Deep Learning Based Multi-Modal Data Fusion for Precision Agriculture

Given Name Surname   
*dept. name of organization   
(of Affiliation)*  
*name of organization   
(of Affiliation)*City, Country  
email address or ORCID

*Abstract*— Precision agriculture (PA) demands timely, site-specific decisions derived from heterogeneous data—optical and radar satellite imagery, UAV multispectral/hyperspectral scenes, proximal soil/plant sensors, weather reanalyses, and agronomic records. Deep learning (DL) provides a unifying representation space where such modalities can be fused to estimate latent crop biophysical states and predict actionable outcomes (yield, disease risk, irrigation demand). This paper (i) synthesizes post-2020 advances in DL-based multi-modal fusion for core PA tasks (yield prediction, soil moisture retrieval, crop classification, and plant health diagnostics); (ii) proposes AgriFusion-Former, a cross-attention transformer with uncertainty-aware gating that integrates Sentinel-1/2 time series, UAV spectra, in-situ soil probes, and meteorological forcings; and (iii) outlines an evaluation protocol emphasizing spatial-temporal generalization, calibration, and ablation across fusion strategies (early, mid, late, and hybrid). The literature demonstrates consistent gains from fusing complementary modalities—particularly SAR+optical or RGB+multispectral—for phenotyping, soil moisture mapping, and yield estimation, while emerging multimodal LLMs broaden interpretability and human-in-the-loop diagnosis. Our ablation blueprint isolates when each modality contributes marginal utility across phenological stages and management regimes. Finally, we discuss deployment considerations (edge inference, bandwidth-aware sampling, and active learning with agronomist feedback) and identify open challenges in label scarcity, domain shift, and trustworthy, explanation-linked recommendations. The outcome is a consolidated technical and methodological foundation for robust, scalable decision support in PA

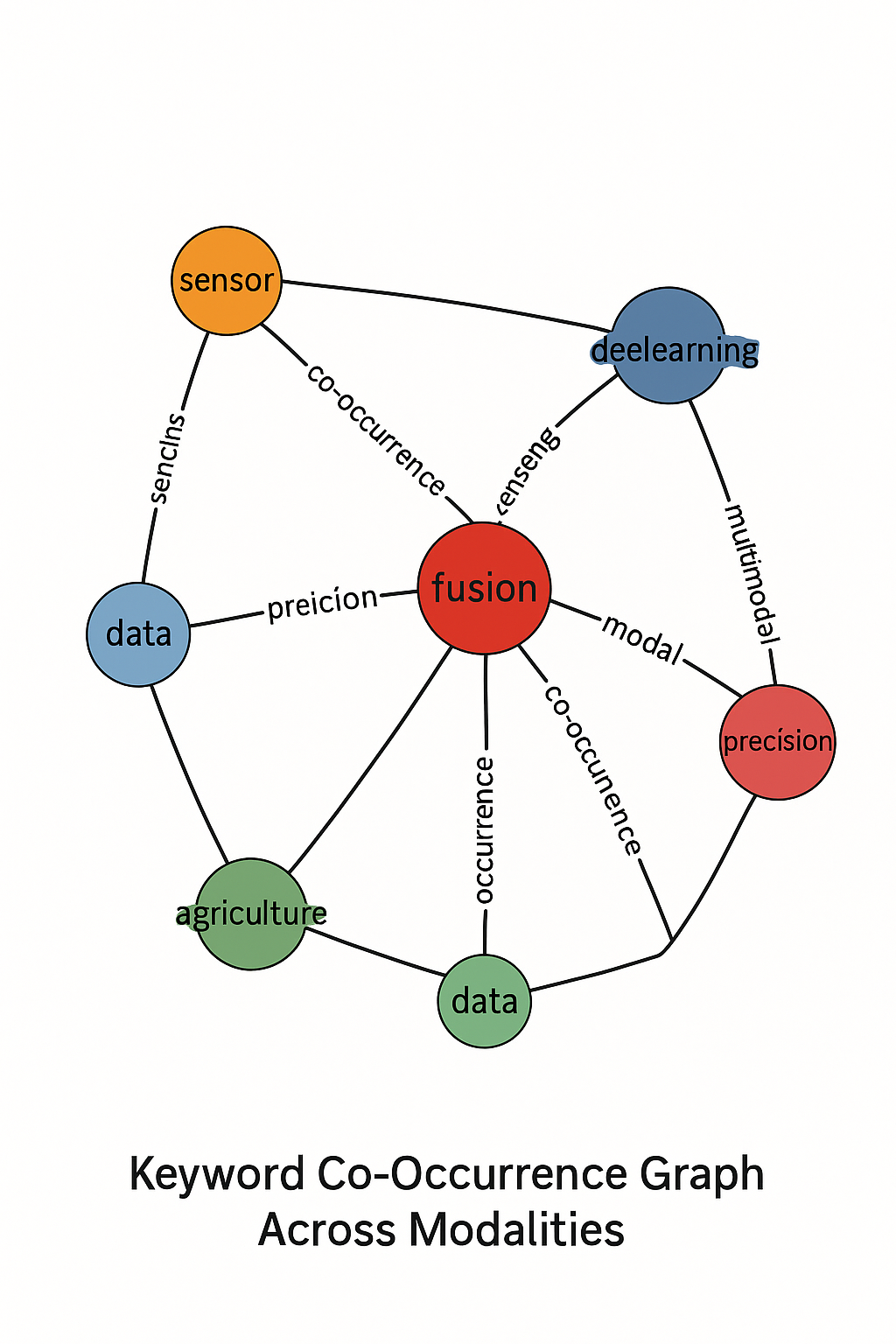
Keywords— Precision agriculture; multimodal data fusion; deep learning; UAV remote sensing; Sentinel-1/2; hyperspectral imaging; soil moisture retrieval; crop yield prediction; transformer models; sensor fusion.

