intelligent agricultural systems, deep learning-based multimodal fusion technologies have emerged as a driving force for crop monitoring, plant management, and resource conservation. This article systematically reviews research progress from three perspectives: technical frameworks, application scenarios, and sustainability-driven challenges. At the technical framework level, it outlines an integrated system encompassing data acquisition, feature fusion, and decision optimization, thereby covering the full pipeline of perception, analysis, and decision making essential for sustainable practices. Regarding application scenarios, it focuses on three major tasks—disease diagnosis, maturity and yield prediction, and weed identification—evaluating how deep learning-driven multisource data integration enhances precision and efficiency in sustainable farming operations. It further discusses the efficient translation of detection outcomes into eco-friendly field practices through agricultural navigation systems, harvesting and plant protection robots, and intelligent resource management strategies based on feedback-driven monitoring. In addressing challenges and future directions, the article highlights key bottlenecks such as data heterogeneity, real-time processing limitations, and insufficient model generalization, and proposes potential solutions including cross-modal generative models and federated learning to support more resilient, sustainable agricultural systems. This work offers a comprehensive three-dimensional analysis across technology, application, and sustainability challenges, providing theoretical insights and practical guidance for the intelligent and sustainable transformation of modern agriculture through multimodal fusion.

Keywords: multimodal fusion; crop monitoring and management; agricultural machinery; deep learning; sustainable agriculture

Academic Editor: Spyros Fountas

Received: 28 April 2025

Revised: 24 May 2025

Accepted: 2 June 2025

Published: 6 June 2025

Citation: 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. <https://doi.org/10.3390/su17125255>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Driven by global population growth and intensified climate change, agricultural production faces dual pressures of ensuring food security and achieving resource sustainability. According to the Food and Agriculture Organization of the United Nations (FAO), global food demand is projected to increase by 60% by 2050, while traditional agriculture, which heavily relies on manual experience and extensive management, is increasingly

inadequate to meet these challenges [1]. In this context, Agriculture 4.0, characterized by the deep integration of the Internet of Things, artificial intelligence, and robotics, is accelerating the transition toward digital and intelligent farming systems [2–4]. However, the complexity and variability of field environments impose significant challenges on intelligent perception and decision-making technologies. The limitations of unimodal approaches in dynamic agricultural scenarios have become increasingly apparent, prompting academic and industrial communities to explore intelligent detection technologies driven by multisource data [5].

Currently, unimodal detection technologies, such as crop classification based on RGB imagery or environmental monitoring using single temperature or humidity sensors, suffer from limited data dimensionality and are highly susceptible to environmental disturbances [6]. For instance, in orchards with dramatic lighting variations, models relying solely on RGB cameras exhibit high misclassification rates [7]; similarly, single soil moisture sensors cannot distinguish between drought-induced and salinity-induced crop wilting, leading to biased irrigation decisions. These shortcomings stem from the inherent incompleteness of unimodal data, where individual sensors can only capture partial features of crops or environments and thus struggle with the multiscale and multidimensional complexity of agricultural fields. Furthermore, data asynchrony and heterogeneity in traditional sensor networks exacerbate the risk of errors in agricultural navigation and pest or disease diagnosis.

To address these challenges, multimodal fusion technologies have been developed to integrate heterogeneous data sources, including RGB images, LiDAR point clouds, hyperspectral imagery, and environmental sensor data, enabling complementary perception and substantially enhancing the robustness and comprehensiveness of detection systems under complex field conditions.

Despite the significant potential of multimodal fusion in agricultural intelligence, three major challenges hinder its practical deployment: (i) The spatiotemporal asynchrony and modal heterogeneity of field data complicate feature alignment and fusion processes, severely limiting model performance; (ii) The computational demands of complex fusion algorithms conflict with the limited edge computing capabilities of agricultural equipment; (iii) Insufficient model generalization across different crops and environmental conditions restricts the scalability of multimodal technologies.

Existing studies predominantly focus on isolated technical aspects, such as sensor design or fusion algorithm optimization, and lack systematic investigations into the coordinated optimization of data acquisition, model construction, and hardware deployment [8–10]. Moreover, research efforts that extend multimodal fusion applications across the entire agricultural management chain remain scarce [11].

The main contributions of this review are as follows:

- (i) A comprehensive deep learning-based multimodal fusion framework is constructed, encompassing data acquisition, feature fusion, and decision optimization.
- (ii) The application of multimodal deep learning in three core areas—crop status detection, intelligent agricultural equipment operations, and resource management—is systematically reviewed.
- (iii) Key challenges, including data heterogeneity, real-time processing bottlenecks, and limited model generalization, are analyzed in depth, and future research directions, such as federated learning, self-supervised pretraining, and dynamic computation frameworks, are proposed.

The remainder of this article is structured as follows: Section 2 presents the basic framework of multimodal fusion technologies in agriculture, covering the full chain from data acquisition to decision optimization; Section 3 reviews applications in crop detection and plant care, including deep learning-based crop status assessment, multimodal

intelligent equipment deployment, and detection-driven resource management and ecological monitoring; Section 4 summarizes current challenges regarding data heterogeneity, real-time constraints, and model generalization, and discusses future development directions; Section 5 concludes the study and offers an outlook. The overall structure of the article is illustrated in Figure 1.

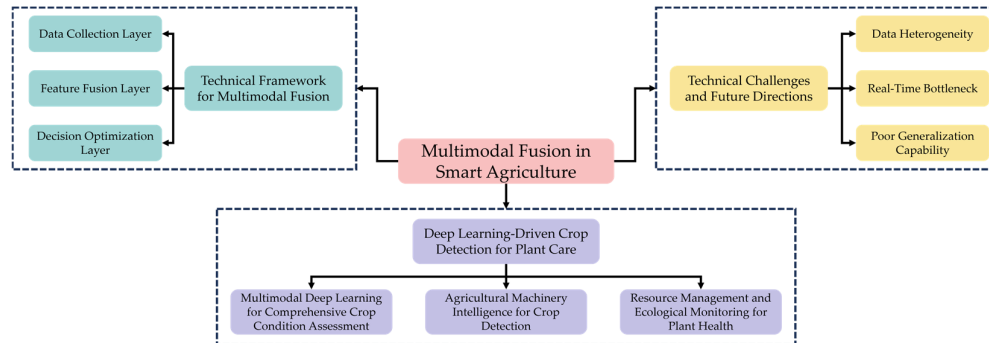


Figure 1. Multimodal fusion in smart agriculture: from sensing to decision making. Note: The phrase “Multimodal Fusion in Smart Agriculture” in the red box represents the overall theme and central concept of the article, emphasizing the macro-level application and significance of multimodal fusion technologies in the context of smart agriculture. In contrast, the phrase “Technical Framework for Multimodal Fusion” in the cyan box specifically refers to the technical implementation pathway detailed in Section 2, which includes three core layers: data collection, feature fusion, and decision optimization.

2. Technical Framework for Multimodal Fusion

The core of multimodal fusion technologies lies in integrating multisource data from different sensors to construct a complete pipeline from perception to decision making. This framework is structured into three layers: data acquisition, feature fusion, and decision optimization. The overall architecture of multimodal fusion is shown in Figure 2.

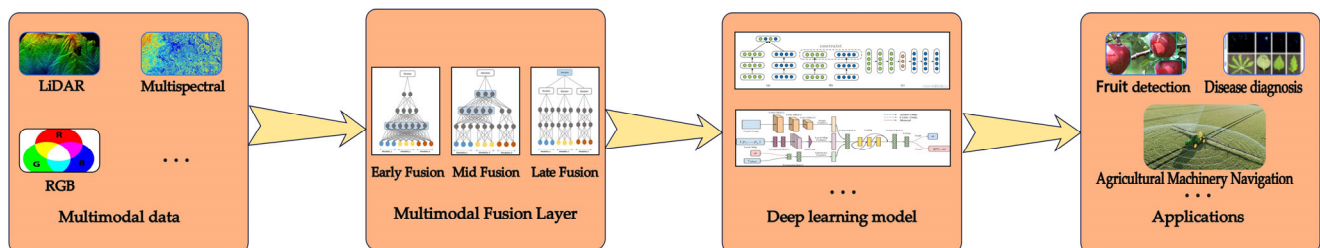


Figure 2. Technical framework for multimodal fusion.

2.1. Data Collection Layer

The foundation of multimodal fusion relies on efficient and coordinated data acquisition technologies. This section addresses two key aspects: an analysis of the technical characteristics of different types of sensors, followed by a discussion on methods for spatiotemporal synchronization and feature mapping of heterogeneous multisource data, aiming to provide standardized inputs for subsequent feature fusion. Effective data acquisition not only ensures the accuracy and reliability of agricultural monitoring but also contributes to sustainable resource utilization by minimizing redundant data collection and reducing energy consumption.

2.1.1. Sensor Types

Agricultural multimodal data acquisition relies on a diverse range of sensor technologies, enabling comprehensive capture of multidimensional information on field environments and crop growth through coordinated sensing across aerial, ground, and subsurface platforms. These sensors provide a rich data foundation essential for the implementation of precision agriculture.

Aerial platforms, primarily unmanned aerial vehicles (UAVs), are equipped with hyperspectral or multispectral cameras and LiDAR systems to achieve large-scale, efficient monitoring. Ground platforms, comprising mobile robots and stationary devices, utilize thermal imagers and RGB cameras for detailed field inspection [12]. Subsurface and near-surface sensors focus on monitoring root zone microenvironments and meteorological dynamics, deploying multiparameter soil sensors to measure variables such as temperature, humidity, electrical conductivity, and pH levels.

The complementary advantages of these sensors form an integrated aerial–ground–subsurface perception network, establishing a robust data basis for multimodal fusion technologies. Table 1 summarizes the advantages, limitations, and application scenarios of the aforementioned sensors, while Figure 3 presents examples of sensor appearances.

Table 1. Comparison and analysis of different sensor types in precision agriculture.

Sensor Type	Advantages	Limitations	Applications	References
Hyperspectral Camera	Accurately identifies crop physiological states and minor biochemical changes	High data volume and high cost, limiting large-scale application	High data volume and high cost, limiting large-scale application	[13–15]
Multispectral Camera	Low cost, portable, suitable for large-area monitoring	Limited data dimensionality, difficulty in capturing subtle changes	Crop classification, large-area field monitoring, detection of group physiological stress	[16–19]
LiDAR	Provides high-precision 3D spatial information, suitable for complex terrain measurement	High equipment cost, complex data processing	Crop height measurement, terrain modeling, 3D digital twin construction	[20]
Thermal Imaging Camera	Identifies irrigation unevenness and early-stage disease regions	Sensitive to environmental temperature changes, may be affected by weather	Irrigation optimization, early disease detection, identification of thermal anomalies	[21]
RGB Camera	Low cost, real-time imaging, high resolution, foundational tool for agricultural monitoring	Only captures visible light information, difficulty in identifying non-visible spectral features	Crop classification, disease detection, basic agricultural monitoring	[22,23]
Soil Multiparameter Sensors	Provides root microenvironment data, supports precision irrigation and fertilization decisions	Limited sensor depth, may not fully reflect soil profile information	Precision irrigation, fertilizer optimization, soil health management	[24,25]

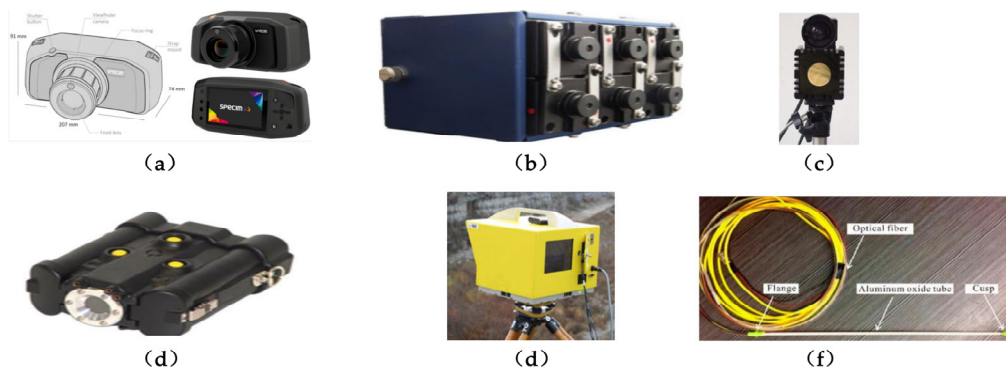


Figure 3. Sensors used in agricultural multimodal data collection: (a) hyperspectral camera [26]; (b) multispectral camera [27]; (c) RGB camera [28]; (d) thermal imaging camera [29]; (e) LiDAR [30]; (f) soil moisture sensor [31].

2.1.2. Data Alignment

The efficient fusion of multisource data is fundamentally challenged by spatiotemporal asynchrony and modality heterogeneity. Differences in platform characteristics and sampling mechanisms among aerial, ground, and subsurface sensors often lead to misalignment in timestamps, spatial coordinates, and semantic features. To address this, a unified framework for spatiotemporal referencing and feature representation must be established to provide a reliable foundation for multimodal fusion in agricultural scenarios. Spatiotemporal synchronization consists of two key processes: timestamp alignment and spatial registration. Timestamp alignment employs high-precision clock synchronization protocols to coordinate the sampling rates of UAVs, ground robots, and soil sensors, combined with interpolation algorithms such as linear interpolation and Kalman filtering to generate temporally consistent data streams. For instance, the USTC FLICAR dataset achieves timestamp deviations within ± 5 ms between UAV-mounted LiDAR and multispectral cameras through GPS-based timing and hardware-trigger mechanisms [32]. Spatial registration utilizes SLAM (Simultaneous Localization and Mapping) or RTK-GPS (Real-Time Kinematic Global Positioning System) to map multisource data into a unified geographic coordinate system. For example, in scenarios involving object-based differentiation and biophysical characterization of three vegetable crops (cabbage, eggplant, and tomato), Nidamanuri et al. [33,34] employed a manually guided spherical fitting algorithm to establish correspondences between multiple scans during the registration of LiDAR point clouds and multispectral images, achieving a recognition accuracy of 92%, approximately 20% higher than that obtained with single-sensor approaches.

Recent advances in deep learning-based registration methods have introduced new solutions for data alignment. In the field of point cloud registration, networks such as PointNet and DCP (Deep Closest Point) automatically learn feature correspondences, effectively overcoming the sensitivity of traditional ICP algorithms to initial pose estimation. Table 2 provides the open-source implementations of these representative networks. These developments provide critical support for precise detection and intelligent decision making in agricultural environments.

Table 2. Open-source resources of representative point cloud registration networks.

Networks	URL
PointNet	https://github.com/charlesq34/pointnet (accessed on 10 May 2025)
DCP	https://github.com/WangYueFt/dcp (accessed on 10 May 2025)

2.2. Feature Fusion Layer

The feature fusion layer serves as the critical nexus between data acquisition and intelligent decision making in multimodal technologies. Its objective is to extract more discriminative fused representations through deep interaction and coordinated optimization of cross-modal features, thereby enhancing precise perception and decision making in agricultural scenarios. This section elaborates on two key dimensions: fusion strategies and deep learning models, which together establish a closed-loop framework from strategy-driven design to model-enabled implementation. By effectively fusing features from multiple modalities, this layer plays a vital role in promoting sustainable agricultural practices by enabling more accurate and timely decision making, which can lead to optimized resource allocation and reduced environmental impact.

2.2.1. Fusion Strategies

Multimodal data fusion strategies can be classified into three categories—early fusion, mid-level fusion, and late fusion—based on the integration stage and feature interaction mechanisms. Each strategy exhibits distinct differences in computational complexity, information retention, and application suitability, requiring careful selection according to agricultural task demands and data characteristics.

