Highlights

- A large-scale multispectral UAV dataset for sequential radish and cabbage growth stages.
- Hierarchical framework (CropMap) using tree-structured label constraints for phenology.
- Green, near-infrared, red edge combination for addressing subtle physiological shifts.
- CropMap outperformed state-of-the-art CNN and transformer baselines.
- Robust canopy-background discrimination in complex, real-world field scenarios.

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

Field-scale crop growth stage mapping using multispectral images and deep hierarchical segmentation

L. Minh Dang [a,b,c], Sufyan Danish [d], Kyungbok Min [d], Gul E. Arzu [d], Lilia Tightiz [d], Han Yong Park [e], Hyoung-Kyu Song [c], Hyeonjoon Moon [d,*]

[a] Institute of Research and Development, Duy Tan University, Da Nang, 550000, Vietnam
[b] Faculty of Information Technology, Duy Tan University, Da Nang, 550000, Vietnam
[c] Department of Information and Communication Engineering and Convergence Engineering for Intelligent Drone, Sejong University, Seoul, 05006, South Korea
[d] Department of Computer Science and Engineering, Sejong University, Seoul, 05006, South Korea
[e] Department of Bioresource Engineering, Sejong University, Seoul, 05006, Republic of Korea

ARTICLE INFO

Keywords:
Hierarchical segmentation
Remote sensing
Growth stage mapping
Multispectral imaging
Precision agriculture

ABSTRACT

Accurate, field-scale mapping of crop growth stages is critical for supply-sensitive vegetable production, where timely harvests require detailed phenological information. Consecutive growth stages often involve rapid and subtle morphological changes and are influenced by challenging open-field conditions, which frequently result in misclassification when stages are treated as independent, discrete categories. To address this issue, CropMap is proposed as a growth-stage mapping framework that integrates Hierarchical Semantic Segmentation Networks (HSSN) with multispectral unmanned aerial vehicles (UAVs) imagery. CropMap incorporates the structured biological progression of crop development into the learning objective through tree-based label constraints, allowing the model to recognize phenological continuity and reduce confusion between adjacent stages. The framework is evaluated on the publicly available National Information Society Agency of Korea (NIA) field crop growth-stage dataset, a large-scale, multi-institutional UAV dataset containing 337,665 multispectral patches across six hierarchically related growth stages of Chinese cabbage and radish, curated by the NIA. CropMap achieves a test-set mean Intersection over Union (mIoU) of 0.5382, representing a 5% relative improvement over the best-performing transformer baseline (SegFormer; mIoU = 0.5124). Performance varies across classes: background separation is strong (IoU = 0.9128) and the rosette stage is well distinguished (IoU = 0.6354), while the leaf expansion stage remains the primary challenge (IoU = 0.3541), reflecting the inherent difficulty of mapping this spectrally and morphologically transitional class. These findings indicate that hierarchy-aware learning reduces inter-stage confusion for most phenological classes, but transitional growth stages remain a significant limitation for field-scale deployment. The framework provides a foundation for stage-resolved crop monitoring to support harvest timing and supply forecasting in high-value vegetable systems.

1. Introduction

Projections from the Food and Agriculture Organization (FAO) estimate that the global population will reach approximately 9.2 billion by 2050 [1]. This demographic trend places immense pressure on the international food system. To ensure food security for this expanding population, global agricultural output must increase by 60% to 70% over current levels. Although research and policy frequently focus on staple grains, supply-sensitive vegetables are equally vital for food security and regional economic stability, particularly in East Asia where they constitute dietary staples. These vegetables exhibit rapid growth and adaptability to diverse climates and seasons, enabling year-round supply and supporting diversified farming systems. In Korea, for instance, napa cabbage and radish are essential ingredients in kimchi, a traditional fermented food that preserves vegetables through winter and serves as a fundamental side dish [2]. Despite their significance, both crops are highly vulnerable to supply and demand imbalances, resulting in considerable price volatility.

Accurate field-scale crop mapping is essential because it provides detailed information on the location and quantity of each crop. This enables reliable yield forecasting, early detection of local shortages or surpluses, and improved coordination of harvest, logistics, and cold-chain capacity with actual demand [3]. Within crop mapping, precise identification of phenological growth stages is critical for optimal harvest

* Corresponding author.
E-mail address: hmoon@sejong.ac.kr (H. Moon).

https://doi.org/10.1016/j.atech.2026.102365
Received 23 March 2026; Received in revised form 10 June 2026; Accepted 2 July 2026

timing, stage-specific yield forecasting, and efficient labor allocation in time-sensitive vegetable production systems. In these systems, accuracy measured in days can determine market outcomes [4]. However, field-scale growth-stage mapping remains significantly less developed than crop-type classification, particularly for supply-sensitive vegetables that exhibit rapid, spectrally subtle, and continuous phenological transitions.

Traditionally, crop growth-stage identification has depended on manual field surveys and expert inspections. These methods are labor-intensive and often subjective, as they rely on the experience of individual surveyors. Ground-based measurements may also be destructive and are challenging to repeat at the temporal frequency necessary for effective time-series monitoring [4,5]. Moreover, delays between data collection and reporting can hinder timely interventions. Automated monitoring systems are therefore crucial for enabling frequent, field-scale assessments of crop development. UAVs equipped with multispectral and high-resolution red-green-blue (RGB) sensors present a promising alternative to both ground surveys and satellite remote sensing. Satellite imagery is often limited by infrequent revisit intervals, atmospheric interference, and coarse spatial resolution, which can result in mixed-pixel effects in small agricultural plots. UAVs address these limitations by providing high-fidelity aerial imagery at centimeter-scale resolution, thereby offering a comprehensive view of key phenological features such as leaf color gradation, canopy architecture, and head formation [6]. These features are challenging to quantify reliably from the ground but are essential for distinguishing early-stage crops from soil, residue, and weed cover.

The increasing adoption of unmanned aerial vehicles (UAVs) has led to a significant rise in the volume and complexity of remote sensing datasets, driving the development of advanced computational methods for agricultural analysis [6,7]. Deep learning (DL), particularly convolutional neural networks (CNNs), has become the prevailing approach for image segmentation tasks. Modern frameworks such as DeepLabv3 [8], U-Net [9], and Fully Convolutional Networks (FCN) [10], achieve state-of-the-art performance by learning hierarchical spatial features directly from data, outperforming classical machine learning techniques that rely on heuristic-based feature selection. Recent research highlights the effectiveness of DL for crop phenotyping and growth monitoring. For instance, Lüling et al. [11] utilized ResNet-101 Mask R-CNN to segment cabbage heads and leaves for non-destructive yield-trait estimation. Yokoyama et al. [12] developed an instance-segmentation dataset containing 17,621 annotated cabbages to estimate key biophysical metrics, including Leaf Area Index (LAI) and biomass. Arab et al. [13] implemented pose-estimation models, such as YOLOv8s and YOLOv11s-pose, to estimate head diameters for yield prediction and spatial variability mapping, while Ye et al. [14] demonstrated that Mask R-CNN extracts and counts individual plants from UAV RGB imagery with greater accuracy and efficiency than object-based image analysis baselines.

A primary limitation of DL-based semantic segmentation is the assumption that target classes are mutually exclusive, which constrains each pixel to a single, isolated category [10,15]. This methodology overlooks the hierarchical relationships that exist between semantically related classes. While recent architectural advancements have emphasized encoder-decoder optimization, receptive field expansion, and attention mechanisms, they often neglect the explicit modeling of semantic structure [16]. Although certain approaches incorporate hierarchical information, these methods are frequently domain-specific [17,18] or require substantial architectural changes, such as the integration of graph neural networks (GNNs) [19], without directly enforcing tree-structured dependencies during inference. To address this limitation, Li et al. [16] proposed Hierarchical Semantic Segmentation Networks (HSSN), which organize classes into a tree hierarchy and associate each pixel with a root-to-leaf path. This framework follows human hierarchical reasoning and facilitates multi-level semantic representation.

2. Related work

2.1. Sensor fusion strategies

Over the last decade, sensor fusion strategies have gained increasing attention for enhancing segmentation robustness in agricultural remote sensing [20]. While multispectral imagery provides valuable signatures across visible and near-infrared bands, the integration of complementary modalities significantly enhances boundary precision and growth-stage discrimination [21]. For example, fusing multispectral and thermal infrared data has enhanced the detection of crop water stress and senescence, processes closely aligned with late-stage phenological transitions [22]. However, most fusion frameworks operate at coarse temporal resolutions or target single crop species, leaving fine-grained, multi-class growth-stage mapping under diverse open-field conditions largely unresolved.

2.2. Transfer learning and pretrained backbones

Transfer learning enables models pretrained on large-scale natural image datasets to efficiently adapt to domain-specific tasks with limited labeled data. Vision Transformer architectures, such as the ImageNet-22K pretrained Swin Transformer [23], exhibit strong transferability for semantic segmentation in high-resolution UAV imagery [24]. These architectures employ hierarchical feature representations and window-based self-attention mechanisms to capture fine-grained textures and global spatial dependencies, which are essential for resolving canopy-scale structures. Nevertheless, while conventional backbones are designed for three-channel RGB inputs, multispectral sensors acquire additional physiologically informative bands beyond the visible spectrum [25].