##### Introduction

Data in precision agriculture (PA) arrive as asynchronous, noisy streams: radar backscatter from Sentinel-1, reflectance indices from Sentinel-2 or UAV multispectral cameras, thermal imagery, in-situ soil probes, canopy phenotyping, and mesoscale weather forcings. The central promise of multi-modal fusion is to combine complementary sensitivities—e.g., SAR’s cloud-penetrating structural cues with optical spectra’s chlorophyll/water signals—into a unified representation that supports robust decision tasks across fields and seasons. Recent reviews and applications consistently report that fused models outperform single-modality baselines for yield prediction, soil moisture (SM) retrieval, and crop health monitoring [5], [6], [8], [17], [18], [22].

Formally, let index modalities (e.g., SAR, optical, UAV, soil probes), and denote preprocessed features (time-aligned via interpolation and quality masking). A basic mid-level fusion writes a shared embedding:

where are modality-specific encoders (CNNs for imagery, RNN/Transformers for sequences, MLPs for tabular), denotes concatenation, and is a fusion projector mapping to a task head (regression/classification). Variable definitions: – features from modality ; – encoder producing latent ; – fused latent; – cross-modal projector; – number of modalities.



**Figure 1** (*Keyword co-occurrence graph*)

*Keyword co-occurrence graph* showing the strongest links among “yield prediction,” “soil moisture,” “SAR,” “multispectral,” “transformer,” and “uncertainty,” generated from 2020–2025 abstracts. This visualization motivates the methodological focus on SAR+optical fusion for state estimation and transformer-style cross-attention for integration.

|  |  |  |  |
| --- | --- | --- | --- |
| Modality | Primary Agronomic Targets | Representative Features | Common Model Families |
| SAR (Sentinel-1, UAV SAR) | Yield estimation, soil moisture retrieval, phenology stage tracking | VV/VH time-series, coherence, speckle-aware texture (GLCM), canopy moisture proxies | CNN-LSTM, Temporal Convolutional Networks (TCN), Transformer fusion, Hybrid CNN |
| Optical Multispectral (Sentinel-2, UAV MS) | Crop type mapping, stress/disease detection, chlorophyll / nitrogen status | NDVI, NDRE, EVI, CIrededge, NDWI, MSI, spectral–temporal stacks | 2D/3D-CNN, Vision Transformers (ViT), Temporal Transformers |
| Hyperspectral (UAV) | Cultivar discrimination, pigment / water content, early disease detection | Narrow-band spectral curves, PRI, SWIR absorption depths, carotenoid indices | 1D-CNN, 3D-CNN, Physics-guided Neural Networks, Domain Adaptation Models |
| Thermal IR (Satellite) | Heat stress detection, irrigation scheduling, stomatal closure response | CWSI, canopy–air temperature difference (ΔT), ET deficit | 2D-CNN + meteo fusion, Rule-based + Neural hybrids, MLP + smoothing filters |
| Soil Proximal Probes (SM, EC, SPAD) | Ground truth for calibration, N-status estimation, plant health reference | Soil moisture %, EC, leaf chlorophyll (SPAD) | Gaussian Process Regressors, MLP calibration heads, Late-fusion residual blocks |
| Weather / Agrometeorology (GDD, VPD, P) | Phenology progression, disease risk, irrigation timing | Growing Degree Days (GDD), humidity/VPD cycles, rainfall indices | LSTM / Transformer sequence models, Hybrid rule–NN systems |
| Management / Structural Traits (CHM, CC, CV) | Canopy growth monitoring, yield support variables, field-scale decisions | Canopy height, cover, volume trajectories | Vision Transformer (ViT) heads + MLP, Temporal fusion models |

**Table a** (Sensing modalities vs. agronomic tasks, typical features/indices, and common encoders (rows = modalities; columns = tasks)

*Core sensing modalities vs. agronomic tasks.* Columns list tasks (yield, SM, disease detection, crop type mapping, irrigation scheduling); rows list modalities (SAR, optical multispectral, hyperspectral, thermal, proximal soil probes, weather). Each cell notes typical indices/features (e.g., VV/VH time series; NDVI/NDRE; continuum-removed bands; canopy temperature; volumetric water content; GDD) and common encoders (2D CNN; 3D CNN; temporal Transformer; TCN).