Early fusion operates at the raw data level by directly aligning and concatenating multimodal inputs to form high-dimensional composite data. Typical methods include data concatenation (stacking temporally and spatially aligned multimodal data) and encoder-based mapping (projecting heterogeneous data into a unified low-dimensional space using convolutional neural networks or autoencoders). For instance, Couprie et al. [35] proposed the first deep learning-based multimodal fusion model by integrating RGB and depth images [36]. Zamani et al. [37] developed an early fusion approach by matching and aligning visible and thermal imagery for weed and rice classification in fields. In agricultural scenarios, early fusion preserves comprehensive raw information and is well-suited for strongly correlated tasks such as crop classification and disease detection. However, its high computational demands limit its feasibility for deployment on edge devices.

Mid-level fusion extracts high-level features from each modality independently and subsequently integrates them through interaction mechanisms. Typical methods include feature concatenation, attention mechanisms (e.g., cross-modal attention or Transformers), and graph neural networks (modeling inter-modal feature relationships through message passing). This approach efficiently exploits multi-level features extracted from deep networks, enhancing model performance. For example, Hu et al. [38] proposed AC-Net, an attention-based feature fusion network for extracting features from RGB and depth images. Similarly, Shang et al. [39] introduced a novel information fusion network that improved segmentation performance by jointly processing RGB and depth features. In agricultural applications, mid-level fusion demonstrates superior capability in handling asynchronous data (e.g., UAV multispectral imagery and ground sensor data) and reducing redundant information, thus improving computational efficiency [40]. However, it relies on high-quality feature extractors and may lose low-level cross-modal associations.

Late fusion aggregates independently analyzed results from different modalities using methods such as weighted voting or Bayesian networks. For instance, Sun et al. [41] proposed FuseSeg, a fusion network that combines RGB and thermal data to improve semantic segmentation in urban scenes. He et al. [42] developed a late fusion strategy based on an enhanced voting method for rice seed variety classification by combining predictions from 2D images and 3D point cloud data, achieving an average accuracy of 97.4%. In agricultural contexts, late fusion enhances model robustness and reliability by integrating the complementary advantages of different modalities, making it particularly suitable for edge computing scenarios with constrained resources. Nonetheless, the loss of cross-modal relational information and dependence on balanced unimodal performance impose a ceiling on the achievable accuracy. A comparative analysis of these three fusion strategies is summarized in Table 3.

Table 3. Comparison of characteristics of early, mid-, and late fusion in multimodal systems.

Characteristic	Fusion Stage	Modal Interaction Depth	Computational Complexity	Robustness	Flexibility	Information Integrity	Applicable Scenarios	References
Early Fusion	Data Input Stage	High	High	Low	Low	Highest	Tasks requiring deep interaction	[43]

Mid Fusion	Feature Extraction Stage	Medium	Medium	High	High	Medium	Tasks combining images and text	[44,45]
Late Fusion	Decision Stage	Low	Low	Highest	Highest	Low	Tasks with strong modality independence	[46–48]

2.2.2. Deep Learning Models

As the cornerstone of multimodal fusion systems, deep learning models facilitate the efficient interpretation and collaborative understanding of heterogeneous agricultural data through cross-modal association strategies such as self-attention mechanisms, dynamic weight allocation, feature space mapping, and adaptive interaction architectures [49,50]. This section highlights four representative model architectures and discusses their applications within agricultural contexts.

(A) Convolutional Neural Networks (CNNs): As foundational models in computer vision, CNNs employ a hierarchical architecture consisting of convolutional layers for local feature extraction, pooling layers for spatial downsampling, and fully connected layers for global feature integration, forming a closed-loop learning mechanism [51]. In agricultural detection tasks, CNNs have been widely applied to problems such as crop disease identification and weed detection [52,53]. For example, Coulibaly et al. [54] proposed a method based on the VGG16 model for recognizing pearl millet downy mildew, achieving an accuracy of 95% and a recall rate of 94.5%. Despite their effectiveness, CNNs face limitations when applied to multimodal data. Due to their reliance on local feature extraction, CNNs often struggle to capture global contextual information, which can constrain performance in multimodal fusion scenarios. Specifically, when integrating heterogeneous data sources such as meteorological data and visual imagery, CNNs may fail to effectively combine features across modalities, reducing model generalization capacity [55].

(B) Generative Adversarial Networks (GANs): GANs consist of two components: a generator that captures the distribution of real data and a discriminator that functions as a binary classifier [56]. Their core principle lies in adversarial training, where the generator and discriminator are iteratively optimized to produce high-quality synthetic data [57]. In agricultural detection, GANs are primarily employed as data augmentation tools to enrich training datasets. Additionally, GANs have been utilized for anomaly detection; by learning the distribution of normal crop features, the generator produces images resembling healthy crops, while the discriminator distinguishes between real and synthetic images to identify abnormal instances. For example, Lu et al. [58] applied a GAN-based augmentation approach in pest detection, achieving an F1 score of 0.95, outperforming conventional methods that yielded an F1 score of 0.92. However, the application of GANs in multimodal fusion remains challenging, as generated samples often suffer from artifacts such as blurriness or unrealistic textures, which are particularly problematic in agricultural scenarios requiring fine-grained visual details [59].

(C) Transformer Models: The Transformer architecture, based on a self-attention mechanism, comprises an encoder and a decoder, each formed by stacking multiple identical layers [60]. Its principal advantage lies in the ability to capture long-range dependencies between different regions within an image [61]. Originally designed for natural language processing, the Transformer has been widely adopted for computer vision tasks. In agricultural detection, Transformer models demonstrate significant potential in crop image classification, object detection, and semantic segmentation, supporting applications such as plant disease diagnosis, yield estimation, and quality assessment [62,63]. For instance, Xu et al. [64] and Xiong et al. [65] successfully applied Transformer-based models for crop yield prediction. Nevertheless, Transformers are inherently less effective than

CNNs at extracting fine-grained local features, which may limit their performance in multimodal tasks requiring detailed spatial information. Furthermore, the training of Transformer models is computationally intensive, demanding large-scale datasets and high hardware resources, while also exhibiting slow convergence, thus constraining their applicability in resource-limited agricultural environments.

(D) Recurrent Neural Networks (RNNs): RNNs, characterized by temporal modeling capabilities and parameter-sharing architectures, provide a fundamental framework for processing sequential data [66]. In agricultural detection, RNNs are particularly suitable for handling time-series data such as crop growth patterns and meteorological measurements. For instance, Chai et al. [67] employed a time-series modeling approach based on NARXNN, a variant of RNN, to estimate the Leaf Area Index (LAI), achieving greater continuity and stability compared to traditional MODIS LAI products. In agricultural applications, RNNs can be integrated with models such as CNNs to enable multimodal data fusion, thereby enhancing adaptability to complex agricultural environments. However, RNNs also face notable limitations [68]. Specifically, when processing long sequences, RNNs are prone to gradient vanishing or explosion, which undermines their ability to capture long-term dependencies critical in multimodal tasks. Additionally, RNNs suffer from low computational efficiency and limited parallelization, resulting in slower training and inference speeds that restrict their use in real-time agricultural detection scenarios.

In addition to the four mainstream models discussed above, several emerging deep learning architectures have shown promising potential in multimodal fusion tasks. For instance, BERT-based vision–language models such as ViLBERT and VisualBERT leverage dual-stream Transformer architectures to achieve effective integration of visual and textual agricultural information [69,70]. LSTM, an improved variant of RNN, offers enhanced stability when handling long agricultural time-series data, such as modeling the relationship between weather variations and crop growth [71,72]. Graph Neural Networks (GNNs) are well-suited for capturing complex spatial relationships in agricultural settings, such as interactions between distributed soil sensors [73,74]. Meanwhile, Diffusion Models, known for their breakthroughs in high-fidelity image generation, are increasingly being explored for agricultural image enhancement and synthesis [75,76]. Although these models are still in the early stages of agricultural application, their unique structural advantages present new opportunities for advancing multimodal fusion research.

The application of these deep learning models in multimodal fusion provides powerful tools for agricultural detection, significantly enhancing both accuracy and efficiency, and accelerating the advancement of intelligent agricultural management. A comparative overview of these four models, in terms of training efficiency, model complexity, and robustness, is presented in Table 4.

Table 4. Comparison of training efficiency, model complexity, and robustness across different models.

Model	Training Efficiency	Model Complexity	Robustness	References
CNNs	High training efficiency, parallelizable processing	Relatively simple structure, fewer parameters, not suitable for sequential data	Moderate robustness to data noise and deformation	[77,78]
GANs	Low training efficiency, complex and unstable training process	Complex structure, many parameters	Good robustness for generation tasks, but unstable training process	[79,80]
Transformer	High training efficiency, parallelizable processing	Relatively complex structure, many parameters	Strong robustness to data noise and deformation	[81,82]
RNNs	Relatively low training efficiency, difficult to parallelize	Relatively complex structure, many parameters	Moderate robustness to data noise and deformation	[83,84]

2.3. Decision Optimization Layer

As the closed-loop decision hub of multimodal fusion technologies, the decision optimization layer transforms high-dimensional semantic information within cross-modal feature spaces into executable agricultural operation instructions through dynamic knowledge mapping and constraint satisfaction mechanisms. Current mainstream decision optimization methods include dynamic knowledge graph-driven reasoning, reinforcement learning-based adaptive strategies, multi-objective optimization modeling, as well as emerging strategies for addressing data imbalance in multimodal systems. Together, these models and mechanisms are essential for sustainable agriculture, as they support precise and efficient monitoring of crop and environmental conditions, optimize resource utilization, and reduce environmental impact.

(A) **Dynamic Knowledge Graph-Driven Reasoning:** Dynamic knowledge graph-driven reasoning constructs agricultural knowledge graphs (e.g., crop growth stages, disease transmission patterns, soil-climate associations) to map multimodal data into semantic nodes and relations, enabling reasoning and decision making through rule engines or graph neural networks (GNNs) [85–88]. For example, Chen et al. [89] built a dynamic knowledge graph based on ontology modeling and a Neo4j graph database, applying it to agricultural informatization, with experimental results demonstrating effectiveness and expected outcomes. The advantage of dynamic knowledge graphs lies in their support for semantic interpretability and transparent decision-making processes; however, challenges include reliance on domain expert knowledge for graph construction and the high cost of updates.

(B) **Reinforcement Learning-Based Adaptive Strategies:** Reinforcement learning (RL) adaptive strategies, based on Markov Decision Processes, learn optimal policies through the interaction among environment states (e.g., crop growth status), actions (e.g., irrigation amounts), and reward functions (e.g., yield maximization) [90,91]. For instance, Sun et al. [92] applied RL for water-saving irrigation in fields based on soil water levels and crop yield, while Chen et al. [93] utilized RL for rice irrigation by integrating short-term weather forecasts and observed meteorological data. The strength of RL lies in its ability to adapt to dynamic environmental changes (e.g., sudden rainfall) and achieve real-time decision optimization, whereas challenges include the complex design of reward functions and the need for extensive training data [94].

(C) **Multi-Objective Optimization Modeling:** In agricultural decision-making scenarios, multi-objective optimization modeling incorporates production benefits (e.g., yield maximization), environmental constraints (e.g., carbon emission control), and resource efficiency (e.g., water use minimization) into a unified framework, formulating a multi-objective optimization problem. Intelligent optimization algorithms, such as genetic algorithms and particle swarm optimization, are then employed to solve for the non-dominated solution sets, offering decision-makers balanced solutions that address multiple objectives [95–98]. For example, Habibi Davijani et al. [99] developed a multi-objective optimization model that, based on production functions, cultivated area, product yield, and revenue per product, determined a comprehensive objective function for water resource allocation, achieving a 54% increase in the economic-to-profit ratio compared to the baseline. The main advantage of this approach is its ability to perform global optimization under complex constraints; however, the computational complexity grows exponentially with the dimensionality of the objective space [100].

(D) **Data Imbalance Handling in Multimodal Fusion:** In the process of multimodal fusion decision making, data imbalance remains a critical challenge that must be addressed [101,102]. In practical agricultural applications, different types of sensors often operate at varying sampling frequencies, resulting in significant disparities in data volume and dimensionality [103]. For instance, meteorological sensors can generate

hundreds of data points per second, while remote sensing imagery or manual observations are updated far less frequently. This discrepancy can lead to models favoring high-frequency data during training, thereby compromising the fairness and accuracy of the fusion results. To tackle this issue, several strategies have been proposed in recent research. First, temporal alignment and resampling techniques are employed to downsample high-frequency data or interpolate low-frequency streams, achieving synchronization across modalities [104]. Second, attention mechanisms are widely integrated into fusion models to dynamically assign weights based on the contextual relevance of each modality, rather than data frequency alone [105]. Third, adversarial training approaches are used to balance the contributions of different data sources during model training [106]. Finally, data augmentation is used to increase the number of low-frequency modality samples, enhancing the model's ability to capture important features from low-frequency data [107,108]. These strategies are crucial for ensuring the stability and generalization ability of multimodal systems, and are an essential part of achieving optimized decision making in agricultural intelligence.

3. Deep Learning-Driven Crop Detection for Plant Care

3.1. Multimodal Deep Learning for Comprehensive Crop Condition Assessment

Crop detection is one of the core tasks in precision agriculture, encompassing key processes such as disease diagnosis, maturity assessment, and yield prediction. Traditional unimodal detection methods, due to their limited data dimensionality and poor environmental robustness, struggle to meet the dynamic changes and diverse demands present in complex farmland scenarios. Multimodal fusion technologies, by integrating spectral, thermal, morphological, and environmental parameters from multiple sources, significantly enhance detection accuracy and robustness, offering a new paradigm for agricultural intelligence. This section systematically examines the technical approaches and application value of multimodal fusion in crop detection, focusing on three dimensions: disease diagnosis, maturity and yield prediction, and weed identification.

3.1.1. Disease Diagnosis

Early and accurate diagnosis of crop diseases is critical for ensuring agricultural productivity and sustaining food security. Traditional models relying on meteorological data (e.g., rainfall) often struggle to differentiate between disease symptoms and environmental stress due to their phenotypic similarities. Multimodal fusion technologies, by integrating spectral, thermal, textural, and environmental parameters, substantially enhance both the accuracy and robustness of disease detection [109,110]. A representative architecture illustrating the integration of RGB, thermal, and spectral data for crop disease detection is shown in Figure 4 [111]. This example demonstrates how heterogeneous inputs can be effectively fused to improve diagnostic accuracy and robustness in complex agricultural environments. For instance, Zhao et al. [112] proposed a Multi-Context Fusion Network (MCFN) that combines visual features (field-acquired images) with contextual information (season, geographical location, temperature, and humidity) for crop disease prediction, achieving a recognition accuracy of 97.5%. Similarly, Selvaraj et al. [113] developed a pixel-based classification approach using multisource satellite imagery (Sentinel-2, PlanetScope, WorldView-2) and UAV data (MicaSense RedEdge) for banana disease detection in complex African landscapes, combining RetinaNet and a custom classifier to achieve a detection accuracy of 92%. By enabling precise and timely disease detection, these technologies support sustainable farming by reducing the need for chemical interventions and minimizing crop losses, which is essential for maintaining ecological balance and ensuring food security.