2.3. Domain adaptation and generalization

Domain adaptation and out-of-distribution generalization remain persistent challenges in agricultural applications [26]. Models trained on localized datasets frequently suffer performance degradation when deployed across new geographies or growing seasons, driven by domain shifts in soil reflectance, cultivar-specific morphology, solar illumination geometry, and canopy architecture [27]. In growth-stage mapping, identical developmental phases often exhibit divergent spectral and structural signatures across altitudinal gradients or seasonal cycles, due to temperature-mediated variations in canopy development. While prior cross-regional studies chose geographically stratified sampling and site-independent validation [28], a critical gap remains for adaptation frameworks that decouple phenological progression from environmental confounders. This deficiency continues to limit the operational scalability of growth-stage models.

2.4. Plant growth stages

Although plant development exhibits visually distinct milestones, phenological progression is fundamentally continuous. Imposing discrete, mutually exclusive categories on this progression exacerbates label ambiguity between adjacent stages, particularly where morphological and spectral transitions occur gradually rather than abruptly [29]. Empirical studies on growth-stage estimation consistently identify boundary confusion between temporally adjacent phases [30]. This limitation arises because canopy reflectance and structural attributes evolve continuously over days or weeks, rendering hard-label classification at single acquisition dates inherently misaligned with underlying biological dynamics [31]. Consequently, models that treat phenological stages as independent categories often yield temporally inconsistent predictions, underscoring the need for a segmentation approach that explicitly encodes ordinal relationships, continuous phenological scales, or probabilistic stage transitions.

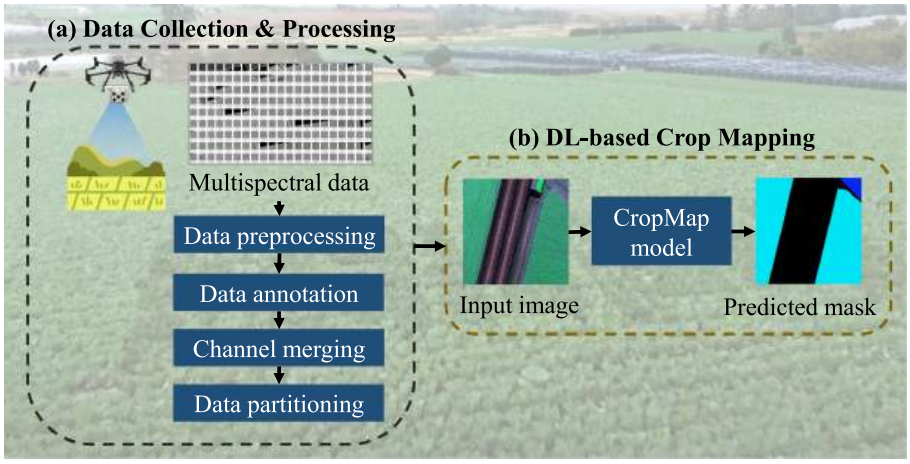

Fig. 1. Schematic representation of the CropMap framework applied to crop growth-stage mapping.

The primary contribution of this work is a hierarchy-aware segmentation framework that explicitly addresses the spectral, structural, and phenological continuity challenges inherent in field-scale crop monitoring. Our specific contributions are:

- Biologically-inspired hierarchical segmentation for crop phenology: The HSSN [16] was used to develop a two-level label taxonomy that encodes temporal adjacency and developmental continuity. Comparative experiments demonstrate that this hierarchy-aware design significantly reduces confusion across adjacent stages, mitigating the hard-boundary mismatch that limits flat segmentation of continuous biological processes.
- Optimal spectral fusion for growth-stage mapping: A Green–Red-Edge–near-infrared (NIR) composite was used to maximize phenological discriminability while preserving compatibility with pretrained ImageNet-22K models. A six-configuration ablation study confirms this composite outperforms RGB, RGB + NIR, and vegetation-index–based inputs, achieving 98.6% of the accuracy of full five-band inputs while minimizing redundancy and computational complexity.
- Standardized benchmark for field-scale phenology: A reproducible benchmark using the NIA multispectral UAV dataset, comprising 337,665 image patches across six developmentally ordered stages. The dataset spans 45 survey zones across four South Korean provinces and covers both summer and autumn cultivation cycles.

The structure of this study is organized into several key sections. Section 3 presents an overview of the CropMap framework. Section 4 introduces a large-scale, pixel-annotated dataset for crop growth-stage mapping. Section 5 describes the core components of CropMap, including hierarchical label modeling and multispectral feature fusion for growth-stage segmentation. Section 6 presents the experimental methodology, evaluation metrics, and comparative results. A thorough analysis of the findings, research constraints, and real-world applications is offered in Section 7. Finally, Section 8 summarizes the study and suggests potential approaches for further research.

3. Framework overview

Fig. 1 presents an overview of CropMap, a framework for mapping crop growth stages from multispectral UAV imagery. The name "CropMap" reflects its purpose: "Crop" denotes the precision-agriculture application domain, and "Map" highlights the hierarchical mapping strategy used to classify and spatially recognize growth-stage information.

CropMap has two sequential stages: data processing and crop mapping. During data processing, raw multispectral UAV images undergo preprocessing to mitigate sensor noise and improve radiometric consistency. Following this refinement, the images undergo manual labeling, in which each pixel is assigned specific information on crop species and phenological development. To prepare for model implementation, the resulting dataset is split into distinct training, validation, and testing subsets, with data augmentation applied exclusively to the training set to improve model generalization under field variability.

In the crop mapping stage, the HSSN is trained to recognize growth stages of Chinese cabbage and radish. The model jointly learns spatio-spectral features while explicitly enforcing hierarchical label constraints to reduce inter-class confusion between visually similar consecutive stages. During inference, CropMap receives raw multispectral inputs and generates georeferenced segmentation masks that spatially identify phenological stages across entire fields, facilitating scalable, stage-resolved monitoring for supply-chain decision support.

4. Crop growth stage mapping dataset

A large-scale UAV-based dataset was employed to enable fine-grained growth-stage mapping of supply-sensitive open-field vegetables. This dataset comprises approximately 337,665 multispectral patches, representing three principal development stages of Chinese cabbage (*Brassica rapa* subsp. *pekinensis*) and radish (*Raphanus sativus*). These crops were chosen for their economic sensitivity to fluctuations in supply and demand within regional fresh-produce markets.

The dataset was curated through a multi-institutional collaboration coordinated by the National Information Society Agency of Korea (NIA) and is publicly available under an open-access license[1]. Spatial Information Co., Ltd.[2] led the dataset construction. Sunyoung General Engineering Co., Ltd.[3] and Dia Co., Ltd. managed data collection. Annotation and preprocessing were performed by New Layer Co., Ltd.[4], Data Edu Co., Ltd.[5], and Muhan Information Technology Co., Ltd.[6].

4.1. Data acquisition and sources

Field data were collected during the 2022 growing season in the primary vegetable-producing regions of South Korea (Fig. 2). Data collection comprises 45 survey zones distributed across 12 localities in four

[1] https://www.nia.or.kr/site/nia_kor/main.do
[2] https://www.geomatic.co.kr/
[3] http://sunyoungeng.co.kr/
[4] http://egis.everlinks.co.kr/
[5] https://www.dataedu.kr/
[6] https://muhanit.kr/

Table 1
Site-level data acquisition summary. The dataset comprises 45 survey zones (001–045), distributed across 12 named localities in four provinces. Note: # denotes the total number of images. N and S indicate north and south, respectively.

Province	Specific localities	Crop(s)	Season	# images	Share
Gangwon-do	Anbandegi, Maebongsan, Gwineomi, Wangsan (Gangneung), Hajang (Samcheok), Taebaek highlands	Chinese cabbage	Summer	20,580	10.41%
Gangwon-do	Pyeongchang-gun, Hoengseong-gun	Radish	Summer	6125	3.10%
Chungcheongnam-do	Boryeong, Hongseong	Chinese cabbage	Autumn	10,565	5.35%
Jeolla-do (N + S)	Gochang, Haenam, Jindo	Chinese cabbage	Autumn	80,730	40.85%
Jeolla-do (N)	Buan	Radish	Autumn	14,140	7.16%
Jeju-do	Seongsan	Radish	Autumn	65,325	33.06%
Jeju-do	(within cabbage zones)	Chinese cabbage	Autumn	145	0.07%
Total	45 survey zones, 12 localities			197,610	100%

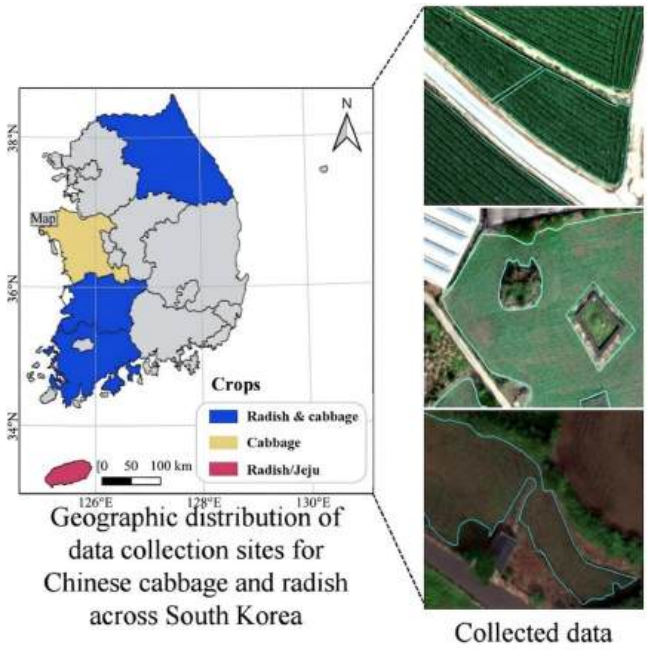

Fig. 2. Geographic distribution of data acquisition zones in South Korea. Note: Markers are color-coded according to the primary crop.

provinces, representing cultivation areas of approximately 16,880 ha for Chinese cabbage and 15,223 ha for radish.