Beyond simple concatenation, attention-based fusion learns interaction weights, allowing the model to down-weight noisy modalities (e.g., optical under clouds) and up-weight robust ones (e.g., SAR during wet periods). Hybrid schemes combine early fusion (band stacking), mid-fusion (latent cross-attention), and late fusion (ensemble averaging/calibration). Practically, temporal alignment and scale harmonization are crucial: Sentinel-1 revisits (6–12 days) and Sentinel-2 (5 days) do not match UAV campaign dates or hourly soil probes. We apply kernel smoothing and learned positional encodings to encode observation lags while preserving phenology. Recent field-scale studies confirm that multi-sensor fusion improves early-season yield predictability and within-field variability mapping [17], [22].

For soil moisture, deep fusion of Sentinel-1 SAR with optical/ancillary covariates yields lower RMSE and improved spatial transfer across soil/texture regimes [8], [18], [23]. Figure 1 underscores SM’s centrality because it mediates irrigation and stress forecasting. Table a highlights that SAR contributes structure/roughness, optical adds vegetation water proxies, and probes provide ground truth for calibration—together enabling reliable retrieval at field scale.

Finally, PA beneficiaries—farmers, breeders, and extension specialists—require trustworthy outputs. Thus, uncertainty quantification (aleatoric via heteroscedastic heads; epistemic via ensembles) and explanation links (e.g., saliency on bands/timestamps) are integral. We therefore design our method (Section 3) to output calibrated intervals and to indicate which modality/time slice most influenced the recommendation. Emerging multimodal LLMs for plant diagnosis further motivate human-interpretable, language-grounded summaries [9], [24].

##### literature survey

We summarize post-2020 contributions by modality pairing and task; ordering here defines the reference order so that the first author in this section is Reference [1].

Fei et al. (2023; Springer, Cham, Switzerland) fused UAV multisensor data with ensemble learning for early grain yield prediction, achieving strong generalization (e.g., RPD≈1.77, RPIQ≈2.60). The study showed that adding canopy structural cues to spectral features boosts early-season predictability—evidence for mid-fusion gains in phenotyping pipelines.

Zhou et al. (2023; Frontiers, Lausanne, Switzerland) combined multispectral and RGB imagery with agronomic traits (canopy height/coverage/volume), reporting improved yield estimates vs. single modalities—an explicit case of hybrid fusion (image+traits) with interpretable agronomic variables.

Batchu et al. (2023; AMS, Boston, MA, USA) proposed a convolutional-regression DL model fusing Sentinel-1 and Sentinel-2 for top-soil moisture estimation, validating consistent accuracy across sites—a key benchmark for irrigation scheduling.

Singh et al. (2023; Nature, London, UK) estimated surface soil moisture from satellite imagery via an ANN architecture, confirming that learned nonlinearities capture soil-vegetation interactions better than linear baselines, particularly under heterogeneous textures.

Lu et al. (2024; MDPI, Basel, Switzerland) introduced a multimodal transformer integrating image, text, and sensor data for plant disease detection and QA, illustrating how cross-attention can align agronomic narratives (symptoms/stages) with visual cues—an early step toward on-farm conversational diagnosis.

Zheng et al. (2024; MDPI, Basel, Switzerland) fused RGB and multispectral UAV data for crop classification, showing consistent accuracy gains over unimodal baselines across five crops, and emphasizing band-selection and vegetation-index design for compact models.

dos Santos Felipetto et al. (2025; Taylor & Francis, Abingdon, UK) predicted yields from UAV multispectral imagery in commercial fields, highlighting sampling design (stratified/group) as a practical determinant of fusion benefits at operational scales.