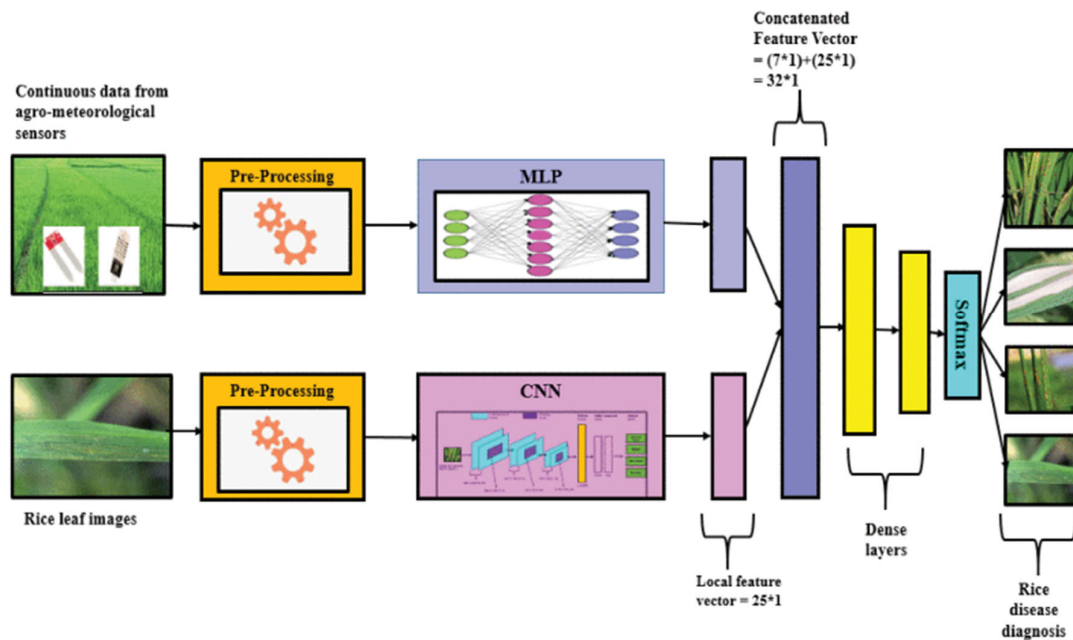


Figure 4. Multimodal rice-fusion architecture [111]. In the figure, “*” refers to multiplication.

3.1.2. Maturity Detection and Yield Prediction

Maturity assessment and yield prediction are pivotal components of precision agriculture, directly influencing harvest timing and market decision making [114,115]. Traditional models based on historical yield statistics lack the responsiveness to sudden droughts or pest outbreaks. Multimodal fusion, by integrating remote sensing imagery, environmental sensor data, and crop phenotypic traits, significantly improves prediction accuracy and timeliness [116,117]. Maimaitijiang et al. [118] utilized UAV-mounted multisensors to extract canopy information, including spectral indices (VIs), structural features (vegetation fraction, canopy height), thermal traits (normalized relative canopy temperature), and texture information, for soybean grain yield prediction. In a related study [119], they further combined RGB, multispectral, and thermal images to extract vegetation traits and employed an ElasticNet regression model for biochemical parameter estimation of soybeans. Similarly, Chu et al. [120] introduced the BBI deep learning model, utilizing datasets from 81 counties in China, including meteorological, area, and rice yield data, demonstrating accurate predictions for both summer and winter rice.

Moreover, Liu et al. [121] achieved a 99.4% classification accuracy for tomato maturity by fusing features from imaging, near-infrared spectroscopy, and tactile modalities into a unified feature set, outperforming single-modality baselines (color imaging: 94.2%; spectroscopy: 87.8%; tactile sensing: 87.2%). Garillos-Manlague et al. [122] proposed a multimodal deep convolutional neural network that concatenated features from visible and hyperspectral imaging to classify papaya maturity into six stages, achieving an F1 score up to 0.90 and a top-2 error rate of 1.45%. In a subsequent study [123], the same team developed an AI-derived nondestructive method using hyperspectral and visible imagery, achieving an improved F1 score of 0.97 through image-specific network fusion. Accurate maturity detection and yield prediction through multimodal fusion allow farmers to optimize harvest times and market strategies, reducing post-harvest losses and ensuring that resources are used efficiently. This not only enhances economic sustainability but also minimizes the environmental footprint by avoiding overproduction and resource waste.

3.1.3. Weed Identification

Efficient weed identification remains a major challenge in farmland management [124,125], directly impacting crop yields and ecological sustainability [126]. Traditional unimodal detection methods are limited in complex environments, often suffering from occlusion and illumination variability [127–129]. Multimodal fusion, by integrating spectral, morphological, spatial, and environmental features, significantly boosts the accuracy and robustness of weed identification, providing critical support for intelligent field management. To provide a clearer understanding of the target detection task, Figure 5 presents common weed species typically found in wheat fields. For example, Xia et al. [130] leveraged multimodal data fusion and deep learning for herbicide resistance scoring by applying three fusion methods based on 3D-CNN and 2D-CNN architectures to UAV-acquired spectral, structural, and textural data, validating the effectiveness of the approach. Xu et al. [131] introduced a dual-path Swin Transformer model for multimodal feature extraction from RGB and depth images, effectively addressing leaf occlusion issues and improving weed detection accuracy to 85.14%. Zamani et al. [37] focused on weed detection in visible and thermal images from paddy fields, developing multiple fusion architectures and demonstrating that a late fusion structure using a Genetic Algorithm (GA) combined with Extreme Learning Machine (ELM) achieved the best performance, with an accuracy of 98.08%. By enhancing the precision of weed identification, these multimodal approaches enable targeted herbicide application, reducing chemical usage and minimizing environmental impact. This is crucial for sustainable weed management, as it helps preserve soil health, reduces water contamination, and supports biodiversity in agricultural ecosystems.

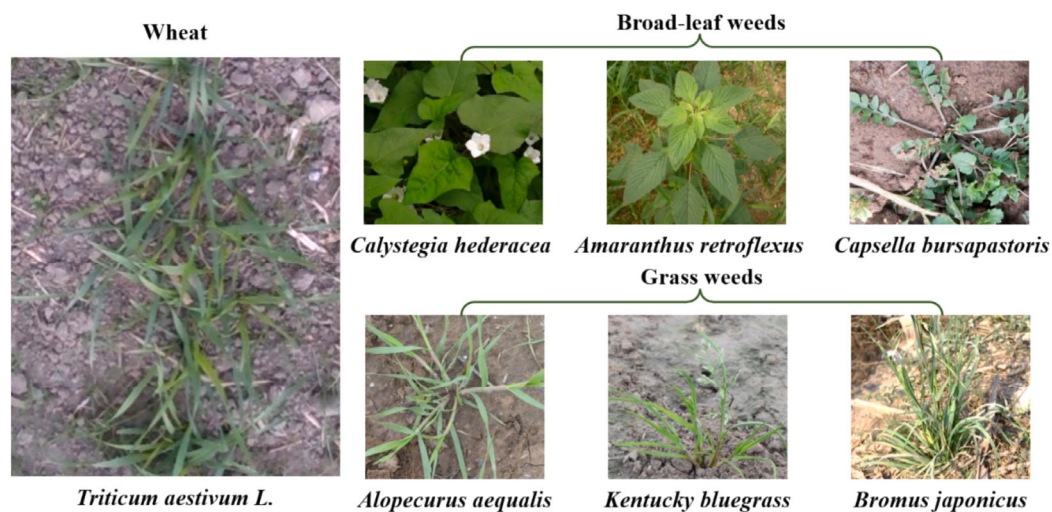


Figure 5. Common weeds in wheat fields [131].

Through systematic analysis of three core application scenarios—disease diagnosis, maturity and yield prediction, and weed identification—this section highlights the transformative impact of multimodal fusion technologies on agricultural detection. The advantages are twofold: enhanced information complementarity, where cross-modal feature interaction mitigates unimodal perception blind spots; and improved environmental adaptability, where dynamic fusion strategies (e.g., contextual embedding, multiscale feature concatenation) significantly strengthen model robustness in complex farmland conditions characterized by illumination variability and occlusions.

Despite remarkable progress, existing research still faces several limitations. First, most studies focus on optimizing single tasks (e.g., disease diagnosis or yield prediction),

with relatively few exploring multitask synergy—for example, simultaneous disease diagnosis and precision weeding remains underexplored. Second, the complexity of multimodal data fusion, particularly in dynamic agricultural environments, poses challenges for efficient real-time data alignment and feature extraction. Third, many models rely on datasets from specific scenarios, and their generalization across regions and crop types requires further validation. Lastly, while multimodal fusion demonstrates superior detection accuracy, its cost-effectiveness and operational feasibility in real-world field deployments warrant additional evaluation.

Future research should prioritize practical and universal solutions, advancing the transition of multimodal fusion technologies from experimental research to large-scale field applications. Table 5 summarizes the applications of multimodal deep learning in comprehensive crop condition assessment. While deep learning-driven multimodal fusion has achieved notable success in disease identification and yield prediction, the realization of real-time field operations also depends on the integration with intelligent agricultural equipment. The next section will focus on the deployment of fusion algorithms in agricultural machinery, harvesting robots, and plant protection robots, completing the loop from “perception” to “execution”.

Table 5. Summary of multimodal deep learning applications in comprehensive crop condition assessment.

Application Scenario	Algorithms	Dataset	Precision	Metric Type	References
Crop Disease Detection	MCFN	Over 50,000 crop disease images and Contextual Information	97.5%	Accuracy	[112]
Banana Localization	RetinaNet, Custom Classifier	Pixel-based multispectral UAV and satellite image dataset for banana classification	92%	Accuracy	[113]
Soybean grain yield prediction	DNN-F1, DNN-F2	UAV-acquired RGB, multispectral, and thermal imaging data	72%	Accuracy	[118]
Estimation of biochemical parameters in soybean	PLSR, SVR, ELR	UAV-acquired RGB, multispectral, and thermal imaging data	22.6%	RMSE	[119]
Rice yield prediction	BBI	summer/winter rice yields, meteorological data, and cultivation areas for 81 counties in Guangxi, China	0.57%	RMSE	[120]
Tomato maturity classification	a fully connected neural network	2568 data collections comprising visual, spectral, and tactile modalities	99.4%	Accuracy	[121]
Papaya growth stage estimation	Deep Convolutional Neural Network	4427 RGB images and 512 hyperspectral (HS) images	90%	F1	[122]
Papaya growth stage estimation	imaging-specific deep convolutional neural networks	Hyperspectral and visible-light images	97%	F1	[123]
Quantification of weed resistance	CNN	Hyperspectral data, High-resolution RGB images and point cloud data	54.7%	RMSE	[130]
Leaf occlusion problem in weed detection	Swin Transform	The multimodal dataset comprises 1288 RGB images and 1288 PHA images, while the multiview dataset contains 692 images	85.14%	Accuracy	[131]
Weed detection in paddy fields	ELM-E	100 pairs of visible and thermal images of rice and weeds	98.08%	Accuracy	[38]

3.2. Agricultural Machinery Intelligence for Crop Detection

Driven by the dual forces of Agriculture 4.0 and labor shortages, agricultural machinery is evolving from simple mechanical substitution towards integrated systems combining multimodal perception, autonomous decision making, and precise execution. Traditional agricultural machinery, reliant on manual operation or preset programs, struggles to adapt to dynamic variables in complex field environments, such as crop growth heterogeneity, terrain undulations, and sudden weather changes. In contrast, multimodal fusion technologies, integrating data from vision sensors, LiDAR, IoT devices, and other sources, endow machinery with real-time environmental perception, autonomous decision making, and precision operation capabilities, thus propelling agriculture from mechanization towards true intelligence. This section focuses on three major application areas: agricultural machinery navigation, harvesting robots, and plant protection robots. These intelligent machinery systems are pivotal to sustainable agricultural development as they enhance operational efficiency, reduce resource consumption, and minimize environmental disruption, supporting the transition towards eco-friendly and resource-efficient farming practices.

3.2.1. Agricultural Machinery Navigation

Achieving autonomous navigation for agricultural machinery in complex field environments is a core requirement of precision agriculture. Traditional GPS-based navigation systems are vulnerable to signal obstructions (e.g., from tall crops or uneven terrain) and dynamic obstacles (e.g., temporary structures or moving personnel), often resulting in path planning deviations [132]. Multimodal fusion technologies, by integrating LiDAR, visual sensors, inertial measurement units (IMU), and high-precision map data, significantly enhance the environmental perception and decision-making capabilities of agricultural machinery in challenging scenarios [133]. Figure 6 illustrates the principles of navigation technology and the hardware architecture of navigation systems.

For example, Zhang et al. [134] designed a field environment point cloud acquisition system based on LiDAR, by mounting LiDAR and GNSS/INS systems on agricultural machinery to capture positional data and generate 3D point clouds of farmland. Through algorithms such as point cloud denoising, segmentation, and clustering, obstacles in the field were successfully identified, and the SLAM framework was employed to optimize the navigation paths. Experimental results demonstrated that this multisensor fusion approach significantly improved navigation accuracy and robustness under complex field conditions. Similarly, Li et al. [135] utilized an extended Kalman filter to integrate environmental information from LiDAR and depth cameras, while posture and acceleration data were obtained from IMUs. A SLAM algorithm based on the fusion of LiDAR, depth cameras, and IMUs was constructed, followed by global path planning using an improved ant colony optimization algorithm, and local planning and obstacle avoidance using the dynamic window approach. Experiments showed that maps generated through multimodal sensor fusion more closely reflected real-world environments, effectively enhancing SLAM accuracy and robustness, while the combination of ant colony optimization and dynamic window methods substantially improved path planning efficiency. These technologies improve the autonomy and precision of agricultural machinery navigation, reducing manual intervention and optimizing fuel consumption. This supports sustainable agriculture.

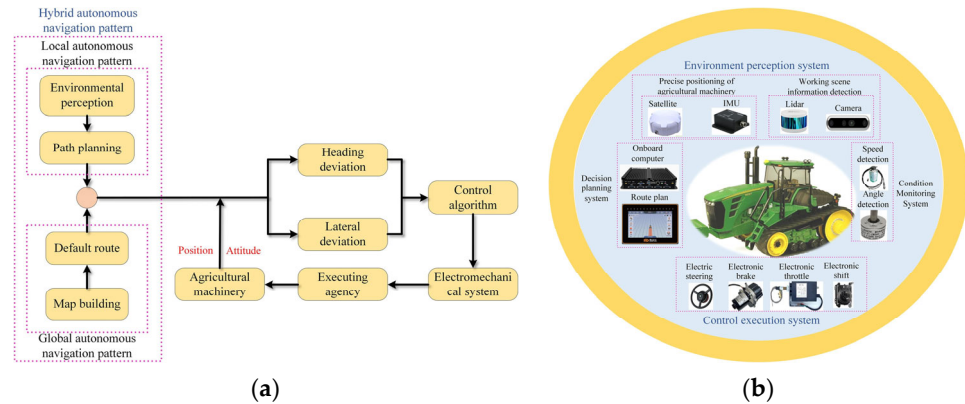


Figure 6. Principle and Hardware Architecture of Agricultural Navigation System [136]. (a) principle of navigation technology; (b) Hardware architecture of agricultural navigation system.

3.2.2. Harvesting Robots

Driven by the advancement of precision agriculture, harvesting robots have emerged as a key technology in agricultural machinery intelligence, offering solutions to labor shortages and enhancing fruit harvesting efficiency [137,138]. Traditional manual harvesting faces challenges such as high labor costs, low efficiency, and seasonal workforce shortages. By integrating multimodal data fusion, harvesting robots combine visual recognition, robotic arm control, and autonomous navigation to accurately assess fruit maturity, position, and morphology, enabling precise and non-destructive harvesting [139,140]. These technologies contribute to sustainable agricultural practices by reducing labor demands and improving harvesting efficiency, which helps minimize resource waste and ensures a more consistent and reliable food supply.