As shown in Table 1, study sites were selected to represent the environmental conditions and phenological stages of Chinese cabbage and radish under commercial cultivation. Summer campaigns (June to August) focused on highland production zones in Gangwon Province (above 700 m a.s.l.). Chinese cabbage sites included the alpine regions of Anbandegi, Maebongsan, and Gwineomi, as well as elevated districts in Gangneung (Wangsan-myeon), Samcheok (Hajang-myeon), and Taebaek (Hasami, Sangsami, and Hwajeon districts). Summer radish cultivation was concentrated in Pyeongchang and Hoengseong counties. Autumn campaigns (September to November) targeted lowland and coastal production regions, primarily in Chungcheongnam-do (Boryeong, Hongseong) and the Jeolla region (Gochang, Haenam, Jindo) for Chinese cabbage, and in Jeollabuk-do (Buan) and Jeju-do (Seongsan) for radish. This multi-regional and multi-season sampling design captures phenological variability across contrasting agroclimatic zones, enabling robust evaluation of cross-environment model generalization.

The dataset covers approximately 16,880 ha of Chinese cabbage and 15,223 ha of radish cultivation. To resolve key developmental transitions, each field was revisited two to three times at approximately 10–14 day intervals, corresponding to major phenological shifts. Imagery was collected using a DJI Matrice 300 RTK UAV [32] equipped with a MicaSense RedEdge-MX multispectral sensor [33], which captures five specific wavebands: blue (475 nm), green (560 nm), red (668 nm), red-edge (717 nm), and near-infrared (842 nm). The drone missions were conducted at a constant altitude of 120 m, resulting in a Ground Sampling Distance (GSD) of 8 cm/pixel. To ensure radiometric consistency, all missions were scheduled between 09:00–11:00 Korea Standard Time under stable solar illumination. By including data gathered under varying sky conditions (from clear to 30% cloud cover), the model's resilience to practical environmental challenges was enhanced. The final output comprises orthorectified, atmospherically corrected multispectral images in georeferenced TIFF format.

4.2. Data preparation

4.2.1. Data pre-processing and orthophoto generation

Initial quality control of raw UAV imagery was performed by the primary data collector to exclude frames affected by motion blur, lens flare, or Global Navigation Satellite System (GNSS) signal loss. Validated images were subsequently processed in Pix4Dmapper [34] following a standardized structure-from-motion (SfM) workflow (Fig. 3). The pipeline comprised initial image alignment using onboard Global Positioning System (GPS), geometric correction, and ground control point (GCP) adjustment.

Radiometric calibration was conducted using a MicaSense calibrated reflectance panel (CRP; nominal reflectance of 0.53–0.57 across all bands), which was deployed immediately before takeoff and after landing for each mission to normalize ambient illumination variability. This procedure produced georeferenced, band-aligned orthomosaics for all five spectral channels. The final orthomosaics were clipped to active field boundaries to reduce edge distortion and flight-line overlap artifacts, thereby retaining only valid crop canopy areas for subsequent analysis.

For model training, the orthomosaics were divided into 480×480 pixel patches. Patches with less than 10% target-crop coverage or with more than 10% non-crop elements (such as bare soil, weeds, or infrastructure) were excluded. Expert annotators performed pixel-wise semantic labeling, assigning each pixel a class label that specified both the crop species and its phenological stage. The resulting dataset consists of georeferenced patches representing the principal growth stages of both crops across diverse field conditions.

4.2.2. Spectral feature engineering

Previous studies demonstrate that the green (G), red-edge (RE), and near-infrared (NIR) spectral bands offer complementary information critical for distinguishing phenological stages in crops [35,36]. Specifically, the green band is highly responsive to chlorophyll content during early growth, the red-edge band detects subtle changes in canopy structure and chlorophyll concentration, and the NIR band is closely associated with biomass accumulation and LAI in dense canopies.

Building on spectral feature engineering approaches in biomass estimation [36] and crop classification [35], a three-channel pseudo-RGB composite was developed by assigning the RE, NIR, and G bands to the Red, Green, and Blue input channels, respectively. This spectral configuration improves the differentiation of sequential growth stages, particularly during rapid transitions, while ensuring compatibility with DL ar-

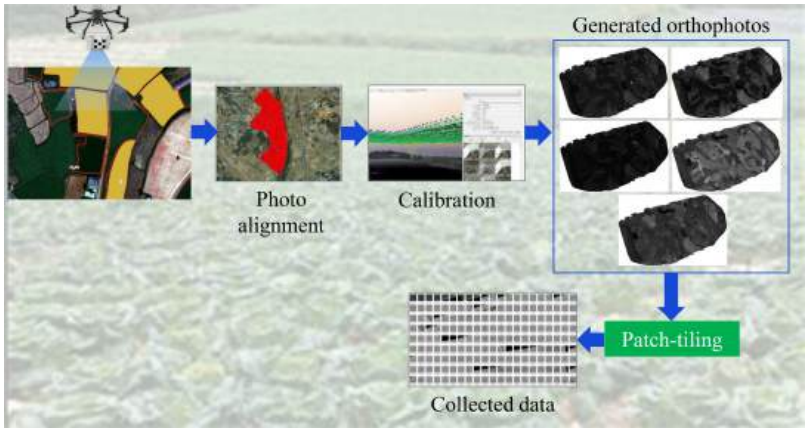

Fig. 3. Primary pre-processing procedures applied to the crop growth dataset.

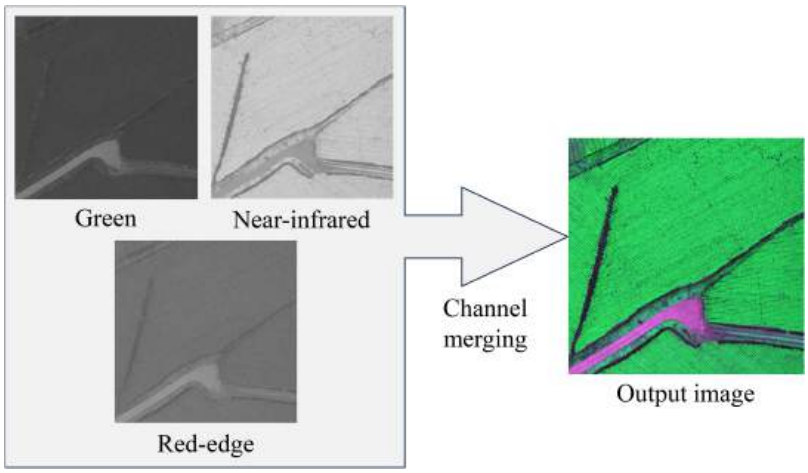

Fig. 4. Multispectral fusion process mapping single-band reflectance into a three-channel composite. The visualization employs pseudo-RGB composites, where the green channel represents green reflectance, the red channel represents near-infrared (NIR), and the blue channel represents red-edge (RE). Note: Bright green areas correspond to healthy, photosynthetically active vegetation, including cabbage, radishes, and surrounding trees. Pink or magenta regions denote bare soil, dirt roads, paths, and unplanted field edges. Black tones indicate shadows, typically cast by trees, or areas with higher soil moisture resulting in increased infrared absorption. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

chitectures pretrained on natural images [37]. The fusion method deliberately omits the red band, as it typically saturates early during canopy closure and contributes minimal additional information for mid-to-late phenological stages in dense crops. Fig. 4 presents the band assignment and the resulting composite imagery.

4.2.3. Post-processing and labeling

A standard quality assurance step was implemented to eliminate geometric artifacts, including stitching errors, lens distortion residuals, and warping induced by GCPs. Subsequently, the refined orthophotos were segmented into individual field plots to enable plot-level analysis and ensure consistent labeling.