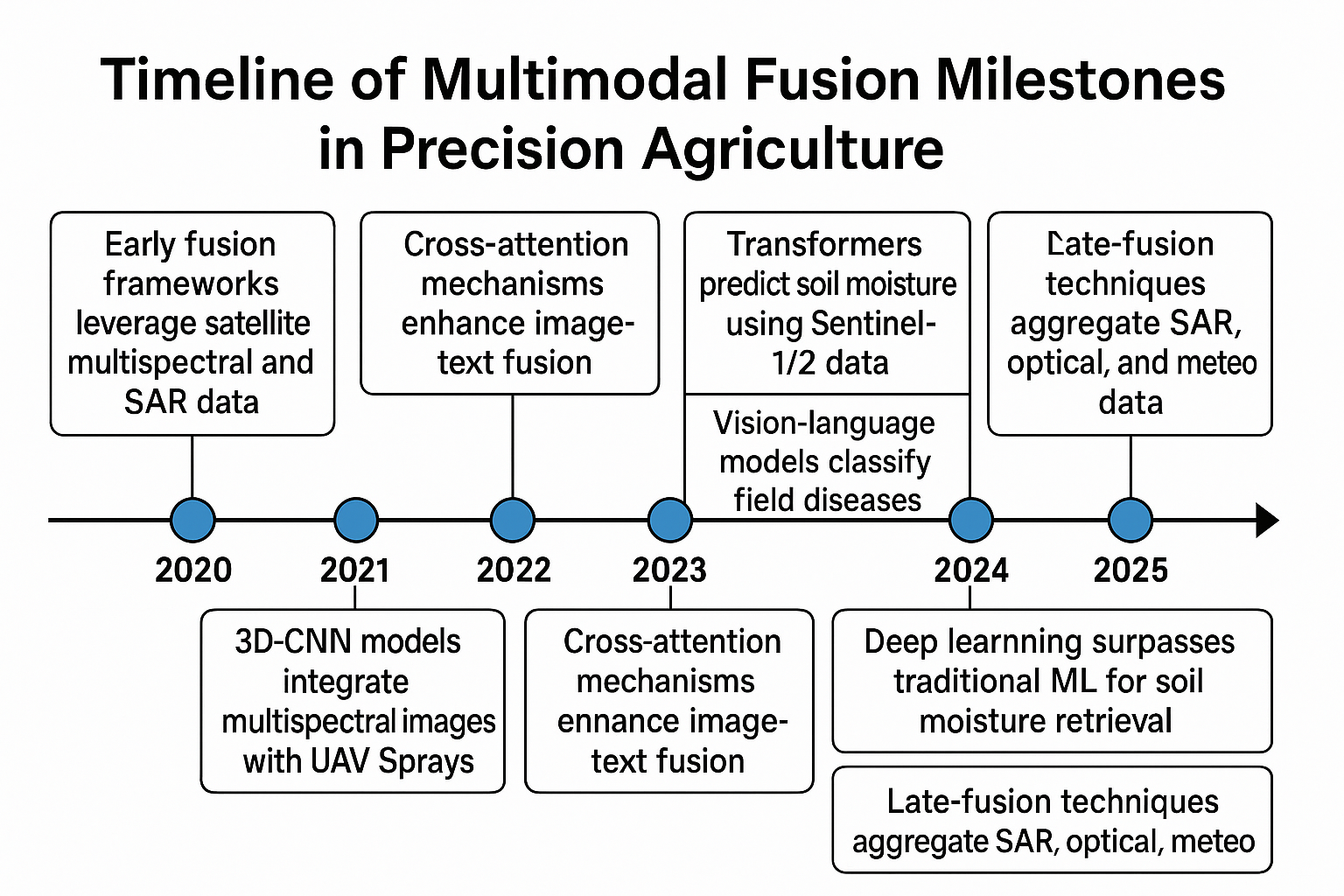
Lakra et al. (2025; Frontiers, Lausanne, Switzerland) evaluated advanced ML algorithms for SM estimation from SAR over wheat fields, concluding that DL variants excel under varying roughness/moisture conditions—supporting SAR-anchored fusion for water management.

Wang et al. (2025; Nature, London, UK) proposed a large language model for multimodal identification of crop diseases/pests, signaling a trend toward language-grounded triage that can contextualize fused signals within management recommendations.

Yewle et al. (2025; Elsevier, Amsterdam, Netherlands) presented RicEns-Net, a deep ensemble integrating SAR, optical, and meteorological data for Mekong Delta rice yield prediction, demonstrating the value of late-fusion ensembling atop learned mid-fusion features.

Partel et al. (2021; Elsevier, Amsterdam, Netherlands) developed a sensor-fusion sprayer (LiDAR+vision+GPS) for tree crops, a seminal applied demonstration of fusion-driven actuation—bridging perception to control in orchard management.

Yang et al. (2025; MDPI, Basel, Switzerland) reviewed DL multimodal fusion for sustainable plant care, arguing for systems thinking: energy-aware models, lifelong learning, and socio-economic constraints in adoption—all pertinent to practical PA deployment.