For instance, Birrell et al. [141] developed an iceberg lettuce harvesting robot utilizing a customized vision and learning system based on convolutional neural networks for classification and localization. A force-feedback-enabled custom end-effector was employed to detect the ground and achieve non-destructive harvesting. Experimental results demonstrated a harvesting success rate of 97% and an average harvesting time of 31.7 ± 32.6 s. Similarly, Silwal et al. [142] designed a seven-degree-of-freedom robotic system capable of precise apple localization and harvesting, achieving an average fruit localization time of 1.5 s, an average picking time of 6 s, and a harvesting success rate of 84%. Figure 7 illustrates the end-effectors used in these two systems.

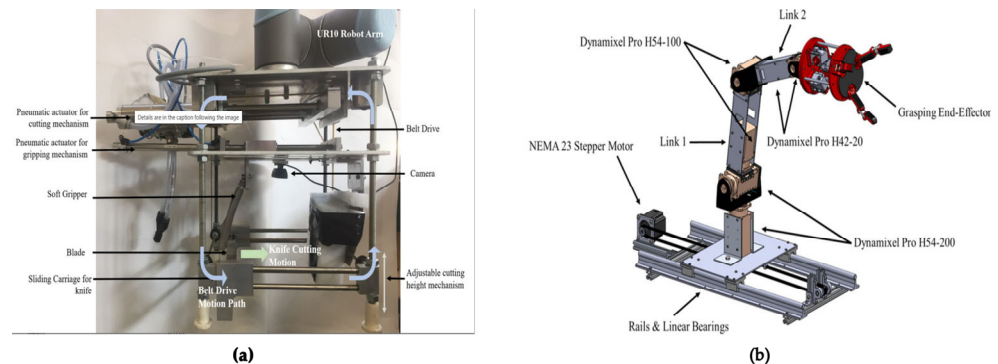


Figure 7. End-effector of harvesting robots: (a) iceberg lettuce picking robot [141]; (b) robotic apple harvester [142].

Additionally, Mao et al. [143] proposed a cucumber harvesting robot recognition method based on deep learning and multi-feature fusion. By employing a multi-path

convolutional neural network (MPCNN) and the I-RELIEF algorithm to select color components, combined with a support vector machine (SVM), the method achieved robust cucumber recognition under complex natural conditions. By analyzing the weights of 15 color components, the most discriminative components (e.g., green, red, and brightness) were selected as input features, significantly improving fruit recognition accuracy and resistance to environmental interference. Experimental results indicated a correct recognition rate exceeding 90% and a false recognition rate below 22%.

3.2.3. Plant Protection Robots

Amid the surge of precision agriculture, plant protection robots are gradually transforming traditional pest management practices through multimodal data fusion and intelligent decision-making capabilities. Conventional pesticide spraying, reliant on manual labor or pre-programmed drone flights, often results in uneven application and excessive chemical usage, leading to resource waste, soil pollution, and ecological imbalance [144,145]. In contrast, modern plant protection robots integrate a range of sensors and meteorological monitoring devices to collect detailed field or orchard data, including terrain, canopy features, and vegetation conditions. Through algorithmic analysis, they generate prescription maps tailored to specific plant needs, enabling differentiated pesticide application based on these maps [146,147]. This establishes an integrated system of perception, decision making, and execution, advancing the shift from broad-spectrum coverage to targeted intervention [148].

For example, Blue River Technology's See & Spray system employs high-resolution cameras and CNNs to distinguish crops from weeds in real-time, enabling targeted herbicide application via high-pressure spraying. This approach has reduced herbicide use in cotton fields by 90%, setting a benchmark for precision spraying [149,150]. Similarly, Chittoor et al. [151] proposed a plant protection robot design methodology based on generative artificial intelligence (Gen-AI) and multimodal data fusion. By combining 3D LiDAR, depth cameras, and inertial measurement units (IMUs), the robot achieves precision spraying for mosquito hotspots. The method optimizes design parameters through multimodal input and feedback mechanisms, ensuring the robot can autonomously navigate and accurately identify target zones in complex urban landscapes. Experimental results demonstrated significant improvements in operational efficiency and environmental safety while reducing chemical usage. Figure 8 illustrates various types of plant protection robots.

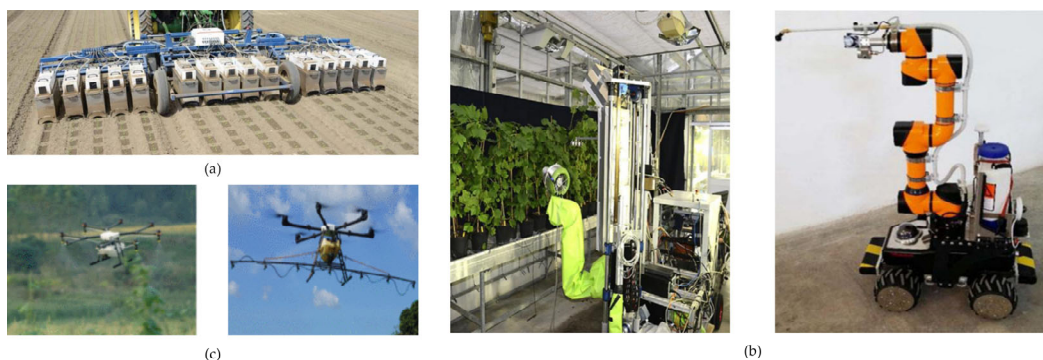


Figure 8. Types of plant protection robots: (a) wheeled plant protection robot [149]; (b) plant protection robot with manipulator [152,153]; (c) plant protection drones [154].

Through the analysis of agricultural navigation systems, harvesting robots, and plant protection robots, this section highlights the transformative role of multimodal data fusion in the intelligentization of agricultural machinery. Significant breakthroughs have

been realized across the perception, decision making, and execution layers. From a technological evolution perspective, three major trends characterize this transition: (1) the extension of perception capabilities from single-modality to cross-modal complementarity, for example, LiDAR compensating for vision-based limitations in 3D terrain modeling, and thermal imaging enhancing RGB-based detection of concealed plant diseases; (2) the shift of decision-making frameworks from rule-based systems to data- and knowledge-driven models, exemplified by the integration of dynamic knowledge graphs and reinforcement learning to optimize obstacle avoidance strategies; and (3) the improvement of execution precision from centimeter level to millimeter level, driven by the adoption of flexible robotic arms and high-precision spraying systems, which substantially reduce operational damage rates. Despite these advancements, several challenges remain, particularly the computational latency caused by the heterogeneity of multimodal data, which continues to hinder real time field deployment. Table 6 summarizes the applications of multimodal fusion technologies in the context of agricultural machinery intelligentization. These advancements in agricultural machinery intelligence are key to promoting sustainable farming practices by enhancing resource efficiency, reducing environmental impact, and improving overall agricultural productivity.

Table 6. Summary of multimodal fusion applications for intelligent agricultural machinery.

Application Scenario	Algorithms	Dataset	Precision	References
Autonomous navigation of agricultural machinery in complex farmland environments	LiDAR, GNSS/INS, Point Cloud Processing Algorithms, SLAM	Self-collected three-dimensional point cloud data of farmland environment	Significant improvement in navigation accuracy, enhanced robustness	[134]
Mobile robot mapping and path planning with multi-sensor information fusion	EKF, Improved Ant Colony Optimization, Dynamic Window Approach, SLAM	LiDAR, Inertial Measurement Unit, Depth Camera	Significant improvement in mapping accuracy and robustness; enhanced path planning efficiency and safety with error within 4 cm	[135]
Iceberg lettuce harvesting robot	Customized Convolutional and CNN	A dataset of 1505 lettuce images annotated with weather conditions, camera height, plant spacing, and maturity stages, complemented by force feedback data	97%	[141]
Apple harvesting robot	Circular Hough Transform, Blob Analysis	RGB and 3D fruit localization data	84%	[142]
Cucumber harvesting robot	MPCNN, I-RELIEF, SVM	Data were collected from a cucumber plantation in Shouguang, Shandong, China, with an image resolution of 1024 × 768 pixels and a total of 218 validated samples	Correct identification rate: >90% with false recognition rate < 22%	[142]
Targeted weeding by plant protection robot	CNN	Field-captured image data, real-time operational telemetry, multi-sensor readings, historical farm records	90% reduction in herbicide usage	[149,150]
Urban Landscape Fumigation	IMU, SLAM	Sensor data from urban environments, prescription maps for targeted spraying, real-time feedback dataset	Reduced chemical usage by optimizing spray, enhanced navigation accuracy during application in complex terrains	[151]

3.3. Resource Management and Ecological Monitoring for Plant Health

Agricultural environment and resource management constitute a core pillar in the transition of precision agriculture from an efficiency-driven to a sustainability-driven model. Traditional extensive farming practices, which have overexploited soil, water

resources, and ecosystems, are approaching critical environmental thresholds. The application of multimodal fusion technologies and digital tools is driving a shift from experience-based to data-driven agricultural management. Intelligent monitoring systems, enabled by comprehensive sensing networks, capture the real-time dynamics of farmland environments, providing a robust data foundation for resource regulation. Full-chain agricultural digitalization bridges information gaps across production and processing stages, facilitating the optimization of resource allocation across time and space. Agricultural ecological sustainability focuses on ecological restoration and resource recycling empowered by technology, addressing the traditional dichotomy between economic growth and environmental protection. These three dimensions—environmental monitoring, digital integration, and ecological restoration—progressively reinforce one another, collectively supporting a green, low-carbon, and efficient circular agricultural system, offering a systematic solution to global food and ecological security.

3.3.1. Smart Monitoring

As the neural hub of precision agriculture, intelligent monitoring systems leverage multimodal sensor networks and real-time data analytics to enable comprehensive dynamic perception and closed-loop management of farmland environments, crop growth, and resource utilization [155–157]. Traditional agricultural monitoring, relying on manual inspections and single-sensor measurements, suffers from data delays and limited coverage. In contrast, intelligent systems integrate multi-source data, such as meteorological, soil, and image information, to establish an integrated monitoring network, providing precise decision-making support for both controlled-environment agriculture and open-field farming [158,159].

In typical applications, intelligent monitoring systems have demonstrated particular effectiveness in precision irrigation and disaster early warning. For precision irrigation, Munaganuri et al. [160] proposed a model named PAMICRM, based on multimodal remote sensing data and machine learning techniques, aimed at optimizing irrigation scheduling and reducing water resource waste. By integrating satellite imagery, high-resolution drone images, and ground sensor data, the model enables real-time monitoring of soil moisture, crop health, and environmental conditions. Experiments showed that PAMICRM improved adjustment accuracy by 8.5% and reduced response latency by 3.5%, effectively minimizing water waste and enhancing resource efficiency. In disaster prevention, Birrell et al. [161] developed a multimodal sensor-based early warning system for monitoring crop diseases and extreme weather events. The system combines RGB-D cameras, LiDAR, and meteorological sensors, employing deep learning models such as YOLOv5 and U-Net to detect early signs of crop diseases and predict potential disaster risks using meteorological data. Experimental results demonstrated that the system could issue warnings up to 24 h in advance, reducing disaster-related losses by more than 30% and significantly improving the accuracy and timeliness of disaster response.

3.3.2. Targeted Resource Allocation

Agricultural ecological sustainability represents one of the most socially valuable goals of multimodal fusion technologies within precision agriculture [162,163]. The negative impacts of conventional extensive farming on soil health, water resources, and biodiversity have become increasingly evident [164,165]. In response, intelligent systems based on multimodal data integration optimize resource use efficiency, reduce chemical inputs, and promote ecological restoration, thus establishing a new agricultural model that harmonizes production, environmental stewardship, and economic development.

In soil health management, the synergistic application of multimodal sensors—such as capacitive moisture probes, near-infrared spectrometers, and electrical conductivity

sensors—enables the dynamic monitoring of soil moisture, nutrient levels, and pollutant concentrations [166]. For instance, Dhakshayani et al. [167] developed a novel multimodal fusion network (M2F-Net) based on deep learning and the Internet of Things (IoT), integrating agrometeorological and imaging data for high-throughput phenotyping to diagnose fertilizer overuse. Experimental results showed that the late-fusion model achieved 91% accuracy, significantly enhancing the classification performance for detecting excessive fertilizer application and helping reduce fertilizer overuse. Bhattacharya et al. [168] proposed a method that seamlessly integrates multimodal data sources—including NPK values, moisture content, image analysis, and geographic information—to provide accurate and customized recommendations for crop and fertilizer management. Their experiments demonstrated a 2.5% improvement in fertilizer recommendation accuracy and a 4.9% increase in crop recommendation accuracy.

In addition, multimodal fusion technologies have shown unique value in water resource management. Kilinc et al. [169] applied a multimodal fusion approach to hydrological time-series prediction by integrating historical flow data, meteorological information, and geographical data. They optimized bidirectional LSTM (Bi-LSTM) and bidirectional GRU (Bi-GRU) models using a particle swarm optimization (PSO) algorithm and introduced a self-attention mechanism to capture complex temporal dependencies. Experimental results demonstrated that the method could stably and accurately predict flow variations, providing reliable support for dynamic irrigation scheduling based on crop water requirements. These technological examples highlight that resource allocation strategies driven by multimodal crop status detection can significantly enhance water and fertilizer use efficiency while reducing environmental risks.

This section systematically illustrates the critical role of multimodal deep learning in agricultural resource management and ecological monitoring. First, *Smart Monitoring* leverages multimodal data fusion to construct real-time sensing networks, enabling intelligent monitoring systems to accurately capture soil moisture dynamics, meteorological variations, and crop conditions. Second, *Targeted Resource Allocation* utilizes crop status assessment results to guide precision fertilization and dynamic irrigation scheduling, with model optimization and self-attention mechanisms enhancing decision-making accuracy and stability. These two modules complement each other, seamlessly linking crop status detection with resource regulation. Together, they not only significantly improve water and fertilizer use efficiency but also provide robust data support for ecological sustainability practices. Despite existing challenges such as sensor latency, data incompatibility, and conflicts in decision-making priorities, the multimodal fusion-based “intelligent agricultural system” is evolving toward greater real-time performance, robustness, and adaptability. Table 7 summarizes the applications and performance metrics of multimodal deep learning in resource management and ecological monitoring.

Table 7. Summary of multimodal deep learning applications in resource management and ecological monitoring.

Application Scenario	Algorithms	Dataset	Precision	References
Irrigation scheduling optimization	GWO	Hydrological datasets across Australia, U.S. Department of Agriculture Dataset Social media data: typically comprises both visual and textual features for flood detection and classification; remote sensing data: generally consists of multispectral satellite imagery for flood detection and mapping.	92.3%	[160]
Flood detection and assessment	CNN, ANN, DBP		96.1%	[161]

Diagnosis of fertilizer overuse	Novel Multimodal Fusion Network	The experiment was conducted in a village in Karaikal, India, where agrometeorological data and image data were collected.	91%	[167]
To provide precise and tailored recommendations for crop and fertilizer management	Graph Convolutional FPMaX Model, Recurrent FPMaX Model	Covering multiple aspects, including crops, soil, meteorology, and fertilizers	The accuracy of crop recommendations was improved by 4.9%; the accuracy of fertilizer recommendations was improved by 2.5 percentage points.	[168]
Hydrological time series forecasting	PSO, Bi-LSTM, Bi-GRU, Self-Attention	A hydrological dataset comprising 3652 daily discharge measurements across the Kızılırmak Basin, Turkey	Root Mean Square Error (RMSE): 0.085; Mean Absolute Error (MAE): 0.040; Coefficient of Determination (R^2): 0.964	[169]

4. Technical Challenges and Future Directions

Although multimodal fusion technologies have introduced groundbreaking solutions for agricultural intelligence, their practical deployment still faces significant challenges. Current obstacles primarily arise from the limited efficiency of multisource data fusion due to heterogeneity, real-time bottlenecks that constrain edge decision making in dynamic environments, and insufficient model generalization, which hampers large-scale cross-domain applications. This section addresses these three core challenges, analyzes their technical roots, and outlines potential pathways for future breakthroughs, offering strategic insights for the development of intelligent agricultural systems.