Semantic annotation was conducted at the polygon level using a web-based platform developed by New Layer Co., Ltd. These polygons were then rasterized into pixel-wise labels for model training. Instead of manual labeling from the beginning, annotators refined pre-existing field boundary polygons from the SmartFarm Map database [38] to align precisely with visible crop extents in each orthomosaic. This constraint-driven workflow reduced arbitrary boundary placement and improved the spatial consistency of the annotation process. Each polygon received a growth-stage label (`GRWH_STEP_CODE`) based on ground truth field surveys conducted on the same date. During these surveys, field teams physically inspected each plot to record crop type, phenological stage, planting density, and anticipated harvest date, ensuring that stage assignments were based on verified on-site conditions rather than solely on visual interpretation of imagery.

The six target growth stages were defined according to explicit botanical criteria specified in the NIA dataset construction guidelines[7]. For Chinese cabbage, *planting* refers to transplanted seedlings exhibiting true leaf emergence and root establishment; *rosette* encompasses outer-leaf vertical elongation before heading initiation; and *heading* includes inner-leaf compaction into a firm head through pre-harvest maturity. For radish, *seedling* indicates post-germination hypocotyl establishment; *leaf expansion* is characterized by approximately 15 erect leaves; and *root enlargement* is defined by rapid taproot thickening concurrent with above-ground leaf senescence. To address transitional or ambiguous regions, standardized protocols were implemented: shadowed areas and non-crop infrastructure were excluded using interior hole polygons, and plots containing mixed crop types were divided into distinct sub-polygons rather than assigned a composite label.

Annotation quality was assessed using a three-level review protocol. The first level confirmed temporal synchronization between imagery and annotations. The second level identified topological errors and duplicate entries after orthomosaic generation. The third level evaluated

[7] https://eng.nia.or.kr

Table 2
Breakdown of annotated patches according to phenological stage for Chinese cabbage (planting, rosette, heading) and radish (seedling, leaf expansion, root enlargement).

Crop	Stage	Images	Description
Cabbage	Planting	22,895 (7%)	Initial phase; emergence of true leaves and root establishment.
	Rosette	107,305 (32%)	Rapid horizontal leaf growth to maximize photosynthesis.
	Heading	89,290 (26%)	Inner leaves overlap to form a compact central head.
Radish	Seedling	59,340 (18%)	Early germination phase; presence of cotyledons.
	Leaf expansion	34,695 (10%)	Increase in foliage biomass to fuel root development.
	Root enlargement	24,140 (7%)	Significant thickening of the taproot and hypocotyl.
	Total	337,665 (100%)	

guideline compliance and completeness through automated polygon-overlay validation tools and mandatory expert visual inspection. The dataset was required to meet quantitative benchmarks of at least 95% semantic labeling accuracy, ≥ 99% syntactic/schema compliance, and a downstream segmentation validation target of mIoU at least 80%. To ensure inter-annotator consistency, all annotators used identical SmartFarm baseline polygons and synchronized field survey records for stage assignment.

Quality assurance and geometric artifact removal were conducted using Pix4Dmapper[8]. Field-plot subdivision was performed in QGIS[9] using field boundary polygons from the custom-developed Almaps application as spatial references.

4.3. Dataset description

To mitigate spatiotemporal data leakage associated with random patch-level splitting, a group-aware partitioning strategy was applied at the orthomosaic level. Each orthomosaic, defined by a unique combination of site, field, and acquisition date, received a group identifier $g = (\texttt{site\_id}, \texttt{field\_id}, \texttt{date})$. All patches derived from the same orthomosaic were allocated exclusively to a single data partition. The unique groups were subsequently divided into training, validation, and test subsets in an 80:10:10 ratio. This orthomosaic-level partitioning ensures that spatially adjacent and temporally correlated samples are restricted to a single subset, thereby preventing data leakage. The final partition consists of 270,133 training patches and 33,766 patches each for validation and testing.

Table 2 presents the dataset distribution by crop type and growth stage. The dataset contains 337,665 pixel-annotated patches organized into six hierarchical growth stages. These categories represent the biological progression from seedling emergence to pre-harvest maturity for both Chinese cabbage and radish.

Fig. 5 illustrates representative samples for each growth stage of Chinese cabbage and radish under diverse field conditions, including different elevations, soil types, and planting densities. The combination of high spatial resolution (8 cm GSD) and large data volume necessitates a model architecture that effectively balances representational capacity and computational efficiency.

5. Methodology

CropMap extends the HSSN framework [16], which models visual scenes with structured, pixel-wise labels organized by a class taxonomy rather than treating categories as independent entities. CropMap directly incorporates its core architectural components. Four domain-specific adaptations are introduced to render HSSN suitable for crop phenology mapping from multispectral UAV imagery:

- Biologically grounded label hierarchy (Table 5): HSSN was originally developed for human-body-part parsing using the Pascal-Person-Part hierarchy, which is based on visual similarity. In the context of crop phenology, no comparable hierarchy exists. Therefore, a three-level agronomic tree (Root → Crop → Phonological stage) is constructed from botanical domain knowledge, encoding the temporal developmental sequence of cabbage and radish as a directed hierarchy. This is the primary domain-specific design contribution of this work.
- Focal modulation in the Tree-Min loss (L_{FTM}, Equation 3): The original HSSN loss employs standard binary cross-entropy combined with tree-consistency constraints. Field-scale phenological datasets exhibit severe class imbalance; for example, leaf expansion comprises only 10% of training patches compared to 32% for rosette, which standard cross-entropy addresses inadequately. Focal modulation ($\gamma = 2.0$) is integrated into the tree-consistent loss to automatically down-weight well-classified background pixels and amplify gradients from ambiguous stage boundaries. This represents a loss-function modification rather than direct adoption.
- Inverse-frequency triplet sampling: HSSN samples pixel triplets uniformly from the training set. To prevent the Tree-Triplet loss from being dominated by overrepresented classes, triplet sampling probabilities are weighted by the inverse of each class's patch frequency. This approach ensures that rare stages, such as leaf expansion and root enlargement, are proportionally represented in the metric learning objective.
- Multispectral pseudo-RGB input (G + NIR + RE): HSSN assumes standard three-channel RGB input. In this adaptation, a physiologically motivated composite is used, mapping Green, NIR, and red-edge bands to the three input channels. This maintains compatibility with ImageNet-22K pretrained Swin weights while incorporating spectral information beyond the visible range. The optimality of this selection over RGB, RGB + N, VI-based composites, and full five-band input is validated in Section 6.1.

Fig. 6 presents the complete training pipeline, which consists of five core components:

- Encoder: Extracts spatial and spectral features from input imagery I using a Swin Transformer backbone [23], resulting in a dense feature tensor $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$. The backbone is initialized with ImageNet-22K weights and subsequently fine-tuned in an end-to-end manner.
- Segmentation head: Projects $\mathbf{F}$ into a structured score map $\mathbf{S} \in \mathbb{R}^{H \times W \times |V|}$ across all hierarchy nodes V, not limited to leaf classes (where $|V|$ is the total number of nodes). This approach formulates hierarchical semantic segmentation as a pixel-wise multi-label classification problem.
- Projection head: A lightweight multi-layer perceptron (MLP) maps per-pixel features to 256-dimensional embeddings during training, facilitating contrastive optimization through the Tree-Triplet loss.

[8] https://support.pix4d.com/hc/pix4dmapper
[9] https://docs.qgis.org/3.44/en/docs/user_manual/working_with_projections/working_with_projections.html

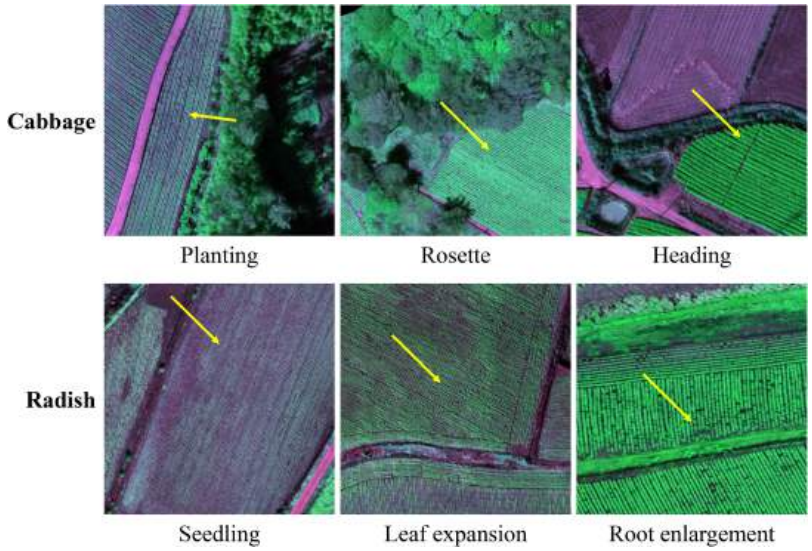

Fig. 5. Representative samples illustrating the growth stages of Chinese cabbage and radish included in the dataset.