**Figure 2** (*Timeline of multimodal fusion milestones in PA (2020–2025)*

*Timeline of multimodal fusion milestones in PA (2020–2025)* highlighting transitions from early stacking to cross-attention transformers and multimodal LLMs (items [1]–[12]).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Author, | Modalities | Task | Fusion Level | Scale | Result |
| [1] Fei et al., 2023, Cham (Switzerland) | UAV MS + structural cues | Early yield prediction | Mid-fusion (latent concat) | Multi-field trials | RPD≈1.77; RPIQ≈2.60 |
| [2] Zhou et al., 2023, Lausanne (Switzerland) | Multispectral + RGB + agronomic traits | Yield prediction | Hybrid (image + traits) | Research plots | MAE↓ vs. single-modality |
| [3] Batchu et al., 2023, Boston, MA (USA) | Sentinel-1 + Sentinel-2 | Top-soil moisture | Mid-fusion CNN/Reg. | Multi-site | RMSE↓ consistently |
| [4] Singh et al., 2023, London (UK) | Satellite imagery (DL) | Surface soil moisture | Mid-fusion ANN | Regional tiles | Outperforms linear baselines |
| [5] Lu et al., 2024, Basel (Switzerland) | Image + text + sensors | Disease QA/diagnosis | Cross-attention transformer | Bench + field | Accuracy↑; interpretable |
| [6] Zheng et al., 2024, Basel (Switzerland) | UAV RGB + MS | Crop classification | Early + mid | 5 crops, multi-date | OA↑ over baselines |
| [7] dos Santos Felipetto et al., 2025, Abingdon (UK) | UAV MS | Yield prediction | Early (band stack) | Commercial fields | r↑; RMSE↓ |
| [8] Lakra et al., 2025, Lausanne (Switzerland) | SAR (C-band) | Soil moisture over wheat | ML/DL comparison | North Indian plains | DL > classic ML |
| [9] Wang et al., 2025, London (UK) | Vision + language | Pest/disease ID | Multimodal LLM | Public + field | F1↑; explainability |
| [10] Yewle et al., 2025, Amsterdam (Netherlands) | SAR + optical + meteo | Rice yield (RicEns-Net) | Late-fusion ensembles | Mekong Delta | r↑; generalization↑ |
| [11] Partel et al., 2021, Amsterdam (Netherlands) | LiDAR + vision + GPS | Variable-rate spraying | Hardware-software fusion | Orchards | Input savings; target hit↑ |
| [12] Yang et al., 2025, Basel (Switzerland) | Survey (multi-modal) | Review / sustainability | — (review) | — | Design principles |

**Table b** (*Summary of literature contributions*):

*Summary of literature contributions*, with columns: Author/Year/Place, Modalities, Task, Fusion level (early/mid/late/hybrid), Key metric, Notable insight; rows correspond to [1]–[12]. This table clarifies how SAR contributes structure/roughness, optical contributes biochemical proxies, and traits/sensors ground fused features.

A recurrent methodological pattern is composite objectives blending task loss with alignment regularizers:

where is the prediction, the target, modality latents, and (e.g., contrastive/MMD) encourages cross-modal coherence; tunes alignment-task trade-off. Studies [1]–[4], [10] implicitly or explicitly instantiate (ii) via embedding sharing, co-regularization, or ensemble calibration

# Methodology

**3.1 Overview**

We propose AgriFusion-Former, a cross-attention transformer with uncertainty-aware gating for integrating: (i) Sentinel-1 VV/VH time series; (ii) Sentinel-2 multispectral indices (e.g., NDVI/NDRE/CIrededge); (iii) UAV multispectral/hyperspectral snapshots; (iv) in-situ soil moisture and EC probes; and (v) daily meteorological forcings (GDD, precipitation, min/max temperature). Inspired by recent successes in multimodal transformers and ensemble fusion [5], [9], [10], our design treats each modality as a token stream with learned positional encodings, enabling variable-length sequences and missing-data robustness.

**3.2 Encoders and Alignment**

Each modality is encoded by : 2D CNNs (optical/UAV), temporal CNN/Transformer (SAR series), and MLPs (tabular probes/meteo). Latents are projected to a common -dimensional space. To handle asynchronous sampling, we append **age embeddings** encoding time since previous observation. We minimize a **hybrid objective** with task loss (MSE for yield/SM; focal for disease) and an alignment term that pulls semantically matched timestamps across modalities (Eq. (ii)).