4.1. Data Heterogeneity

Data heterogeneity constitutes a fundamental challenge for multimodal agricultural systems, rooted in the diverse modalities, mismatched spatiotemporal scales, and semantic fragmentation inherent in agricultural environments [170,171]. It manifests across three levels: modal heterogeneity (structural differences among visual, spectral, and physical sensor data), spatiotemporal heterogeneity (temporal mismatch between instantaneous UAV imagery and continuous soil sensor sampling), and semantic heterogeneity (nonlinear relationships between crop height features from LiDAR point clouds and chlorophyll content from multispectral imagery) [172–179]. This multilayered heterogeneity results in two major dilemmas for fusion models: first, a mapping gap between high-dimensional input data and low-dimensional semantic labels, which forces models to rely on strong assumptions that sacrifice microenvironment specificity; second, spatiotemporal asynchrony among data sources (e.g., hourly meteorological updates versus second-level visual data collection) that can introduce causal inference biases. In practical applications, different types of sensors often collect data on different physical quantities, such as temperature, humidity, and light intensity. These data typically differ in units, dimensions, and statistical distributions, and directly fusing them may lead to inaccurate results [180].

Although spatiotemporal synchronization techniques (as shown in Section 2.1.2) improve data alignment through hardware clock synchronization and coordinate registration, they primarily achieve shallow geometric-level alignment and fail to address deep semantic heterogeneity. For example, while time-series data from soil moisture sensors and spatial features from UAV multispectral imagery can be aligned within a common coordinate system, the delayed influence of soil moisture on crop transpiration may render feature interaction ineffective. Furthermore, the dynamic evolution of agricultural environments (e.g., sudden weather changes) leads to continuous data distribution shifts,

causing static alignment models to lose accuracy under environmental perturbations. In agricultural contexts, data heterogeneity not only increases preprocessing complexity but also directly impacts the efficacy of multisource information fusion, thereby exacerbating real-time performance bottlenecks.

Future breakthroughs may emerge from two directions: dynamic adaptive architectures and cross-modal causal reasoning. Dynamic adaptive architectures emphasize models' online learning capabilities to handle data distribution shifts, for instance, by dynamically adjusting multimodal fusion weights using meta-learning frameworks or incorporating memory-augmented networks to store historical environmental states and predict data evolution trends. Cross-modal causal reasoning aims to disentangle confounding factors and identify causal mechanisms between crop growth and multimodal signals, rather than relying solely on statistical correlations.

From a broader perspective, data heterogeneity is not merely a technical obstacle but an intrinsic property of multimodal agricultural systems. The governance goal should not be to eliminate heterogeneity, but rather to exploit it for difference-driven decision making. The transition from multimodal fusion to multimodal collaboration may represent the critical pathway for advancing multimodal agricultural systems into an era of adaptive intelligence.

4.2. Real-Time Bottlenecks

The real-time bottleneck represents a critical barrier to the deployment of multimodal agricultural systems from theoretical development to field implementation, stemming from the contradiction between the explosive data processing demands and the limited edge computing resources available in agricultural environments [181]. In dynamic decision-making scenarios, such as autonomous obstacle avoidance by agricultural machinery or disaster emergency response, systems are required to complete multimodal data acquisition, fusion, and decision making within milliseconds. However, the computational capacity, communication bandwidth, and energy supply of edge devices in the field often fail to meet these stringent demands. This tension is further exacerbated in complex agricultural settings, where the parallel fusion of LiDAR point clouds from ground robots and time-series data from soil sensors leads to an exponential increase in computational complexity. Even with the adoption of lightweight model compression techniques, the computational burden at cross-modal interaction layers may still exceed the processing capabilities of embedded processors, causing cumulative decision delays that can ultimately trigger control failures or operational risks. In practical applications, such as when a drone is flying at low altitude, the airflow generated by the propellers may cause changes in the shape of weed leaves, such as curling or rolling. These deformations can increase the difficulty of weed identification. If real-time processing bottlenecks prevent the model from handling these dynamically changing images promptly, the accuracy of weed recognition may be reduced [182].

Current technological strategies to address real-time challenges mainly involve model compression, optimization of edge computing architectures, and hardware acceleration [183–185]. Nevertheless, each approach faces inherent limitations. Although model compression techniques, such as knowledge distillation and network pruning, reduce the number of parameters, they often significantly impair generalization performance due to the long-tail distribution of agricultural data, making lightweight models vulnerable during extreme weather events [186]. While edge computing alleviates computational load through task offloading, sparse network coverage and strict data privacy requirements in agricultural fields constrain the use of cloud-based solutions [187]. Specialized hardware accelerators (e.g., TPUs, NPUs) can enhance computational power but are limited by the harsh environmental conditions and cost-sensitivity of agricultural

equipment, hindering widespread adoption. Moreover, real-time bottlenecks are closely tied to generalization limitations. Given the highly dynamic nature of agricultural environments, insufficient real-time performance not only reduces the efficiency of edge decision making but also undermines the model's ability to capture critical features in evolving scenarios, leading to unstable performance when adapting to new conditions.

Future breakthroughs to overcome the real-time bottleneck may hinge on the development of dynamic adaptive computing frameworks. These frameworks emphasize online resource awareness and task priority scheduling to achieve elastic adjustment of computational granularity. For instance, high-precision multimodal fusion could be performed when device resources are ample, while automatic downgrading to rapid single-modal inference could occur during power shortages or sudden data surges, with incremental learning techniques maintaining continuity of model performance. Innovations in communication-computation coupling are also essential. Semantic communication technologies envisioned for 6G networks could transform raw data transmission into high-level feature transmission, significantly reducing the data load on wireless links [188,189]. Additionally, the integration of federated learning and edge inference offers the potential to achieve real-time decision making while safeguarding data privacy through distributed model aggregation [190–193]. Another promising direction is the incorporation of multi-horizontal prediction, where the system makes short-term sequential forecasts across multiple time steps in parallel, rather than relying solely on recursive single-step inference. This strategy can reduce cumulative latency and mitigate the error propagation typically associated with real-time temporal models [194]. In agricultural scenarios, multi-horizontal prediction could be particularly useful for tasks such as disease progression forecasting or irrigation scheduling under rapidly changing conditions [195]. By leveraging partially overlapping feature windows and parallel processing, it may help alleviate computational bottlenecks while preserving prediction continuity and robustness. Further exploration of lightweight, multi-step forecasting architectures tailored for edge deployment could offer a valuable solution to real-time constraints in multimodal agricultural systems [196,197].

4.3. *Insufficient Generalization Capacity*

Limited generalization capacity is a core bottleneck restricting the large-scale deployment of multimodal agricultural systems, fundamentally rooted in the strong heterogeneity of agricultural environments and the spatiotemporal non-stationarity of data distributions [198,199]. Traditional machine learning models rely on the assumption of independent and identically distributed data; however, the dynamic evolution of farmland environments (e.g., gradual soil type changes, shifting climate patterns) and the regional specificity of crop growth (e.g., varietal differences) jointly lead to generalization failures. This deficiency becomes particularly pronounced in cross-region and cross-crop transfer scenarios [200]. At a deeper level, the challenge lies in the disconnect between data-driven surface correlations and the causal complexity of agricultural systems. Existing models often rely on statistical correlations to map features to labels, yet struggle to distinguish confounding factors, resulting in misinterpretation of causal mechanisms and breakdowns in cross-domain generalization [201]. In real-world applications, existing deep learning-based plant disease recognition systems often experience performance degradation under new field conditions and previously unseen data. This is primarily due to the limited generalization ability of deep neural networks when domain shifts occur. Such domain shifts may be caused by factors like disease type, infection stage, lighting, background, and planting time. Moreover, most current plant disease recognition methods operate under the closed-set assumption, which is unrealistic in practice, as the target domain may contain numerous images, only a small portion of which belong to the

categories known during training. In such cases, the system may misclassify new data, leading to suboptimal performance [202].

Current strategies for enhancing generalization primarily revolve around domain adaptation and data augmentation techniques [203–206]. However, the unique characteristics of agricultural scenarios present significant challenges to conventional domain adaptation methods (e.g., adversarial training, feature alignment) [207]. These methods typically assume latent commonalities between source and target domains; yet in crop phenotyping tasks, the high plasticity of plant morphology (e.g., significant phenotypic variations of the same cultivar under different water stress conditions) leads to sparse domain-invariant features [208]. Under such circumstances, forced feature alignment can undermine the discriminative essence of phenotypic traits. Although data augmentation techniques (e.g., GAN-generated synthetic images) help mitigate small sample issues, they often fail to reconstruct the complex physical constraints of agricultural systems (e.g., the dynamic coupling between root development and soil texture), resulting in a semantic gap between synthetic and real-world data [209].

Future research in agricultural intelligence is likely to focus on the synergistic integration of self-supervised pretraining and meta-learning frameworks. Specifically, self-supervised pretraining can extract universal feature representations from large-scale unlabeled multimodal data (e.g., satellite time-series imagery), capturing both the physical laws of farmland environments and the biological rhythms of crops, while significantly reducing dependence on manual annotations [210,211]. Meanwhile, meta-learning can establish task generalization mechanisms, enabling models to transfer across diverse scenarios; for example, enabling fine-tuning with only a few labeled samples of new crop varieties to effectively address small-sample generalization challenges [212–214]. An illustrative example is the Task-Informed Meta-Learning method proposed by Tseng et al. [215], which dynamically adjusts model weights based on task metadata (e.g., geographic location, crop type), substantially enhancing performance in cross-crop transfer tasks. To improve readability and provide a more systematic understanding, Table 8 summarizes the key bottlenecks identified in this section along with their corresponding improvement strategies and enabling technologies.

Table 8. Summary of major bottlenecks and corresponding solutions in multimodal agricultural systems.

Bottleneck	Problem Description	Proposed Solutions	Key Technologies
Data Heterogeneity	Multimodal data vary in structure, spatiotemporal resolution, and semantic meaning, making integration challenging.	Dynamic adaptive architectures; cross-modal causal reasoning	Meta-learning, memory-augmented networks, causal graphs
Real-Time Processing Bottlenecks	Limited computational resources on edge devices hinder high-frequency, real-time decision making in the field.	Dynamic computation frameworks; semantic communication; federated learning	Model compression, 6G semantic transmission, edge inference
Insufficient Generalization Capacity	Models perform poorly when transferred to new regions or crop types due to strong environmental variability.	Self-supervised pretraining; meta-learning	Large-scale unsupervised feature learning, task metadata adaptation

5. Conclusions

Multimodal fusion technology has emerged as a key driver of the smart transformation of Agriculture 4.0, offering innovative solutions to complex challenges such as crop detection, autonomous machinery operation, and resource management through the deep integration of heterogeneous data sources. This review systematically outlines the development trajectory of multimodal fusion in agricultural intelligence from three perspectives: technical frameworks, application practices, and core challenges, highlighting its potential and value across the full chain of perception, analysis, and decision making.

Currently, the deployment of multimodal fusion in agricultural scenarios faces significant obstacles, including the following: (i) difficulties in feature alignment and semantic fusion caused by the heterogeneity of multisource data; (ii) the contradiction between limited edge computing resources and the demands of real-time processing of high-dimensional data; and (iii) the poor generalization of models due to the strong heterogeneity of agricultural environments. These challenges reflect the inherent complexity of agricultural systems and place greater demands on existing deep learning-based fusion frameworks.

In the future, emerging technologies such as cross-modal generative modeling, federated learning, and causal inference are expected to further empower multimodal fusion in agricultural intelligence. Meanwhile, the deep integration of full-chain agricultural digitization with ecological management holds the promise of reshaping the collaborative development of resources, production, and the environment, accelerating the evolution of agricultural systems from precision farming toward autonomy and green sustainability. Ultimately, multimodal deep fusion technologies are poised to play a pivotal role in advancing global food security, ecological protection, and carbon neutrality goals.

This review provides theoretical insights and practical guidance for the application of multimodal fusion in agricultural intelligence. Future research should focus on the collaborative optimization of data acquisition, model design, and hardware deployment to facilitate the transition from laboratory prototypes to large-scale field implementations, thereby fostering the development of resilient, efficient, and sustainable smart agricultural systems.

Author Contributions: Conceptualization, P.H. and W.-H.S.; methodology, Z.-X.Y., P.H. and W.-H.S.; formal analysis, Z.-X.Y., Y.L. and R.-F.W.; investigation, Z.-X.Y., Y.L. and R.-F.W.; resources, P.H. and W.-H.S.; writing—original draft preparation, Z.-X.Y., Y.L. and R.-F.W.; writing—review and editing, P.H. and W.-H.S.; visualization, Z.-X.Y., Y.L. and R.-F.W.; supervision, P.H. and W.-H.S.; project administration, P.H. and W.-H.S.; funding acquisition, W.-H.S. All authors have read and agreed to the published version of the manuscript.

Funding: The research was funded by National Natural Science Foundation of China [grant number 32371991] and the 2115 Talent Development Program of China Agricultural University.

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

References

1. Falcon, W.P.; Naylor, R.L.; Shankar, N.D. Rethinking global food demand for 2050. *Popul. Dev. Rev.* **2022**, *48*, 921–957.
2. Klerkx, L.; Jakku, E.; Labarthe, P. A review of social science on digital agriculture, smart farming and agriculture 4.0: New contributions and a future research agenda. *NJAS-Wagen J. Life Sci.* **2019**, *90*, 100315.
3. Javaid, M.; Haleem, A.; Singh, R.P.; Suman, R. Enhancing smart farming through the applications of Agriculture 4.0 technologies. *Int. J. Intell. Netw.* **2022**, *3*, 150–164.
4. Wang, R.; Su, W. The application of deep learning in the whole potato production Chain: A Comprehensive review. *Agriculture* **2024**, *14*, 1225.