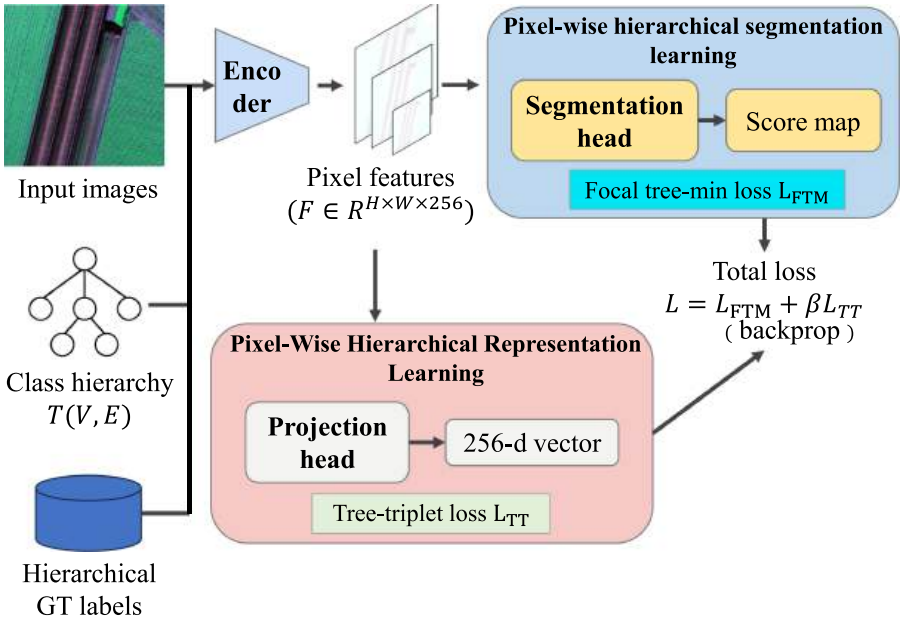

Fig. 6. Schematic overview of the CropMap training pipeline utilized for crop growth-stage mapping.

- Focal tree-min loss: Enforces hierarchy-coherent predictions by estimating a hierarchy-consistent score map that satisfies the positive and negative T-properties, while applying a focal modulating factor to emphasize challenging pixels. Predictions that violate hierarchy constraints receive explicit penalties.
- Tree-triplet loss: Reshapes the embedding space using hierarchy-induced margin constraints. This loss reduces distances between embeddings of semantically similar pixels (proximal in the hierarchy) and increases distances for dissimilar pixels, thereby aligning the representation space with structured class relationships.

Table 3 summarizes the CropMap architecture and data flow from multispectral input to hierarchical prediction. The framework processes a three-channel pseudo-RGB input image $I \in \mathbb{R}^{H\times W\times 3}$ (green, red-edge, NIR) using an encoder f_{ENC}, which is a Swin Transformer Small backbone [23] pretrained on ImageNet-22K. This encoder extracts hierarchical multi-scale features $F \in \mathbb{R}^{H\times W\times C}$. During training, the projection head f_{PROJ} (a 1×1 convolution, ReLU, and L_2 normalization) maps F to a 256-dimensional pixel embedding space $E \in \mathbb{R}^{H\times W\times 256}$ for contrastive optimization with the Tree-Triplet loss. The segmentation head f_{SEG} uses an upsampling decoder to convert F into a score map $S \in \mathbb{R}^{H\times W\times |V|}$, where $|V|$ is the number of nodes in the vegetable growth-stage hierarchy. Sigmoid activation produces per-node probabilities, enabling pixel-wise multi-label prediction while enforcing hierarchical consistency. During inference, only the encoder and segmentation head are used; the projection head is omitted as it is only required for representation learning during training.

5.1. Encoder

The encoder functions as the principal feature extractor, transforming raw multispectral inputs into hierarchical representations that capture both fine-grained morphological characteristics and field-scale contextual patterns essential for growth-stage discrimination. To address the complex spatial structure of open-field vegetable cultivation, CropMap employs the Swin Transformer (Swin-Small) backbone [23],

Table 3
Architecture description of the CropMap framework.

Component	Operations	Channels	Output size
Input	Image I	3	$H \times W \times 3$
Encoder (f_{ENC})	Swin-small backbone	C	$H/32 \times W/32 \times C$
Projection head (f_{PROJ})	1×1 Conv $\rightarrow$ ReLU $\rightarrow$ L_2 Norm	256	$H/32 \times W/32 \times 256$
Segmentation head (f_{SEG})	MaskFormer decoder	$\lvert V \rvert$ (All hierarchy classes)	$H \times W \times \lvert V \rvert$

Table 4
Multi-scale feature extraction stages in the Swin-Small encoder. Each stage downsamples spatial resolution while increasing feature dimensionality.

Stage	Resolution	Channels	Layers	Attention heads
0 (Patch Embed)	$\frac{H}{4} \times \frac{W}{4}$	96	–	–
1	$\frac{H}{4} \times \frac{W}{4}$	96	2	3
2	$\frac{H}{8} \times \frac{W}{8}$	192	2	6
3	$\frac{H}{16} \times \frac{W}{16}$	384	18	12
4	$\frac{H}{32} \times \frac{W}{32}$	768	2	24

initialized with ImageNet-22K pretrained weights to leverage transfer learning from large-scale visual recognition tasks.

For an input image $I \in \mathbb{R}^{H \times W \times 3}$, the encoder divides it into non-overlapping 4×4 patches, which are then linearly projected into an embedding space. The network consists of four hierarchical stages (Table 4), each comprising Swin Transformer blocks with shifted window attention, followed by patch merging. Spatial resolution is progressively reduced while channel dimensionality increases, resulting in a multi-scale feature pyramid that preserves high-resolution details for subtle stage transitions and aggregates coarse contextual information for robust crop-background separation.

Formally, the encoder f_{ENC} maps the input to a dense feature tensor at the original resolution through upsampling of deep features:

$$F = f_{\text{ENC}}(I) \in \mathbb{R}^{H \times W \times C}, \tag{1}$$

where each pixel is represented by an embedding in $\mathbb{R}^C$. In addition to generic feature extraction, CropMap regularizes the encoder output through a tree-triplet loss, aligning the embedding space with the growth-stage hierarchy. This loss enforces hierarchy-induced margins by drawing semantically similar stages, such as rosette and heading of cabbage, closer together, while separating distant categories, including crop, background, or different crop types.

This multi-scale design is particularly effective for vegetable phenology mapping. Shallow network stages resolve leaf-level textural variations that are critical for distinguishing early vegetative phases. In contrast, deeper stages capture canopy-scale structural patterns, such as head compactness, which differentiate late-stage maturity classes. The resulting feature hierarchy supports subsequent hierarchical segmentation and maintains computational efficiency suitable for UAV-based inference.

5.2. Hierarchy-aware segmentation learning

Table 5 presents the hierarchical label taxonomy used for growth-stage segmentation. This taxonomy is structured as a directed tree with three levels: (i) a root node that partitions pixels into Background (non-crop elements) and Crops; (ii) two internal nodes representing crop categories (Radish and Chinese cabbage); and (iii) six leaf nodes corresponding to specific growth stages (Radish: seedling, leaf expansion, root enlargement; Chinese cabbage: planting, rosette, heading). During inference, each pixel receives a root-to-leaf path assignment rather than a single label. This approach enforces two essential biological constraints: (1) a pixel identified as crop must be assigned to exactly one crop type, and (2) any crop pixel must be assigned precisely one valid growth stage within its respective crop category. Any violations of these constraints are explicitly penalized to ensure hierarchy-coherent predictions that accurately represent the continuous progression of plant development.

5.2.1. Pixel-wise hierarchical segmentation learning

The segmentation head utilizes the MaskFormer decoder architecture [39] to perform hierarchical semantic segmentation. With encoder features $F \in \mathbb{R}^{H \times W \times C}$ as input, the head generates a score map $S \in \mathbb{R}^{H \times W \times |V|}$, where $|V| = 10$. This map includes two root-level nodes (Background, Crops), two internal crop-category nodes (Radish, Cabbage), and six leaf-stage nodes (three radish stages and three cabbage stages). In contrast to conventional segmentation approaches that predict mutually exclusive leaf labels, CropMap formulates growth-stage mapping as pixel-wise multi-label classification. Each pixel receives confidence scores for all nodes along its root-to-leaf path within the hierarchy.

For each class $v \in V$ and spatial location, the head outputs sigmoid-activated confidence scores $s_v \in [0, 1]$. To ensure logical consistency with the hierarchy during inference, the raw scores are converted into hierarchy-coherent predictions p_v using Tree-Min constraints [16]:

$$p_v = \begin{cases} \min_{u \in \mathcal{A}_v \cup \{v\}} (s_u) & \text{if } l_v = 1, \\ 1 - \min_{u \in \mathcal{C}_v \cup \{v\}} (1 - s_u) & \text{if } l_v = 0. \end{cases} \tag{2}$$

where $\mathcal{A}_v$ denotes the set of ancestor classes and $\mathcal{C}_v$ represents descendant classes.

- Positive constraint (p_v): A child node cannot possess higher confidence than any ancestor. This enforces the logical hierarchy in which a specific growth stage (for example, "Heading") cannot be present unless the plant exists within its parent category (for example, "Chinese cabbage").
- Negative constraint ($1 - p_v$): This constraint propagates rejection downward. For instance, if a parent node is negative, all of its descendants must also be negative.