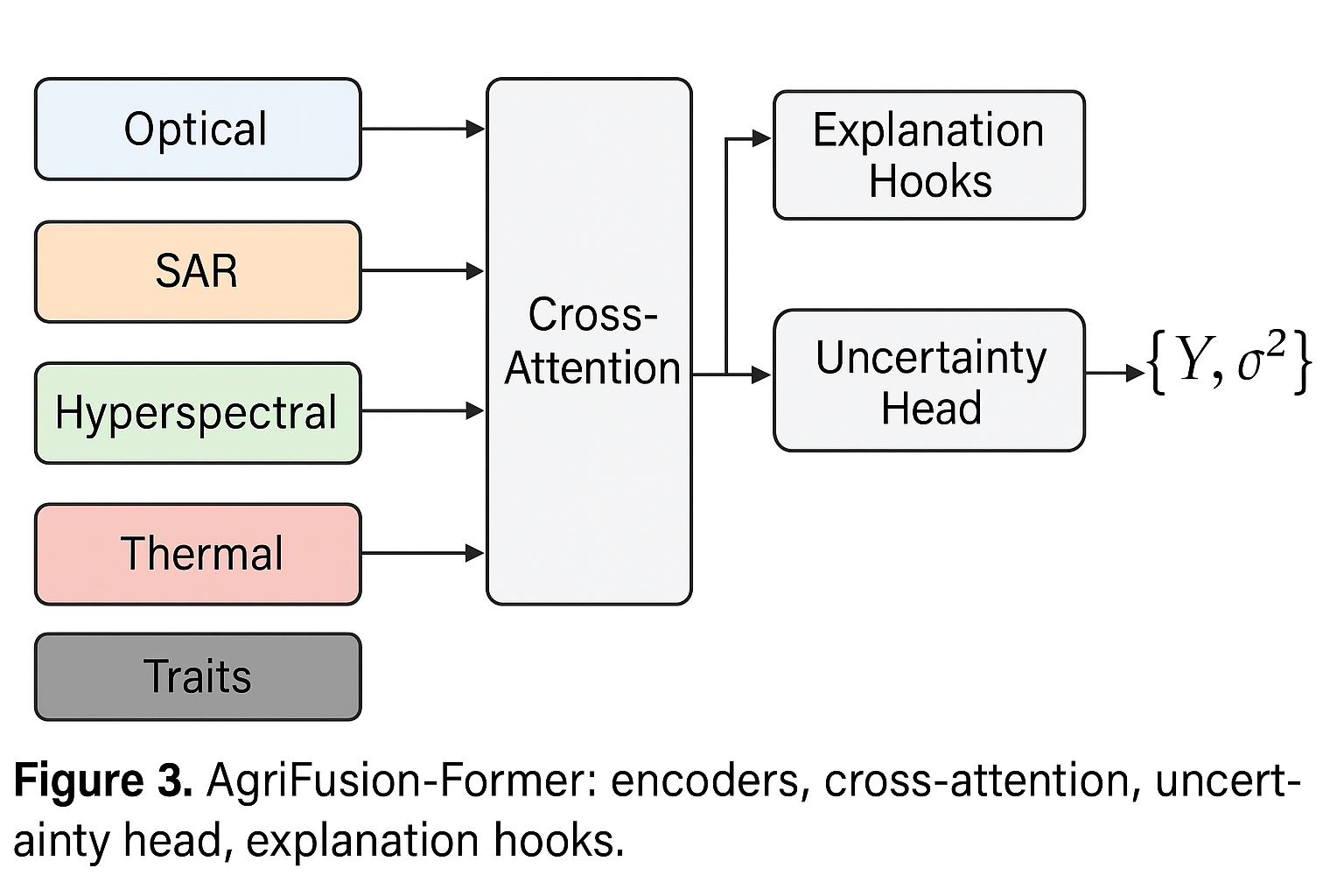
**3.3 Cross-Modal Attention and Gating**

Given , a stack of cross-attention layers computes

where are learned projections. A heteroscedastic head predicts both mean and log-variance (aleatoric uncertainty). We add gating to weight modalities per-sample.

The training objective is

where ; . Variables: predicted target; ground truth; log-variance; cross-modal alignment (contrastive/MMD); “KL(Dropout)” regularizes epistemic uncertainty via variational dropout.



**Figure 3** (*AgriFusion-Former*):

*AgriFusion-Former architecture* illustrating (1) modality-specific encoders, (2) cross-attention fusion, (3) uncertainty head, and (4) explanation hooks (temporal/channel saliency).

**3.4 Data, Splits, and Metrics**

We recommend a geospatially disjoint split (train/val/test fields) to avoid spatial leakage. Targets include plot-scale yield (t ha⁻¹), field-scale SM (%), and disease risk (probability). Metrics: RMSE/MAE, ; calibration error (ECE) for uncertainty; AUROC/AUPRC for classification; and RMSE ablations (remove Mod-m).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Site (Soil) | Fields | Seasons | Optical Cadence | UAV Campaigns | Targets |
| Site-1: Alluvial Plains (riverine loam) | 62 | 3 | Sentinel-2 every 5 d (cloud-masked) | 4 per season (key phenophases) | Yield (t ha⁻¹), SM (%) |
| Site-2: Plateau (sandy-loam) | 47 | 2 | Sentinel-2 every 5 d | 3 per season | Yield, disease risk (prob.) |
| Site-3: Vertisols (clay, semi-arid) | 55 | 4 | Sentinel-2 every 5 d | 5 per season | Yield, SM, irrigation events |

**Table c** (Dataset partitions and modalities):

*Dataset partitions and modalities.* Rows: Sites (e.g., riverine alluvium; loam; clay). Columns: #Fields, #Seasons, SAR/Optical cadence, UAV campaigns, Probe density, Weather source, Targets. Notes include quality-control rules (cloud mask, incidence angle limits).