5. Tu, Y.; Wang, R.; Su, W. Active disturbance rejection control—New trends in agricultural cybernetics in the future: A comprehensive review. *Machines* **2025**, *13*, 111.
6. Fu, L.; Gao, F.; Wu, J.; Li, R.; Karkee, M.; Zhang, Q. Application of consumer RGB-D cameras for fruit detection and localization in field: A critical review. *Comput. Electron. Agric.* **2020**, *177*, 105687.
7. Gené-Mola, J.; Llorens, J.; Rosell-Polo, J.R.; Gregorio, E.; Arnó, J.; Solanelles, F.; Martínez-Casasnovas, J.A.; Escolà, A. Assessing the performance of rgb-d sensors for 3d fruit crop canopy characterization under different operating and lighting conditions. *Sensors* **2020**, *20*, 7072.
8. Li, W.; Du, Z.; Xu, X.; Bai, Z.; Han, J.; Cui, M.; Li, D. A review of aquaculture: From single modality analysis to multimodality fusion. *Comput. Electron. Agric.* **2024**, *226*, 109367.
9. El Sakka, M.; Ivanovici, M.; Chaari, L.; Mothe, J. A review of CNN applications in smart agriculture using multimodal data. *Sensors* **2025**, *25*, 472.
10. Lahat, D.; Adali, T.; Jutten, C. Multimodal data fusion: An overview of methods, challenges, and prospects. *Proc. IEEE Inst. Electr. Electron. Eng.* **2015**, *103*, 1449–1477.
11. 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.
12. Vadivambal, R.; Jayas, D.S. Applications of thermal imaging in agriculture and food industry—A review. *Food Bioproc. Tech.* **2011**, *4*, 186–199.
13. Polk, S.L.; Chan, A.H.; Cui, K.; Plemmons, R.J.; Coomes, D.A.; Murphy, J.M. Unsupervised detection of ash dieback disease (*Hymenoscyphus fraxineus*) using diffusion-based hyperspectral image clustering. In Proceeding of the IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 2287–2290.
14. Polk, S.L.; Cui, K.; Chan, A.H.; Coomes, D.A.; Plemmons, R.J.; Murphy, J.M. Unsupervised diffusion and volume maximization-based clustering of hyperspectral images. *Remote Sens.* **2023**, *15*, 1053.
15. Cui, K.; Li, R.; Polk, S.L.; Lin, Y.; Zhang, H.; Murphy, J.M.; Plemmons, R.J.; Chan, R.H. Superpixel-based and spatially-regularized diffusion learning for unsupervised hyperspectral image clustering. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–18.
16. Deng, L.; Mao, Z.; Li, X.; Hu, Z.; Duan, F.; Yan, Y. UAV-based multispectral remote sensing for precision agriculture: A comparison between different cameras. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 124–136.
17. Olson, D.; Anderson, J. Review on unmanned aerial vehicles, remote sensors, imagery processing, and their applications in agriculture. *Agron. J.* **2021**, *113*, 971–992.
18. Shi, J.; Bai, Y.; Diao, Z.; Zhou, J.; Yao, X.; Zhang, B. Row detection BASED navigation and guidance for agricultural robots and autonomous vehicles in row-crop fields: Methods and applications. *Agronomy* **2023**, *13*, 1780.
19. Moreno, H.; Valero, C.; Bengochea-Guevara, J.M.; Ribeiro, Á.; Garrido-Izard, M.; Andújar, D. On-ground vineyard reconstruction using a LiDAR-based automated system. *Sensors* **2020**, *20*, 1102.
20. Wang, X.; Pan, H.; Guo, K.; Yang, X.; Luo, S. The evolution of LiDAR and its application in high precision measurement. *IOP Conf. Ser. Earth Environ. Sci.* **2020**, *502*, 12008. <https://doi.org/10.1088/1755-1315/502/1/012008>.
21. Ishimwe, R.; Abutaleb, K.; Ahmed, F. Applications of thermal imaging in agriculture—A review. *Adv. Remote Sens.* **2014**, *3*, 128–140.
22. Xu, X.; Wang, L.; Shu, M.; Liang, X.; Ghafoor, A.Z.; Liu, Y.; Ma, Y.; Zhu, J. Detection and counting of maize leaves based on two-stage deep learning with UAV-based RGB image. *Remote Sens.* **2022**, *14*, 5388.
23. Li, L.; Qiao, J.; Yao, J.; Li, J.; Li, L. Automatic freezing-tolerant rapeseed material recognition using UAV images and deep learning. *Plant Methods* **2022**, *18*, 1–13.
24. Lu, Y.; Liu, M.; Li, C.; Liu, X.; Cao, C.; Li, X.; Kan, Z. Precision fertilization and irrigation: Progress and applications. *AgriEngineering* **2022**, *4*, 626–655.
25. Ahmad, U.; Alvino, A.; Marino, S. Solar fertigation: A sustainable and smart IoT-based irrigation and fertilization system for efficient water and nutrient management. *Agronomy* **2022**, *12*, 1012.
26. Behmann, J.; Acebron, K.; Emin, D.; Bennertz, S.; Matsubara, S.; Thomas, S.; Bohnenkamp, D.; Kuska, M.T.; Jussila, J.; Salo, H. Specim IQ: Evaluation of a new, miniaturized handheld hyperspectral camera and its application for plant phenotyping and disease detection. *Sensors* **2018**, *18*, 441.
27. Pozo, S.D.; Rodríguez-González, P.; Hernández-López, D.; Felipe-García, B. Vicarious radiometric calibration of a multispectral camera on board an unmanned aerial system. *Remote Sens.* **2014**, *6*, 1918–1937.
28. Van den Bergh, M.; Van Gool, L. Combining RGB and ToF cameras for real-time 3D hand gesture interaction. In Proceeding of the 2011 IEEE Workshop on Applications Of Computer Vision (WACV), Kona, HI, USA, 5–7 January 2011; pp. 66–72.

29. Szajewska, A. Development of the thermal imaging camera (TIC) technology. *Procedia Eng.* **2017**, *172*, 1067–1072.
30. Harrap, R.; Lato, M. An overview of LIDAR: Collection to application. *NGI Publ.* **2010**, *2*, 1–9.
31. Leone, M. Advances in fiber optic sensors for soil moisture monitoring: A review. *Results Opt.* **2022**, *7*, 100213.
32. Wang, Z.; Liu, Y.; Duan, Y.; Li, X.; Zhang, X.; Ji, J.; Dong, E.; Zhang, Y. USTC FLICAR: A sensors fusion dataset of LiDAR-inertial-camera for heavy-duty autonomous aerial work robots. *Int. J. Robot. Res.* **2023**, *42*, 1015–1047.
33. Nidamanuri, R.R.; Jayakumari, R.; Ramiya, A.M.; Astor, T.; Wachendorf, M.; Buerkert, A. High-resolution multispectral imagery and LiDAR point cloud fusion for the discrimination and biophysical characterisation of vegetable crops at different levels of nitrogen. *Biosyst. Eng.* **2022**, *222*, 177–195.
34. Liu, J.; Liang, X.; Hyyppä, J.; Yu, X.; Lehtomäki, M.; Pyörälä, J.; Zhu, L.; Wang, Y.; Chen, R. Automated matching of multiple terrestrial laser scans for stem mapping without the use of artificial references. *Int. J. Appl. Earth Obs. Geoinf.* **2017**, *56*, 13–23.
35. Couprie, C.; Farabet, C.; Najman, L.; LeCun, Y. Indoor semantic segmentation using depth information. *arXiv* **2013**, arXiv:1301.3572.
36. Zhao, F.; Zhang, C.; Geng, B. Deep multimodal data fusion. *ACM Comput. Surv.* **2024**, *56*, 1–36.
37. Zamani, S.A.; Baleghi, Y. Early/late fusion structures with optimized feature selection for weed detection using visible and thermal images of paddy fields. *Precis. Agric.* **2023**, *24*, 482–510.
38. Hu, X.; Yang, K.; Fei, L.; Wang, K. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In Proceeding of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1440–1444.
39. Hung, S.; Lo, S.; Hang, H. Incorporating luminance, depth and color information by a fusion-based network for semantic segmentation. In Proceeding of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 2374–2378.
40. Man, Q. Fusion of Hyperspectral and LiDAR Data for Urban Land Use Classification. Ph.D. Dissertation, School of Geographic Sciences at East China Normal University, Shanghai, China, 2015.
41. Sun, Y.; Zuo, W.; Yun, P.; Wang, H.; Liu, M. FuseSeg: Semantic segmentation of urban scenes based on RGB and thermal data fusion. *IEEE Trans. Autom. Sci. Eng.* **2020**, *18*, 1000–1011.
42. He, X.; Cai, Q.; Zou, X.; Li, H.; Feng, X.; Yin, W.; Qian, Y. Multi-modal late fusion rice seed variety classification based on an improved voting method. *Agriculture* **2023**, *13*, 597.
43. Gadzicki, K.; Khamsehashari, R.; Zetzsche, C. Early vs late fusion in multimodal convolutional neural networks. In Proceeding of the 2020 IEEE 23rd International Conference on Information Fusion (FUSION), Rustenburg, South Africa, 6–9 July 2020; pp. 1–6.
44. Wang, Y.; Huang, W.; Sun, F.; Xu, T.; Rong, Y.; Huang, J. Deep multimodal fusion by channel exchanging. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 4835–4845.
45. Huang, N.; Jiao, Q.; Zhang, Q.; Han, J. Middle-level feature fusion for lightweight RGB-D salient object detection. *IEEE Trans. Image Process.* **2022**, *31*, 6621–6634.
46. Mangai, U.G.; Samanta, S.; Das, S.; Chowdhury, P.R. A survey of decision fusion and feature fusion strategies for pattern classification. *IETE Tech. Rev.* **2010**, *27*, 293–307.
47. Sinha, A.; Chen, H.; Danu, D.G.; Kirubarajan, T.; Farooq, M. Estimation and decision fusion: A survey. *Neurocomputing* **2008**, *71*, 2650–2656.
48. Jeon, B.; Landgrebe, D.A. Decision fusion approach for multitemporal classification. *IEEE Trans. Geosci. Remote Sens.* **1999**, *37*, 1227–1233.
49. Kamilaris, A.; Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* **2018**, *147*, 70–90.
50. Zhu, N.; Liu, X.; Liu, Z.; Hu, K.; Wang, Y.; Tan, J.; Huang, M.; Zhu, Q.; Ji, X.; Jiang, Y. Deep learning for smart agriculture: Concepts, tools, applications, and opportunities. *Int. J. Agric. Biol. Eng.* **2018**, *11*, 32–44.
51. Grm, K.; Štruc, V.; Artiges, A.; Caron, M.; Ekenel, H.K. Strengths and weaknesses of deep learning models for face recognition against image degradations. *IET Biom.* **2018**, *7*, 81–89.
52. Kamilaris, A.; Prenafeta-Boldú, F.X. A review of the use of convolutional neural networks in agriculture. *J. Agric. Sci.* **2018**, *156*, 312–322.
53. Zhao, C.-T.; Wang, R.-F.; Tu, Y.-H.; Pang, X.-X.; Su, W.-H. Automatic lettuce weed detection and classification based on optimized convolutional neural networks for robotic weed control. *Agronomy* **2024**, *14*, 2838. Available online: https://proceedings.neurips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html (accessed on 20 April 2025).

54. Coulibaly, S.; Kamsu-Foguem, B.; Kamissoko, D.; Traore, D. Deep neural networks with transfer learning in millet crop images. *Comput. Ind.* **2019**, *108*, 115–120.
55. Pan, C.; Qu, Y.; Yao, Y.; Wang, M. HybridGNN: A Self-Supervised graph neural network for efficient maximum matching in bipartite graphs. *Symmetry* **2024**, *16*, 1631.
56. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*.
57. Lu, Y.; Chen, D.; Olaniyi, E.; Huang, Y. Generative adversarial networks (GANs) for image augmentation in agriculture: A systematic review. *Comput. Electron. Agric.* **2022**, *200*, 107208.
58. Lu, C.; Rustia, D.J.A.; Lin, T. Generative adversarial network based image augmentation for insect pest classification enhancement. *IFAC-PapersOnLine* **2019**, *52*, 1–5.
59. Rizvi, S.K.J.; Azad, M.A.; Fraz, M.M. Spectrum of advancements and developments in multidisciplinary domains for generative adversarial networks (GANs). *Arch. Comput. Methods Eng.* **2021**, *28*, 4503–4521.
60. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
61. Zhou, G.; Rui-Feng, W. The Heterogeneous Network Community Detection Model Based on Self-Attention. *Symmetry* **2025**, *17*, 432.
62. Xie, W.; Zhao, M.; Liu, Y.; Yang, D.; Huang, K.; Fan, C.; Wang, Z. Recent advances in Transformer technology for agriculture: A comprehensive survey. *Eng. Appl. Artif. Intell.* **2024**, *138*, 109412.
63. Wang, Z.; Wang, R.; Wang, M.; Lai, T.; Zhang, M. Self-supervised transformer-based pre-training method with General Plant Infection dataset. In Proceeding of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Xiamen, China, 13–15 October 2024; pp. 189–202.
64. Xu, J.; Zhu, Y.; Zhong, R.; Lin, Z.; Xu, J.; Jiang, H.; Huang, J.; Li, H.; Lin, T. DeepCropMapping: A multi-temporal deep learning approach with improved spatial generalizability for dynamic corn and soybean mapping. *Remote Sens. Environ.* **2020**, *247*, 111946.
65. Xiong, X.; Zhong, R.; Tian, Q.; Huang, J.; Zhu, L.; Yang, Y.; Lin, T. Daily DeepCropNet: A hierarchical deep learning approach with daily time series of vegetation indices and climatic variables for corn yield estimation. *ISPRS J. Photogramm. Remote Sens.* **2024**, *209*, 249–264.
66. Li, W.; Wang, C.; Cheng, G.; Song, Q. Optimum-statistical Collaboration Towards General and Efficient Black-box Optimization. *Trans. Mach. Learn. Res.* **2023**. Available online: <https://par.nsf.gov/servlets/purl/10418406> (accessed on 20 April 2025).
67. Chai, L.; Qu, Y.; Zhang, L.; Liang, S.; Wang, J. Estimating time-series leaf area index based on recurrent nonlinear autoregressive neural networks with exogenous inputs. *Int. J. Remote Sens.* **2012**, *33*, 5712–5731.
68. Fang, W.; Chen, Y.; Xue, Q. Survey on research of RNN-based spatio-temporal sequence prediction algorithms. *J. Big Data* **2021**, *3*, 97.
69. Cao, Y.; Sun, Z.; Li, L.; Mo, W. A study of sentiment analysis algorithms for agricultural product reviews based on improved bert model. *Symmetry* **2022**, *14*, 1604.
70. Sayeed, M.S.; Mohan, V.; Muthu, K.S. Bert: A review of applications in sentiment analysis. *HighTech Innov. J.* **2023**, *4*, 453–462.
71. Murugesan, R.; Mishra, E.; Krishnan, A.H. Forecasting agricultural commodities prices using deep learning-based models: Basic LSTM, bi-LSTM, stacked LSTM, CNN LSTM, and convolutional LSTM. *Int. J. Sustain. Agric. Manag. Inform.* **2022**, *8*, 242–277.
72. Attri, I.; Awasthi, L.K.; Sharma, T.P.; Rathee, P. A review of deep learning techniques used in agriculture. *Ecol. Inform.* **2023**, *77*, 102217.
73. Ayesha Barvin, P.; Sampradeepraj, T. Crop recommendation systems based on soil and environmental factors using graph convolution neural network: A systematic literature review. *Eng. Proc.* **2023**, *58*, 97.
74. Gupta, A.; Singh, A. Agri-gnn: A novel genotypic-topological graph neural network framework built on graphsage for optimized yield prediction. *arXiv* **2023**, arXiv:2310.13037.
75. Manzano, R.M.; Pérez, J.E. Theoretical framework and methods for the analysis of the adoption-diffusion of innovations in agriculture: A bibliometric review. *Boletín De La Asoc. De Geógrafos Españoles* **2023**, *96*, 4.
76. Macours, K. Farmers' demand and the traits and diffusion of agricultural innovations in developing countries. *Annu. Rev. Resour. Econ.* **2019**, *11*, 483–499.
77. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 1–74.