During training, the head is optimized using the Focal Tree-Min loss $\mathcal{L}^{\text{FTM}}$. This loss function integrates hierarchy-consistent scores with focal modulation to address class imbalance and to emphasize ambiguous boundary pixels:

$$\mathcal{L}^{FTM}(p) = -\sum_{v \in V} \left[l_v (1 - p_v)^\gamma \log(p_v) + (1 - l_v) p_v^\gamma \log(1 - p_v) \right], \tag{3}$$

Here, $\gamma = 2.0$ determines the focusing strength. The modulating factors $(1 - p_v)^\gamma$ and p_v^γ reduce the influence of well-classified pixels and enhance gradients from challenging examples. This approach is particularly important for accurately distinguishing subtle transitions between consecutive growth stages.

5.2.2. Pixel-wise hierarchical representation learning

The projection head enables hierarchical representation learning by mapping encoder features to a structured embedding space that captures semantic relationships within the crop growth-stage hierarchy. It is implemented as a lightweight multilayer perceptron (MLP) consisting of two 1×1 convolutional layers with ReLU activation and L_2 normalization. This module transforms encoder features $F \in \mathbb{R}^{H \times W \times C}$ into 256-dimensional pixel embeddings $E \in \mathbb{R}^{H \times W \times 256}$. The design maintains spatial resolution and introduces minimal computational overhead, which is essential for dense prediction tasks.

Table 5
Class hierarchy employed in this study (root → crop → phenological stage). The background is represented as a root-level leaf class.

Level	Parent	Node (class)	Type	Pixel label mapping
0	–	Background	leaf	Non-crop pixels
0	–	Crops	internal	All crop pixels belong to this subtree
1	Crops	Radish	internal	Crop category
2	Radish	Seedling	leaf	Radish stage label
2	Radish	Leaf expansion	leaf	Radish stage label
2	Radish	Root enlargement	leaf	Radish stage label
1	Crops	Chinese cabbage	internal	Crop category
2	Chinese cabbage	Planting	leaf	Cabbage stage label
2	Chinese cabbage	Rosette	leaf	Cabbage stage label
2	Chinese cabbage	Heading	leaf	Cabbage stage label

During training, the projection head is used in conjunction with the Tree-Triplet loss $\mathcal{L}^{\mathrm{TT}}$ to organize the embedding space according to hierarchical semantics. Pixel triplets $\{i, i^+, i^-\}$ are sampled according to the following criteria:

- The anchor i and positive i^+ are selected from the same or hierarchically proximate nodes, such as consecutive growth stages of Chinese cabbage;
- Negative i^- belongs to a semantically distant node (e.g., radish or background categories);
- Sampling probabilities are adjusted using inverse class frequency to address class imbalance.

The loss minimizes the cosine distance between anchor–positive pairs while maximizing separation from negatives:

$$\mathcal{L}^{\mathrm{TT}}(i, i^+, i^-) = \max\{\delta(i, i^+) - \delta(i, i^-) + m, 0\}, \tag{4}$$

where $\delta(\cdot, \cdot)$ denotes cosine distance,

$$\delta(x, y) = 1 - \frac{x \cdot y}{\|x\|\|y\|}, \tag{5}$$

and i, i^+, and i^- are pixel embeddings extracted from E.

The margin $m(u, v)$ is hierarchy-aware and scales with semantic distance in the taxonomy:

$$m(u, v) = m_c + 0.5\,\psi(u, v), \qquad m_c = 0.1, \tag{6}$$

with normalized tree distance $\psi(u, v)$ calculated as:

$$\psi(u, v) = \frac{d_{\mathrm{tree}}(u, v)}{D}, \tag{7}$$

Here, $d_{\mathrm{tree}}(u, v)$ denotes the shortest-path length between nodes u and v in the hierarchy $\mathcal{T}$, and D represents the diameter of the hierarchy tree, defined as the maximum shortest-path length between any two nodes. This approach enforces greater separation for distant categories, such as different crop types, compared to closely related classes, such as consecutive growth stages. As a result, the embedding geometry is directly aligned with biological progression.

The projection head is removed during inference. The encoder maintains hierarchy-aware representations through end-to-end training with $\mathcal{L}^{\mathrm{TT}}$, resulting in improved segmentation accuracy without increasing deployment overhead. This design supports real-time, field-scale inference on resource-constrained UAV platforms.

5.3. Implementation description

5.3.1. Training objective

CropMap is optimized via a multi-task objective that jointly learns hierarchy-consistent segmentation and semantically structured representations. The total loss combines the Focal Tree-Min loss $\mathcal{L}^{\mathrm{FTM}}$ for hierarchical pixel classification and the Tree-Triplet loss $\mathcal{L}^{\mathrm{TT}}$ for embedding regularization:

$$\mathcal{L} = \mathcal{L}^{FTM} + \beta(t)\,\mathcal{L}^{TT}. \tag{8}$$

The focusing parameter in $\mathcal{L}^{FTM}$ is set to $\gamma = 2$ to emphasize difficult growth-stage transitions. The balancing coefficient $\beta(t)$ is scheduled with cosine annealing within $[0, 0.5]$ to control the influence of metric learning over time. For instance, early training prioritizes accurate segmentation, mid-training increases the contribution of $\mathcal{L}^{TT}$ to shape the embedding space, and late training reduces it to promote stable convergence.

To isolate and quantify the contribution of each hierarchical learning component, we conduct a systematic loss ablation study in this section. Four configurations are evaluated under identical training protocols: (i) **CropMap-Flat**: Swin-Small backbone with a MaskFormer decoder trained using standard cross-entropy loss without tree-structured constraints. This configuration serves as the non-hierarchical reference; (ii) **CropMap-FTM**: trained exclusively with the Focal Tree-Min loss $\mathcal{L}^{\mathrm{FTM}}$ (with the triplet weighting coefficient $\beta(t) = 0$ to disable the Tree-Triplet component), isolating the effect of hierarchical classification supervision; (iii) **CropMap-TT**: trained with the Tree-Triplet loss $\mathcal{L}_{\mathrm{TT}}$ combined with standard cross-entropy (replacing $\mathcal{L}^{\mathrm{FTM}}$), isolating the contribution of hierarchical metric learning to feature representation; and (iv) **CropMap-Full**: the complete framework integrating both $\mathcal{L}^{\mathrm{FTM}}$ and $\mathcal{L}_{\mathrm{TT}}$. This ablation design explicitly tests whether hierarchical label constraints and structured embedding regularization reduce confusion between biologically adjacent growth stages relative to a flat classification baseline.

5.3.2. Experimental configuration

We developed the proposed framework using PyTorch 1.7.1, conducting the training phase on a Linux-based system equipped with two NVIDIA RTX A6000 GPUs (48 GB VRAM per unit). Baselines, including U-Net [9], DeepLabv3 [8], MaskFormer [39], and SegFormer [40], were implemented through MMSegmentation [41] framework.

All models were trained under identical conditions to ensure fair comparison. The Swin-Small encoder was optimized with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$), initial learning rate 6×10^{-5}, and weight decay 0.01. Learning rates followed a polynomial decay schedule with power $\alpha = 0.9$ and a linear warmup over the first 10% of iterations. All backbones were initialized from ImageNet-22K pretrained weights. Training employed 480×480 pixel crops sampled from orthomosaic patches, batch size 64, and 40 epochs. Augmentation included random horizontal flipping (50% probability) and random scaling with ratios sampled uniformly from $[0.5, 2.0]$. No color jittering was applied to preserve radiometric integrity of multispectral inputs.

To ensure comparability at the backbone level, all baseline models utilize the Swin-Small encoder, which is identical to the CropMap backbone, and are initialized with the same ImageNet-22K pretrained weights. Each model is trained using an identical protocol, including the same data pipeline, augmentation policy, optimizer configuration, learning rate schedule, batch size, and number of epochs. Additional analyses are presented in subsequent sections: Section 6.1 systematically ablates spectral input configurations, Section 6.3 examines hierarchical learning mechanisms, and Table 9 (Section 6.5) benchmarks archi-

tectural complexity and computational efficiency, including parameter counts and inference latency.

5.4. Evaluation metrics

Segmentation performance was quantified using the standard semantic segmentation metric of mean Intersection over Union (mIoU). For each growth-stage class c, the class-conditional IoU is computed from the pixel-level confusion matrix:

$$\mathrm{IoU}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}. \quad (9)$$

where TP_c, FP_c, and FN_c denote true positives, false positives, and false negatives for class c, respectively. The mean IoU aggregates performance across all N leaf-stage classes:

$$\mathrm{mIoU} = \frac{1}{N}\sum_{c=1}^{N} \mathrm{IoU}_c. \quad (10)$$

mIoU serves as the primary evaluation metric for three key reasons: (i) it equally penalizes both over-segmentation (false positives) and under-segmentation (false negatives), (ii) it is robust to class imbalance, which is critical for the dataset used in this study where early/late stages often occupy smaller areas than mid-season canopies, and (iii) it directly measures segmentation accuracy at stage boundaries, which is essential for harvest timing decisions.