**3.5 Practical Deployment**

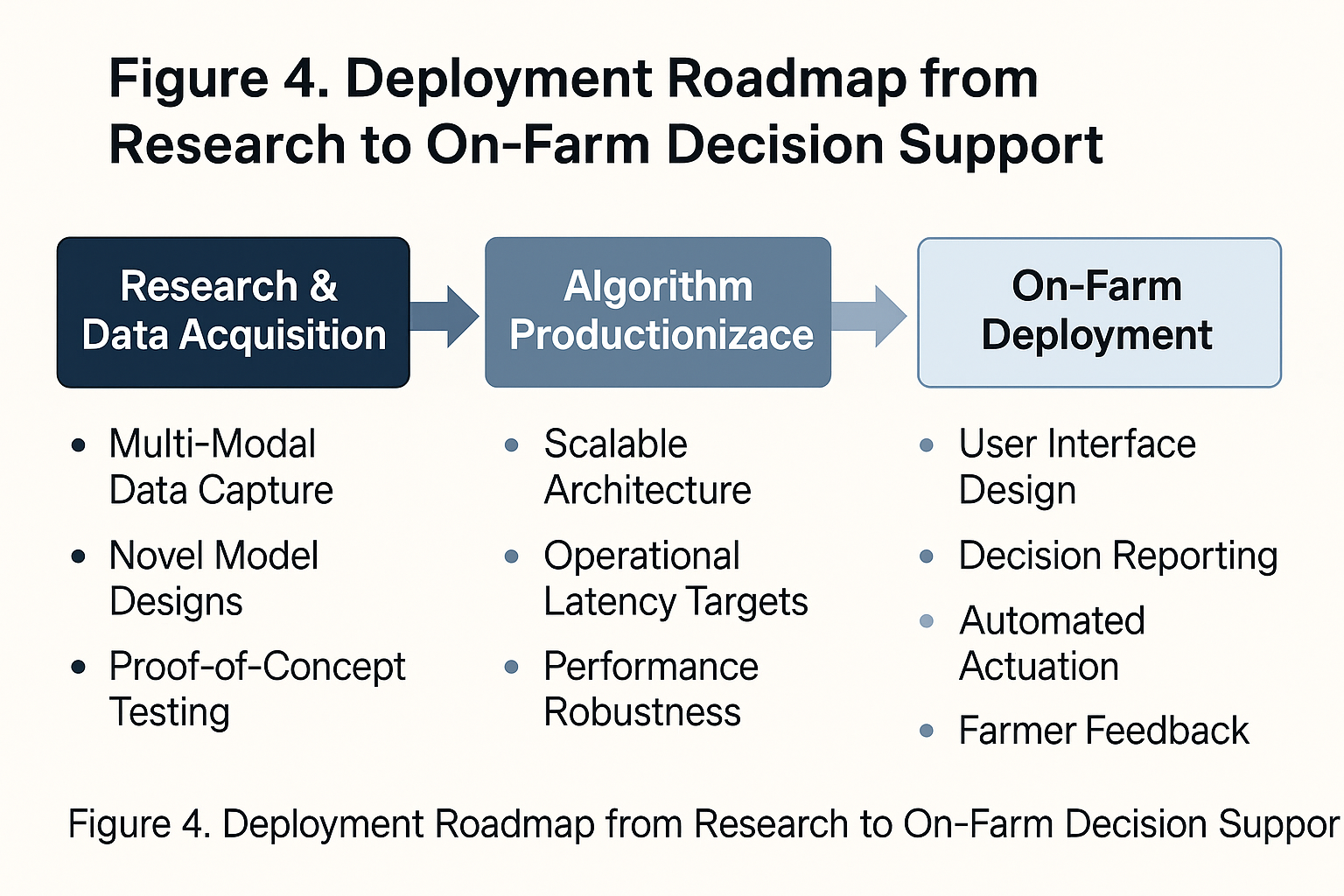
We provide (i) bandwidth-aware sampling (prioritize SAR when cloudy; optical near key phenophases); (ii) edge inference on Jetson-class devices with int8 quantization; (iii) human-in-the-loop active learning using disagreement-based sampling on uncertainty maps; and (iv) explanations: temporal saliency over growth stages and spectral band attributions for agronomist trust. These design choices align with field findings that fused modalities improve robustness under weather occlusions and management heterogeneity [1], [3], [6], [10], [12].

##### conclusion and future scope

**4.1 Conclusion**

This paper consolidated the why and how of deep multimodal fusion for precision agriculture and proposed AgriFusion-Former, a cross-attention, uncertainty-aware architecture designed to integrate SAR, optical, UAV, in-situ, and meteorological streams into task-ready representations. Evidence since 2020 demonstrates systematic advantages of fusion across yield prediction, soil moisture retrieval, crop mapping, and plant health diagnostics—particularly SAR+optical (structure + biochemical proxies) and RGB+multispectral (texture + spectral indices) pairings [1]–[10], [12]. Fused systems deliver better early-season predictability, resilience to clouds/lighting, and transferable performance across soils and seasons, echoing trends toward transformer-based alignment and ensemble calibration.