78. Chua, L.O.; Roska, T. The CNN paradigm. *IEEE Trans. Circuits Syst. I Fundam. Theory Appl.* **1993**, *40*, 147–156.
79. Pearton, S.J.; Zolper, J.C.; Shul, R.J.; Ren, F. GaN: Processing, defects, and devices. *J. Appl. Phys.* **1999**, *86*, 1–78.
80. Borji, A. Pros and cons of GAN evaluation measures: New developments. *Comput. Vis. Image Underst.* **2022**, *215*, 103329.
81. Min, E.; Chen, R.; Bian, Y.; Xu, T.; Zhao, K.; Huang, W.; Zhao, P.; Huang, J.; Ananiadou, S.; Rong, Y. Transformer for graphs: An overview from architecture perspective. *arXiv* **2022**, arXiv:2202.08455.
82. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y. A survey on vision transformer. *IEEE Trans. Pattern. Anal. Mach. Intell* **2022**, *45*, 87–110.
83. Sherstinsky, A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys. D Nonlinear Phenom.* **2020**, *404*, 132306.
84. Yin, W.; Kann, K.; Yu, M.; Schütze, H. Comparative study of CNN and RNN for natural language processing. *arXiv* **2017**, arXiv:1702.01923.
85. Pareja, A.; Domeniconi, G.; Chen, J.; Ma, T.; Suzumura, T.; Kanezashi, H.; Kaler, T.; Schardl, T.; Leiserson, C. Evolvegc: Evolving graph convolutional networks for dynamic graphs. In Proceeding of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 5363–5370.
86. Sankar, A.; Wu, Y.; Gou, L.; Zhang, W.; Yang, H. Dynamic graph representation learning via self-attention networks. *arXiv* **2018**, arXiv:1812.09430.
87. Yang, L.; Chatelain, C.; Adam, S. Dynamic graph representation learning with neural networks: A survey. *IEEE Access* **2024**, *12*, 43460–43484.
88. Zhou, G.; Wang, R.; Cui, K. A local perspective-based model for overlapping community detection. *arXiv* **2025**, arXiv:2503.21558.
89. Chen, Y.; Xing, X. Constructing dynamic knowledge graph based on ontology modeling and neo4j graph database. In Proceeding of the 2022 5th International Conference on Artificial Intelligence and Big Data (ICAIBD), Fuzhou, China, 8–10 July 2022; pp. 522–525.
90. Kaelbling, L.P.; Littman, M.L.; Moore, A.W. Reinforcement learning: A survey. *J. Artif. Intell. Res.* **1996**, *4*, 237–285.
91. Wiering, M.A.; Van Otterlo, M. Reinforcement learning. *Adapt. Learn. Optim.* **2012**, *12*, 729.
92. Sun, L.; Yang, Y.; Hu, J.; Porter, D.; Marek, T.; Hillyer, C. Reinforcement learning control for water-efficient agricultural irrigation. In Proceeding of the 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC), Guangzhou, China, 12–15 December 2017; pp. 1334–1341.
93. Chen, M.; Cui, Y.; Wang, X.; Xie, H.; Liu, F.; Luo, T.; Zheng, S.; Luo, Y. A reinforcement learning approach to irrigation decision-making for rice using weather forecasts. *Agric. Water Manag.* **2021**, *250*, 106838.
94. Renaudo, E.; Girard, B.; Chatila, R.; Khamassi, M. Respective advantages and disadvantages of model-based and model-free reinforcement learning in a robotics neuro-inspired cognitive architecture. *Procedia Comput. Sci.* **2015**, *71*, 178–184.
95. Jain, S.; Ramesh, D.; Bhattacharya, D. A multi-objective algorithm for crop pattern optimization in agriculture. *Appl. Soft. Comput.* **2021**, *112*, 107772.
96. Groot, J.C.; Oomen, G.J.; Rossing, W.A. Multi-objective optimization and design of farming systems. *Agric. Syst.* **2012**, *110*, 63–77.
97. Li, Z.; Sun, C.; Wang, H.; Wang, R. Hybrid optimization of phase masks: Integrating Non-Iterative methods with simulated annealing and validation via tomographic measurements. *Symmetry* **2025**, *17*, 530.
98. Qin, Y.; Tu, Y.; Li, T.; Ni, Y.; Wang, R.; Wang, H. Deep learning for sustainable agriculture: A systematic review on applications in lettuce cultivation. *Sustainability* **2025**, *17*, 3190.
99. Habibi Davijani, M.; Banihabib, M.E.; Nadjafzadeh Anvar, A.; Hashemi, S.R. Multi-objective optimization model for the allocation of water resources in arid regions based on the maximization of socioeconomic efficiency. *Water Resour. Manag.* **2016**, *30*, 927–946.
100. Zhou, Y.; Fan, H. Research on multi objective optimization model of sustainable agriculture industrial structure based on genetic algorithm. *J. Intell Fuzzy Syst.* **2018**, *35*, 2901–2907.
101. Li, Q.; Yu, G.; Wang, J.; Liu, Y. A deep multimodal generative and fusion framework for class-imbalanced multimodal data. *Multimed. Tools Appl.* **2020**, *79*, 25023–25050.
102. Zhang, Q.; Wei, Y.; Han, Z.; Fu, H.; Peng, X.; Deng, C.; Hu, Q.; Xu, C.; Wen, J.; Hu, D. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv* **2024**, arXiv:2404.18947.
103. Pawłowski, M.; Wróblewska, A.; Sysko-Romańczuk, S. Effective techniques for multimodal data fusion: A comparative analysis. *Sensors* **2023**, *23*, 2381.

104. Xu, K.; Yu, Z.; Wang, X.; Mi, M.B.; Yao, A. Enhancing video super-resolution via implicit resampling-based alignment. In Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 2546–2555.
105. Zhu, H.; Wang, Z.; Shi, Y.; Hua, Y.; Xu, G.; Deng, L. Multimodal Fusion Method Based on Self-Attention Mechanism. *Wireless Commun. Mob. Comput.* **2020**, *2020*, 8843186.
106. Shang, Y.; Gao, C.; Chen, J.; Jin, D.; Ma, H.; Li, Y. Enhancing adversarial robustness of multi-modal recommendation via modality balancing. In Proceeding of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 6274–6282.
107. Li, J.; Xu, M.; Xiang, L.; Chen, D.; Zhuang, W.; Yin, X.; Li, Z. Foundation models in smart agriculture: Basics, opportunities, and challenges. *Comput. Electron. Agric.* **2024**, *222*, 109032.
108. Yu, K.; Xu, W.; Zhang, C.; Dai, Z.; Ding, J.; Yue, Y.; Zhang, Y.; Wu, Y. ITFNet-API: Image and Text Based Multi-Scale Cross-Modal Feature Fusion Network for Agricultural Pest Identification. 2023. Available online: <https://www.researchsquare.com/article/rs-3589884/v1> (accessed on 20 April 2025).
109. Pantazi, X.; Moshou, D.; Bochtis, D. *Intelligent Data Mining and Fusion Systems in Agriculture*; Academic Press: Cambridge, MA, USA, 2019; ISBN 0128143924.
110. Yashodha, G.; Shalini, D. An integrated approach for predicting and broadcasting tea leaf disease at early stage using IoT with machine learning—A review. *Mater. Today Proc.* **2021**, *37*, 484–488.
111. Patil, R.R.; Kumar, S. Rice-fusion: A multimodality data fusion framework for rice disease diagnosis. *IEEE Access* **2022**, *10*, 5207–5222.
112. Zhao, Y.; Liu, L.; Xie, C.; Wang, R.; Wang, F.; Bu, Y.; Zhang, S. An effective automatic system deployed in agricultural Internet of Things using Multi-Context Fusion Network towards crop disease recognition in the wild. *Appl. Soft Comput.* **2020**, *89*, 106128.
113. Selvaraj, M.G.; Vergara, A.; Montenegro, F.; Ruiz, H.A.; Safari, N.; Raymaekers, D.; Ocimati, W.; Ntamwira, J.; Tits, L.; Omondi, A.B. Detection of banana plants and their major diseases through aerial images and machine learning methods: A case study in DR Congo and Republic of Benin. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 110–124.
114. Li, B.; Lecourt, J.; Bishop, G. Advances in non-destructive early assessment of fruit ripeness towards defining optimal time of harvest and yield prediction—A review. *Plants* **2018**, *7*, 3.
115. Surya Prabha, D.; Satheesh Kumar, J. Assessment of banana fruit maturity by image processing technique. *J. Food Sci. Technol.* **2015**, *52*, 1316–1327.
116. Radu, V.; Tong, C.; Bhattacharya, S.; Lane, N.D.; Mascolo, C.; Marina, M.K.; Kawsar, F. Multimodal deep learning for activity and context recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *1*, 1–27.
117. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444.
118. Maimaitijiang, M.; Sagan, V.; Sidike, P.; Hartling, S.; Esposito, F.; Fritschi, F.B. Soybean yield prediction from UAV using multimodal data fusion and deep learning. *Remote Sens. Environ.* **2020**, *237*, 111599.
119. Maimaitijiang, M.; Ghulam, A.; Sidike, P.; Hartling, S.; Maimaitiyiming, M.; Peterson, K.; Shavers, E.; Fishman, J.; Peterson, J.; Kadam, S. Unmanned Aerial System (UAS)-based phenotyping of soybean using multi-sensor data fusion and extreme learning machine. *ISPRS J. Photogramm. Remote Sens.* **2017**, *134*, 43–58.
120. Chu, Z.; Yu, J. An end-to-end model for rice yield prediction using deep learning fusion. *Comput. Electron. Agric.* **2020**, *174*, 105471.
121. Liu, Y.; Wei, C.; Yoon, S.; Ni, X.; Wang, W.; Liu, Y.; Wang, D.; Wang, X.; Guo, X. Development of multimodal fusion technology for tomato maturity assessment. *Sensors* **2024**, *24*, 2467.
122. Garillos-Manliguez, C.A.; Chiang, J.Y. Multimodal deep learning and visible-light and hyperspectral imaging for fruit maturity estimation. *Sensors* **2021**, *21*, 1288.
123. Garillos-Manliguez, C.A.; Chiang, J.Y. Multimodal deep learning via late fusion for non-destructive papaya fruit maturity classification. In Proceeding of the 2021 18th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE), Mexico City, Mexico, 10–12 November 2021; pp. 1–6.
124. Colbach, N.; Fernier, A.; Le Corre, V.; Messéan, A.; Darmency, H. Simulating changes in cropping practises in conventional and glyphosate-tolerant maize. I. Effects on weeds. *Environ. Sci. Pollut. Res. Int.* **2017**, *24*, 11582–11600.
125. Wang, R.; Tu, Y.; Chen, Z.; Zhao, C.; Su, W. A Lettpoint-Yolov11l Based Intelligent Robot for Precision Intra-Row Weeds Control in Lettuce. Available at SSRN 5162748 2025. Available online: <https://ssrn.com/abstract=5162748> (accessed on 20 April 2025).
126. Krähmer, H.; Andreasen, C.; Economou-Antonaka, G.; Holec, J.; Kalivas, D.; Kolářová, M.; Novák, R.; Panozzo, S.; Pinke, G.; Salonen, J. Weed surveys and weed mapping in Europe: State of the art and future tasks. *Crop. Prot.* **2020**, *129*, 105010.

127. Eide, A.; Koparan, C.; Zhang, Y.; Ostlie, M.; Howatt, K.; Sun, X. UAV-assisted thermal infrared and multispectral imaging of weed canopies for glyphosate resistance detection. *Remote Sens.* **2021**, *13*, 4606.
128. Eide, A.; Zhang, Y.; Koparan, C.; Stenger, J.; Ostlie, M.; Howatt, K.; Bajwa, S.; Sun, X. Image based thermal sensing for glyphosate resistant weed identification in greenhouse conditions. *Comput. Electron. Agric.* **2021**, *188*, 106348.
129. Yang, Z.; Xia, W.; Chu, H.; Su, W.; Wang, R.; Wang, H. A comprehensive review of deep learning applications in cotton industry: From field monitoring to smart processing. *Plants* **2025**, *14*, 1481.
130. Xia, F.; Lou, Z.; Sun, D.; Li, H.; Quan, L. Weed resistance assessment through airborne multimodal data fusion and deep learning: A novel approach towards sustainable agriculture. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *120*, 103352.
131. Xu, K.; Xie, Q.; Zhu, Y.; Cao, W.; Ni, J. Effective Multi-Species weed detection in complex wheat fields using Multi-Modal and Multi-View image fusion. *Comput. Electron. Agric.* **2025**, *230*, 109924.
132. Bechar, A.; Vigneault, C. Agricultural robots for field operations: Concepts and components. *Biosyst. Eng.* **2016**, *149*, 94–111.
133. Teng, H.; Wang, Y.; Song, X.; Karydis, K. Multimodal dataset for localization, mapping and crop monitoring in citrus tree farms. In Proceeding of the International Symposium on Visual Computing, Lake Tahoe, NV, USA, 16–18 October 2023; pp. 571–582.
134. Man, Z.; Yuhan, J.I.; Shichao, L.I.; Ruyue, C.; Hongzhen, X.U.; Zhenqian, Z. Research progress of agricultural machinery navigation technology. *Nongye Jixie Xuebao/Trans. Chin. Soc. Agric. Mach.* **2020**, *51*, 4.
135. Li, A.; Cao, J.; Li, S.; Huang, Z.; Wang, J.; Liu, G. Map construction and path planning method for a mobile robot based on multi-sensor information fusion. *Appl. Sci.* **2022**, *12*, 2913.
136. Xie, B.; Jin, Y.; Faheem, M.; Gao, W.; Liu, J.; Jiang, H.; Cai, L.; Li, Y. Research progress of autonomous navigation technology for multi-agricultural scenes. *Comput. Electron. Agric.* **2023**, *211*, 107963.
137. Dong, Q.; Murakami, T.; Nakashima, Y. Recalculating the agricultural labor force in China. *China Econ. J.* **2018**, *11*, 151–169.
138. Krishnan, A.; Swarna, S. Robotics, IoT, and AI in the automation of agricultural industry: A review. In Proceeding of the 2020 IEEE Bangalore Humanitarian Technology Conference (B-HTC), Vijayapur, Karnataka, India, 8–10 October 2020; pp. 1–6.
139. Bu, L.; Hu, G.; Chen, C.; Sugirbay, A.; Chen, J. Experimental and simulation analysis of optimum picking patterns for robotic apple harvesting. *Sci. Hortic.* **2020**, *261*, 108937.
140. Ji, W.; Qian, Z.; Xu, B.; Chen, G.; Zhao, D. Apple viscoelastic complex model for bruise damage analysis in constant velocity grasping by gripper. *Comput. Electron. Agric.* **2019**, *162*, 907–920.
141. Birrell, S.; Hughes, J.; Cai, J.Y.; Iida, F. A field-tested robotic harvesting system for iceberg lettuce. *J. Field Robot.* **2020**, *37*, 225–245.
142. Silwal, A.; Davidson, J.R.; Karkee, M.; Mo, C.; Zhang, Q.; Lewis, K. Design, integration, and field evaluation of a robotic apple harvester. *J. Field Robot.* **2017**, *34*, 1140–1159.
143. Mao, S.; Li, Y.; Ma, Y.; Zhang, B.; Zhou, J.; Wang, K. Automatic cucumber recognition algorithm for harvesting robots in the natural environment using deep learning and multi-feature fusion. *Comput. Electron. Agric.* **2020**, *170*, 105254.
144. Wu, Z.; Wang, Z.; Spohrer, K.; Schock, S.; He, X.; Müller, J. Non-contact leaf wetness measurement with laser-induced light reflection and RGB imaging. *Biosyst. Eng.* **2024**, *244*, 42–52.
145. Zou, K.; Ge, L.; Zhou, H.; Zhang, C.; Li, W. Broccoli seedling pest damage degree evaluation based on machine learning combined with color and shape features. *Inf. Process. Agric.* **2021**, *8*, 505–514.
146. Ye, K.; Hu, G.; Tong, Z.; Xu, Y.; Zheng, J. Key intelligent pesticide prescription spraying technologies for the control of pests, diseases, and weeds: A review. *Agriculture* **2025**, *15*, 81.
147. Yin, D.; Chen, S.; Pei, W.; Shen, B. Design of map-based indoor variable weed spraying system. *Trans. Chin. Soc. Agric. Eng.* **2011**, *27*, 131–135.
148. Maslekar, N.V.; Kulkarni, K.P.; Chakravarthy, A.K. Application of unmanned aerial vehicles (UAVs) for pest surveillance, monitoring and management. In *Innovative Pest Management Approaches for the 21st Century: Harnessing Automated Unmanned Technologies*; Springer: Singapore, 2020; pp. 27–45.
149. Chostner, B. See & spray: The next generation of weed control. *Resour. Mag.* **2017**, *24*, 4–5.
150. Mumtaz, N.; Nazar, M. Artificial intelligence robotics in agriculture: See & spray. *J. Intell. Pervasive Soft Comput.* **2022**, *1*, 21–24.
151. Chittoor, P.K.; Dandumahanti, B.P.; Veerajagadheswar, P.; Samarakoon, S.B.P.; Muthugala, M.V.J.; Elara, M.R. Developing an urban landscape fumigation service robot: A Machine-Learned, Gen-AI-Based design trade study. *Appl. Sci.* **2025**, *15*, 2061.
152. Oberti, R.; Marchi, M.; Tirelli, P.; Calcante, A.; Iriti, M.; Tona, E.; Hočevár, M.; Baur, J.; Pfaff, J.; Schütz, C. Selective spraying of grapevines for disease control using a modular agricultural robot. *Biosyst. Eng.* **2016**, *146*, 203–215.
153. Adamides, G.; Katsanos, C.; Constantinou, I.; Christou, G.; Xenos, M.; Hadzilacos, T.; Edan, Y. Design and development of a semi-autonomous agricultural vineyard sprayer: Human–robot interaction aspects. *J. Field Robot.* **2017**, *34*, 1407–1426.