6. Results

6.1. Spectral band contribution analysis

A systematic ablation study was conducted to isolate the contribution of individual spectral bands to growth-stage discrimination. Six input configurations were evaluated: (i) visible bands only (RGB: blue, green, red), (ii) RGB augmented with near-infrared (RGB + N), (iii) physiologically optimized triplets (green, NIR, red-edge; G + N + RE and red, NIR, red-edge; R + N + RE), (iv) a Vegetation Index (VI)-based configuration (green, NIR, Normalized Difference Vegetation Index (NDVI); G + N + NDVI), and (v) full five-band input using all raw spectral channels (blue, green, red, near-infrared, red-edge; B + G + R + N + RE) via a learnable 1×1 projection layer mapping 5 channels to 3 before the Swin-Small backbone.

The VI-based configuration combines raw Green and near-infrared (NIR) bands with NDVI as the three input channels, rather than utilizing multiple vegetation indices. Raw Green and NIR bands were retained because they preserve absolute reflectance magnitudes and spatial texture information that normalized ratio indices inherently discard. This spatial data is critical for pixel-wise boundary recognition at phenological transitions [42]. NDVI was selected as the index channel because it dynamically captures canopy closure progression from sparse planting-stage coverage to the dense, closed canopy of the heading stage, representing the most diagnostically important spectral gradient across the plant growth stages [43].

Other VI-related indexes, such as the Normalized Difference Red Edge Index (NDRE), was not included as a replacement channel because the model accepts three-channel input to maintain compatibility with ImageNet-22K pretrained Swin-Small weights, the red-edge band is already present as a raw reflectance channel in the G + N + RE configuration [44]. Including both NDVI and NDRE would introduce spectral redundancy under dense vegetative cover where both indices converge toward saturation.

The extended ablation results presented in Table 6 demonstrate a consistent and interpretable performance hierarchy across all six spectral configurations. The RGB baseline yields the lowest mIoU (0.3873), indicating that visible wavelengths alone are insufficient for resolving chlorophyll-driven transitions between consecutive green-canopy stages. The R + N + RE configuration (mIoU = 0.4355) performs notably worse than RGB + N (0.4854) despite the inclusion of the red-edge band. This outcome isolates the role of the green channel: when Green is omitted and replaced with Red alongside NIR and RE, overall discrimination declines. This finding demonstrates that the red band's early saturation during canopy closure actively impairs phenological stage separation, rather than being neutral. The G + N + NDVI configuration (mIoU = 0.5114) surpasses RGB + N by +0.0260, confirming that the addition of a normalized difference vegetation index alongside raw Green and NIR bands captures additional physiological information beyond raw reflectance intensity. However, G + N + RE (mIoU = 0.5376) consistently outperforms G + N + NDVI by +0.0262 across all six classes, with the largest margins observed for leaf expansion (+0.0251, +7.6%) and seedling (+0.0271, +5.3%). This difference arises because NDVI, as a normalized ratio, compresses the full reflectance signal into a scalar index, discarding the absolute reflectance magnitude and spatial texture information preserved by the raw red-edge band. For pixel-wise boundary delineation at phenological transitions, this spatial texture is as diagnostically important as the physiological index value. The five-band configuration achieves the highest mIoU (0.5452), confirming that incorporating all available spectral channels provides the maximum discriminative information for growth-stage segmentation.

Although the five-band configuration achieves the highest overall mIoU (0.5452), its margin over G + N + RE remains consistently narrow: +0.0076 mIoU overall, with per-class gains ranging from +0.0069 (leaf expansion) to +0.0084 (planting). This uniform improvement is smaller than the gap between any other adjacent configurations in the ranking. Such uniformity indicates that the learnable 1×1 projection layer, initialized to replicate G + N + RE weighting, converges to a solution in which the blue and red channels contribute only marginal complementary information beyond what G, NIR, and RE already provide. The projection effectively suppresses those channels and approaches the G + N + RE solution. Therefore, G + N + RE remains the recommended operational configuration for three reasons. First, it achieves 98.6% of the five-band mIoU without architectural modification, maintaining full compatibility with standard ImageNet-22K pretrained Swin Transformer weights and eliminating the need for a learnable projection layer, which introduces additional parameters and training sensitivity. Second, the three-channel pseudo-RGB format is directly compatible with the full ecosystem of pretrained vision backbones, enabling straightforward transfer learning and future architectural upgrades without changes to the input pipeline. Third, the 0.0076 mIoU advantage of five-band inputs is unlikely to result in operationally meaningful differences in harvest timing or supply forecasting decisions, which are the primary use cases motivating this study. This is particularly true given the larger uncertainty introduced by the leaf expansion class (IoU = 0.3541 under G + N + RE), which remains the dominant bottleneck for all configurations regardless of spectral input. Collectively, these findings validate the physiological band-selection rationale: Green, NIR, and red-edge together capture the spectral variance most critical to phenological discrimination in dense vegetable crops, while the remaining bands are largely redundant for this specific task.

6.2. CropMap performance evaluation

CropMap exhibits robust convergence and effective learning of hierarchical growth-stage features across 40 training epochs. Fig. 7 illustrates that validation accuracy increases rapidly during the first 10 epochs, rising from 77% to 88%, and subsequently stabilizes at 95%. Concurrently, training loss decreases steadily from 0.21 to 0.09. The close correspondence between training and validation metrics suggests effective regularization and minimal overfitting.

Per-class segmentation performance on the test set (Table 7) highlights both the strengths and limitations of fine-grained crop growth-stage discrimination. CropMap achieves high background separation (IoU = 0.9128), effectively distinguishing crop canopy from soil, plastic mulch, and weeds. Among the growth stages, the rosette phase attains

Table 6
Ablation study evaluating spectral input configurations on the test set. Results are presented as per-class Intersection over Union (IoU) and mean IoU (mIoU), with configurations ordered by mIoU. Note: R denotes red, G green, B blue, N near-infrared, and RE red-edge channels. In the 5-band configuration, a learnable 1×1 projection layer reduces five channels to three prior to the Swin-Small backbone. The VI-based composite utilizes NDVI = (N-R)/(N+R) as input channels. The G+N+NDVI configuration retains raw Green and NIR bands for spatial texture preservation and uses NDVI as the third channel to capture canopy closure dynamics.

Class	RGB	G+N+NDVI	RGB+N	R+N+RE	G+N+RE	5-band
Planting	0.3942	0.5165	0.4915	0.4415	0.5436	0.5520
Rosette	0.4855	0.6112	0.5822	0.5340	0.6354	0.6430
Heading	0.4412	0.5638	0.5401	0.4895	0.5908	0.5980
Seedling	0.3855	0.5075	0.4832	0.4320	0.5346	0.542
Leaf expansion	0.2044	0.3290	0.3012	0.2535	0.3541	0.3610
Root enlargement	0.4128	0.5405	0.5144	0.4625	0.5671	0.5750
Average	0.3873	0.5114	0.4854	0.4355	0.5376	0.5452