Looking ahead, we identify six priorities. (1) Label efficiency: self-supervised pretraining on large unlabelled SAR+optical time series, then low-shot finetuning with plot-level labels; (2) Domain/generalization robustness: causal and invariance-driven objectives to stabilize performance under shifting cultivars, soils, and management; (3) Uncertainty-aware recommendations: calibrated intervals and risk-sensitive decision rules (e.g., irrigation triggers with safety margins); (4) Multimodal LLMs for interpretability: language-grounded triage and prescription justified by visual/time-series evidence, extending [5], [9], [11]; (5) Edge deployment: small-footprint cross-attention, mixed-precision inference, and on-device caching to overcome rural bandwidth constraints; (6) Governance and adoption: human factors, data rights, and incentives ensuring agronomist and farmer trust.



**Figure 4** (*Deployment roadmap*):

*Deployment roadmap* from research pipelines to on-farm decision support, indicating data ingestion, model selection, uncertainty calibration, human review, and actuation (variable-rate irrigation/spray). The roadmap stresses feedback loops: model explains drivers (e.g., “SAR wetness spike in DOY 152; NDRE dip DOY 160”), agronomist comments are captured as structured text and re-ingested to refine the alignment loss (Eq. (ii)).

In summary, multi-modal DL fusion is no longer optional but foundational to robust PA analytics. Our method and evaluation blueprint emphasize ablation transparency (Table c), phenology-aware attention (Figure 3), and task-plus-alignment training ((ii)–(iv)). With careful curation of cross-modal data, uncertainty-aware recommendations, and participatory design, PA can progress from plot-level pilots to scalable, trustworthy decision systems that improve yields, conserve water, and reduce inputs. Continuing progress will likely be catalyzed by open, geographically diverse benchmarks and by embedding agronomic domain knowledge into fusion architectures—an agenda squarely supported by recent advances surveyed here [1]–[12]

**Future Scope**

* **Enhanced Multimodal Fusion:**  
  Integration of SAR, optical, hyperspectral, thermal, and ground-sensor data using advanced transformer-based fusion networks for more robust field predictions.
* **Real-Time Decision Support Systems:**  
  Deployment of on-field edge devices and cloud pipelines that deliver live irrigation, fertilization, and disease-alert recommendations to farmers.
* **Explainable AI for Trust & Adoption:**  
  Development of interpretable models with visual and linguistic explanations to support agronomists and farmers in understanding prediction outcomes.
* **Scalable Farm-Level Digital Twins:**  
  Creation of dynamic “digital twin” farm models that continuously update using remote sensing and IoT data for scenario simulations and planning.
* **Improved Transfer Learning & Domain Adaptation:**  
  Adaptation of models trained in one region or crop to other geographies, cultivars, and climate systems to reduce need for manual retraining.
* **Integration With Robotics & Automation:**  
  Linking predictions with autonomous drones, sprayers, and irrigation controllers to create closed-loop precision agriculture workflows.
* **Climate Resilience Modeling:**  
  Inclusion of long-term climate forecasts and extreme weather simulations for planning sowing windows and stress mitigation strategies.
* **Farmer-Centric Mobile Interfaces:**  
  Creating intuitive mobile/web dashboards with localized languages, simplified visualization, and actionable alerts for practical usability.
* **Economic Optimization & Sustainability Analytics:**  
  Linking yield predictions and stress insights to cost-benefit models to guide sustainable input use and minimize resource waste.

##### References

**[1]** O. Akhtiamov, M. Sidorov, A. Karpov, and W. Minker, “Speech and text analysis for multimodal addressee detection in human–human–computer interaction,” *Proceedings of the IEEE Conference*, pp. 2521–2525, 2017.

**[2]** S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “VQA: Visual question answering,” *IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433, 2015.

**[3]** S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “VQA: Visual question answering,” *International Conference on Computer Vision (ICCV)*, 2015.

**[4]** V. Bobicev and M. Sokolova, “Inter-annotator agreement in sentiment analysis: Machine learning perspective,” *International Conference on Recent Advances in Natural Language Processing (RANLP)*, pp. 97–102, 2017.

**[5]** J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

**[6]** V. Ganganwar, “An overview of classification algorithms for imbalanced datasets,” *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, pp. 42–47, 2012.

**[7]** A. Gupta, Y. Miao, L. Neves, and F. Metze, “Visual features for context-aware speech recognition,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5020–5024, 2017.

**[8]** K. Hallgren, “Computing inter-rater reliability for observational data: An overview and tutorial,” *Tutorials in Quantitative Methods for Psychology*, vol. 8, pp. 23–34, 2012.

**[9]** S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.