154. Lochan, K.; Khan, A.; Elsayed, I.; Suthar, B.; Seneviratne, L.; Hussain, I. Advancements in precision spraying of agricultural robots: A comprehensive Review. *IEEE Access* **2024**, *12*, 129447–129483.
155. Prathibha, S.R.; Hongal, A.; Jyothi, M.P. IoT based monitoring system in smart agriculture. In Proceeding of the 2017 International Conference on Recent Advances in Electronics and Communication Technology (ICRAECT), Bangalore, India, 16–17 March 2017; pp. 81–84.
156. Wan, S.; Zhao, K.; Lu, Z.; Li, J.; Lu, T.; Wang, H. A modularized IoT monitoring system with edge-computing for aquaponics. *Sensors* **2022**, *22*, 9260.
157. Li, L.; Li, J.; Wang, H.; Georgieva, T.; Ferentinos, K.P.; Arvanitis, K.G.; Sygrimis, N.A. Sustainable energy management of solar greenhouses using open weather data on MACQU platform. *Int. J. Agric. Biol. Eng.* **2018**, *11*, 74–82.
158. Suma, N.; Samson, S.R.; Saranya, S.; Shanmugapriya, G.; Subhashri, R. IOT based smart agriculture monitoring system. *Int. J. Recent Innov. Trends Comput. Commun.* **2017**, *5*, 177–181.
159. Boobalan, J.; Jacintha, V.; Nagarajan, J.; Thangayogesh, K.; Tamilarasu, S. An IOT based agriculture monitoring system. In Proceeding of the 2018 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 3–5 April 2018; pp. 594–598.
160. Munaganuri, R.K.; Yamarthi, N.R. PAMICRM: Improving precision agriculture through multimodal image analysis for crop water requirement estimation using multidomain remote sensing data samples. *IEEE Access* **2024**. <https://doi.org/10.1109/ACCESS.2024.3386552>.
161. Mubaiwa, O.; Chilo, D. Female genital mutilation (FGM/C) in garissa and isiolo, kenya: Impacts on education and livelihoods in the context of cultural norms and food insecurity. *Societies* **2025**, *15*, 43.
162. Li, L.; Zhang, Q.; Huang, D. A review of imaging techniques for plant phenotyping. *Sensors* **2014**, *14*, 20078–20111.
163. Gordon, W.B.; Whitney, D.A.; Raney, R.J. Nitrogen management in furrow irrigated, ridge-tilled corn. *J. Prod. Agric.* **1993**, *6*, 213–217.
164. Shafi, U.; Mumtaz, R.; García-Nieto, J.; Hassan, S.A.; Zaidi, S.A.R.; Iqbal, N. Precision agriculture techniques and practices: From considerations to applications. *Sensors* **2019**, *19*, 3796.
165. Yuan, H.; Cheng, M.; Pang, S.; Li, L.; Wang, H.; NA, S. Construction and performance experiment of integrated water and fertilization irrigation recycling system. *Trans. Chin. Soc. Agric. Eng.* **2014**, *30*, 72–78.
166. Wang, H.; Fu, Q.; Meng, F.; Mei, S.; Wang, J.; Li, L. Optimal design and experiment of fertilizer EC regulation based on subsection control algorithm of fuzzy and PI. *Trans. Chin. Soc. Agric. Eng.* **2016**, *32*, 110–116.
167. Dhakshayani, J.; Surendiran, B. M2F-Net: A deep learning-based multimodal classification with high-throughput phenotyping for identification of overabundance of fertilizers. *Agriculture* **2023**, *13*, 1238.
168. Bhattacharya, S.; Pandey, M. PCFRIMDS: Smart Next-Generation approach for precision crop and fertilizer recommendations using integrated multimodal data fusion for sustainable agriculture. *IEEE Trans. Consum. Electron.* **2024**, *70*, 6250–6261.
169. Kilinc, H.C.; Apak, S.; Ozkan, F.; Ergin, M.E.; Yurtsever, A. Multimodal Fusion of optimized GRU-LSTM with self-attention layer for Hydrological Time Series forecasting. *Water Resour. Manag.* **2024**, *38*, 6045–6062.
170. Bianchi, P.; Jakubowicz, J.; Roueff, F. Linear precoders for the detection of a Gaussian process in wireless sensors networks. *IEEE Trans. Signal. Process.* **2010**, *59*, 882–894.
171. Wang, L. Heterogeneous data and big data analytics. *Autom. Control. Inf. Sci.* **2017**, *3*, 8–15.
172. Peng, Y.; Bian, J.; Xu, J. Fedmm: Federated multi-modal learning with modality heterogeneity in computational pathology. In Proceeding of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 1696–1700.
173. Wei, Y.; Yuan, S.; Yang, R.; Shen, L.; Li, Z.; Wang, L.; Chen, M. Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. In Proceeding of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, ON, Canada, 9–14 July 2023; pp. 5240–5252.
174. Bertrand, C.; Burel, F.; Baudry, J. Spatial and temporal heterogeneity of the crop mosaic influences carabid beetles in agricultural landscapes. *Landsc. Ecol.* **2016**, *31*, 451–466.
175. Xu, L.; Jiang, J.; Du, J. The dual effects of environmental regulation and financial support for agriculture on agricultural green development: Spatial spillover effects and spatio-temporal heterogeneity. *Appl. Sci.* **2022**, *12*, 11609.
176. Dutilleul, P. *Spatio-Temporal Heterogeneity: Concepts and Analyses*; Cambridge University Press: Cambridge, UK, 2011; ISBN 0521791278.
177. Bakshi, W.J.; Shafi, M. Semantic Heterogeneity-An overview. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* **2018**, *4*, 197–200.

178. Wang, T.; Murphy, K.E. Semantic heterogeneity in multidatabase systems: A review and a proposed meta-data structure. *J. Database Manag.* **2004**, *15*, 71–87.
179. Luo, Z.; Yang, W.; Yuan, Y.; Gou, R.; Li, X. Semantic segmentation of agricultural images: A survey. *Inf. Process. Agric.* **2024**, *11*, 172–186.
180. Liu, Z.; Zhang, W.; Lin, S.; Quek, T.Q. Heterogeneous sensor data fusion by deep multimodal encoding. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 479–491.
181. Jiao, T.; Guo, C.; Feng, X.; Chen, Y.; Song, J. A comprehensive survey on deep learning Multi-Modal fusion: Methods, technologies and applications. *Comput. Mater. Contin.* **2024**, *80*, 1–35.
182. Rai, N.; Zhang, Y.; Villamil, M.; Howatt, K.; Ostlie, M.; Sun, X. Agricultural weed identification in images and videos by integrating optimized deep learning architecture on an edge computing technology. *Comput. Electron. Agric.* **2024**, *216*, 108442.
183. Ficuciello, F.; Falco, P.; Calinon, S. A brief survey on the role of dimensionality reduction in manipulation learning and control. *IEEE Robot. Autom. Lett.* **2018**, *3*, 2608–2615.
184. Munir, A.; Blasch, E.; Kwon, J.; Kong, J.; Aved, A. Artificial intelligence and data fusion at the edge. *IEEE Aerosp. Electron. Syst. Mag.* **2021**, *36*, 62–78.
185. Ajili, M.T.; Hara-Azumi, Y. Multimodal neural network acceleration on a hybrid CPU-FPGA architecture: A case study. *IEEE Access* **2022**, *10*, 9603–9617.
186. Bultmann, S.; Quenzel, J.; Behnke, S. Real-time multi-modal semantic fusion on unmanned aerial vehicles. In Proceeding of the 2021 European Conference on Mobile Robots (ECMR), Bonn, Germany, 31 August–3 September 2021; pp. 1–8.
187. Jani, Y.; Jani, A.; Prajapati, K.; Windsor, C. Leveraging multimodal ai in edge computing for real-time decision-making. *Computing* **2023**, *1*, 2.
188. Pokhrel, S.R.; Choi, J. Understand-before-talk (UBT): A semantic communication approach to 6G networks. *IEEE Trans. Veh. Technol.* **2022**, *72*, 3544–3556.
189. Yang, W.; Du, H.; Liew, Z.Q.; Lim, W.Y.B.; Xiong, Z.; Niyato, D.; Chi, X.; Shen, X.; Miao, C. Semantic communications for future internet: Fundamentals, applications, and challenges. *IEEE Commun. Surv. Tutor.* **2022**, *25*, 213–250.
190. Li, L.; Fan, Y.; Tse, M.; Lin, K. A review of applications in federated learning. *Comput. Ind. Eng.* **2020**, *149*, 106854.
191. Mammen, P.M. Federated learning: Opportunities and challenges. *arXiv* **2021**, arXiv:2101.05428.
192. Xu, X.; Ding, Y.; Hu, S.X.; Niemier, M.; Cong, J.; Hu, Y.; Shi, Y. Scaling for edge inference of deep neural networks. *Nat. Electron.* **2018**, *1*, 216–222.
193. Shuvo, M.M.H.; Islam, S.K.; Cheng, J.; Morshed, B.I. Efficient acceleration of deep learning inference on resource-constrained edge devices: A review. *Proc. IEEE Inst. Electr. Electron. Eng.* **2022**, *111*, 42–91.
194. Geng, X.; He, X.; Hu, M.; Bi, M.; Teng, X.; Wu, C. Multi-attention network with redundant information filtering for multi-horizon forecasting in multivariate time series. *Expert Syst. Appl.* **2024**, *257*, 125062.
195. Dayal, A.; Bonthu, S.; Saripalle, P.; Mohan, R. Deep learning for multi-horizon water level forecasting in KRS reservoir, India. *Results Eng.* **2024**, *21*, 101828.
196. Kaur, A.; Goyal, P.; Rajhans, R.; Agarwal, L.; Goyal, N. Fusion of multivariate time series meteorological and static soil data for multistage crop yield prediction using multi-head self attention network. *Expert Syst. Appl.* **2023**, *226*, 120098.
197. Deforce, B.; Baesens, B.; Diels, J.; Serral Asensio, E. Forecasting sensor-data in smart agriculture with temporal fusion transformers. *Trans. Comput. Sci. Comput. Intell.* **2022**. Available online: <https://lirias.kuleuven.be/retrieve/668581> (accessed on 20 April 2025).
198. Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv* **2016**, arXiv:1611.03530.
199. Jakubovitz, D.; Giryas, R.; Rodrigues, M.R. Generalization error in deep learning. In Proceeding of the Compressed Sensing and its Applications: Third International Matheon Conference, Berlin, Germany, 4–8 December 2017; Birkhäuser: Basel, Switzerland, 2019; pp. 153–193.
200. Ahmad, A.; El Gamal, A.; Saraswat, D. Toward generalization of deep learning-based plant disease identification under controlled and field conditions. *IEEE Access* **2023**, *11*, 9042–9057.
201. Wang, D.; Cao, W.; Zhang, F.; Li, Z.; Xu, S.; Wu, X. A review of deep learning in multiscale agricultural sensing. *Remote Sens.* **2022**, *14*, 559.
202. Fuentes, A.; Yoon, S.; Kim, T.; Park, D.S. Open set self and across domain adaptation for tomato disease recognition with deep learning techniques. *Front. Plant. Sci.* **2021**, *12*, 758027.

203. Qi, F.; Yang, X.; Xu, C. A unified framework for multimodal domain adaptation. In Proceeding of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 429–437.
204. Singhal, P.; Walambe, R.; Ramanna, S.; Kotecha, K. Domain adaptation: Challenges, methods, datasets, and applications. *IEEE Access* **2023**, *11*, 6973–7020.
205. Shorten, C.; Khoshgoftaar, T.M.; Furht, B. Text data augmentation for deep learning. *J. Big Data* **2021**, *8*, 101.
206. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 1–48.
207. Alijani, S.; Fayyad, J.; Najjaran, H. Vision transformers in domain adaptation and domain generalization: A study of robustness. *Neural Comput. Appl.* **2024**, *36*, 17979–18007.
208. Wilson, G.; Cook, D.J. A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol.* **2020**, *11*, 1–46.
209. Oubara, A.; Wu, F.; Amamra, A.; Yang, G. Survey on remote sensing data augmentation: Advances, challenges, and future perspectives. In Proceeding of the International Conference on Computing Systems and Applications, Algiers, Algeria, 17–18 May 2022; pp. 95–104.
210. Mao, H.H. A survey on self-supervised pre-training for sequential transfer learning in neural networks. *arXiv* **2020**, arXiv:2007.00800.
211. Berg, P.; Pham, M.; Courty, N. Self-supervised learning for scene classification in remote sensing: Current state of the art and perspectives. *Remote Sens.* **2022**, *14*, 3995.
212. Andrychowicz, M.; Denil, M.; Gomez, S.; Hoffman, M.W.; Pfau, D.; Schaul, T.; Shillingford, B.; De Freitas, N. Learning to learn by gradient descent by gradient descent. *Adv. Neural Inf. Process. Syst.* **2016**, *29*. Available online: <https://proceedings.neurips.cc/paper/2016/hash/fb87582825f9d28a8d42c5e5e5e8b23d-Abstract.html> (accessed on 20 April 2025).
213. Wang, J.X. Meta-learning in natural and artificial intelligence. *Curr. Opin. Behav. Sci.* **2021**, *38*, 90–95.
214. De Andrade Porto, J.V.; Dorsa, A.C.; de Moraes Weber, V.A.; de Andrade Porto, K.R.; Pistori, H. Usage of few-shot learning and meta-learning in agriculture: A literature review. *Smart Agric. Technol.* **2023**, *5*, 100307.
215. Tseng, G.; Kerner, H.; Rolnick, D. TIML: Task-informed meta-learning for agriculture. *arXiv* **2022**, arXiv:2202.02124.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.