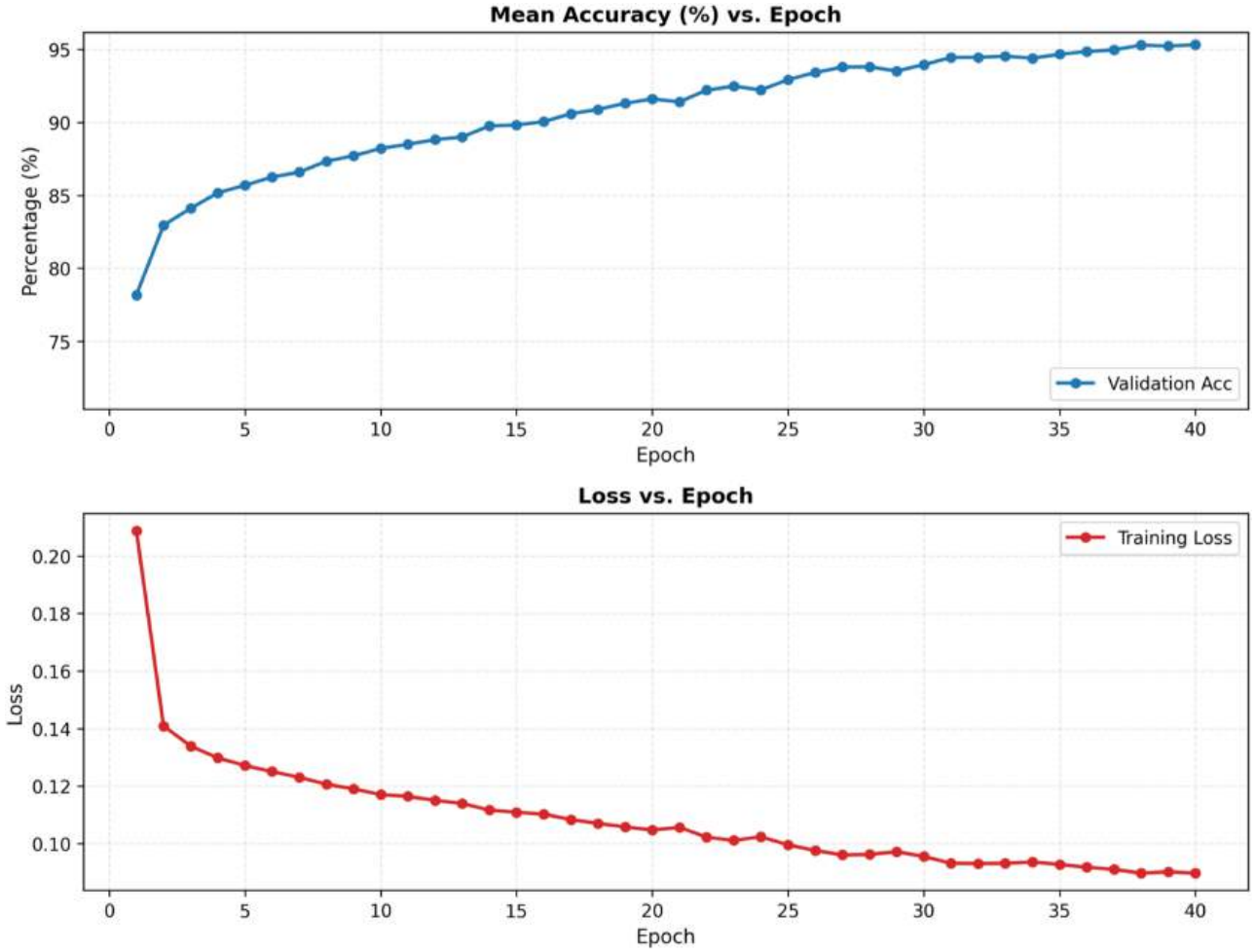

Fig. 7. Training and validation performance of the CropMap framework across 40 epochs, presenting validation accuracy (top) and training loss (bottom).

the highest IoU (0.6354), attributable to its distinct canopy structure and pronounced spectral characteristics. In contrast, the leaf expansion stage demonstrates the lowest IoU (0.3541), as it represents a transitional phenological period with limited morphological differentiation from adjacent seedling and rosette phases. These findings are consistent with biological observations, where rapid vegetative transitions result in subtle spectral changes that pose challenges for hierarchical modeling.

The per-class results presented in Table 7 indicate that the peak validation accuracy of approximately 95% is largely attributable to robust background classification (IoU = 0.9128; Precision = 0.9331). These findings demonstrate that CropMap effectively differentiates vegetation from background materials commonly present in field imagery, such as bare soil, plastic mulch, and weeds, providing a basis for finer-grained growth-stage delineation. By initially stabilizing canopy extraction, the framework minimizes environmental clutter and allows subsequent hi-

Table 7
CropMap performance for all classes in the test set.

Class	IoU	Precision	Recall
Background	0.9128	0.9331	0.9767
Planting	0.5436	0.7791	0.7017
Rosette	0.6354	0.8901	0.7479
Heading	0.5908	0.7041	0.8739
Seedling	0.5346	0.7910	0.6763
Leaf expansion	0.3541	0.5218	0.5678
Root enlargement	0.5671	0.7765	0.7337

Table 8
Component-wise loss ablation demonstrates the contribution of hierarchical learning. Results are reported as per-class IoU and mIoU on the test set (mean ± standard deviation across three seeds). CropMap-Flat corresponds to MaskFormer with a Swin-Small backbone. Note: Adjacent-stage confusion rate = percentage of misclassified crop-stage pixels that are assigned to a phenologically adjacent class (e.g., Seedling predicted as Leaf expansion, or Rosette predicted as Planting/Heading), computed from the confusion matrix on the test set. A lower rate indicates that the model is more biologically coherent in its errors. Adjacent pairs: (Planting, Rosette), (Rosette, Heading) for cabbage; (Seedling, Leaf expansion), (Leaf expansion, Root enlargement) for radish.

Class	CropMap-Flat	CropMap- FTM	CropMap- TT	CropMap- Full
Planting	0.4820 ± 0.0184	0.4744 ± 0.0232	0.5194 ± 0.0051	0.5408 ± 0.0451
Rosette	0.5734 ± 0.0098	0.5703 ± 0.0481	0.6003 ± 0.0063	0.6308 ± 0.0085
Heading	0.5394 ± 0.0124	0.5966 ± 0.0524	0.6048 ± 0.0074	0.6047 ± 0.0051
Seedling	0.4242 ± 0.0150	0.4331 ± 0.0128	0.4976 ± 0.0043	0.5294 ± 0.0561
Leaf expansion	0.3272 ± 0.0136	0.3349 ± 0.0390	0.3471 ± 0.0179	0.3520 ± 0.0030
Root enlargement	0.4816 ± 0.0055	0.4784 ± 0.0330	0.4946 ± 0.0189	0.5715 ± 0.0061
mIoU	0.4713 ± 0.0056	0.4813 ± 0.0214	0.5106 ± 0.0045	0.5382 ± 0.0122
Adj.-stage conf. rate	64.03 ± 0.67%	56.36 ± 0.75%	55.39 ± 0.39%	53.13 ± 0.50%

erarchical constraints to focus on the subtle physiological differences between sequential growth stages.

6.3. Ablation study

Table 8 presents the component-wise ablation study that validates the hierarchical learning mechanisms in CropMap, as described in Section 5.3.1. The baseline architecture, CropMap-Flat serves as a non-hierarchical reference with a mean Intersection over Union (mIoU) of 0.4713. Incorporating only the Focal Tree-Min loss (CropMap-FTM) results in a modest improvement to 0.4813 mIoU (+0.0100, +2.12%). This improvement is primarily observed in the heading class (+0.0572), where the tree-structured loss utilizes parent-level semantic signals to resolve morphological ambiguities at the rosette-to-heading boundary. However, CropMap-FTM slightly reduces performance in early vegetative stages, such as planting and rosette, suggesting that optimization with tree-min constraints without robust embedding regularizers may introduce label-assignment conflicts when parent-class confidence is low. In contrast, applying the Tree-Triplet loss in isolation (CropMap-TT) substantially improves performance to 0.5106 mIoU (+0.0393, +8.34%), outperforming CropMap-FTM by a factor of 3.9 and delivering consistent gains across all six growth stages. This comprehensive improvement indicates that regularizing the latent space with hierarchy-aware distances provides a stronger inductive bias than classification-level constraints alone, compelling the encoder to learn biologically coherent representations that benefit all downstream decision boundaries.

CropMap-Full, which integrates both loss objectives, achieves the highest performance at 0.5382 mIoU, representing a total gain of +0.0669 (+14.20%) over the baseline. Importantly, this combined gain surpasses the linear sum of the individual components (+0.0493) by +0.0176 mIoU, indicating a strong architectural synergy. This mutual reinforcement is most evident in the seedling (+0.1052) and root enlargement (+0.0899) classes, suggesting that effective mapping of distinct developmental subtrees requires both structurally coherent embeddings and hierarchically constrained label assignments.

The adjacent-stage confusion rate offers explicit empirical evidence of this biological coherence. While CropMap-Flat confines 64.03% of its misclassifications to phenologically adjacent stages, this rate decreases progressively across the variants to 56.36% (FTM), 55.39% (TT), and 53.13% (Full). This downward trend, which coincides with improvements in mIoU, confirms that tree-structured losses are the primary factor driving this improvement rather than differences in backbone architecture or data processing.

6.4. Visualization of crop mapping using the CropMap framework

Figs. 8 and 9 present representative segmentation outputs of CropMap for cabbage and radish fields, respectively. For Chinese cabbage (Fig. 8), the framework accurately captures phenological progression across three distinct stages: sparse seedling establishment during planting, rosette formation with expanding leaf area, and dense head development at heading maturity. The predicted masks exhibit reasonable spatial correspondence with ground-truth annotations, preserving row-wise field structure and canopy continuity. Some over-segmentation is observed at rosette-stage boundaries where individual plants begin to merge, a transition that is also reflected in the moderate IoU (0.6354) for this class. Despite this, the predictions remain largely coherent with the ground-truth spatial structure.

For radish (Fig. 9), CropMap demonstrates adaptation to distinct morphological dynamics. During the seedling stage, the model accurately identifies sparse, spatially isolated vegetation clusters against complex soil backgrounds. At the root enlargement stage, it generates spatially contiguous masks corresponding to fully developed canopies. The leaf expansion stage represents the most challenging class (IoU = 0.3541; Precision = 0.5218; Recall = 0.5678) and constitutes the primary failure mode of the framework, and its low accuracy limits the operational utility of radish growth-stage mapping at the current performance level. This performance limitation arises from three interrelated factors. First, pronounced class imbalance restricts the supervisory signal: leaf expansion is the second-rarest category (34,695 training patches), with 3.1 times fewer samples than the dominant rosette stage (107,305 patches). Second, this stage occupies a narrow and transient phenological window. Defined by a single morphological threshold (approximately 15 erect leaves), it serves as a transitional phase between seedling establishment and root enlargement, exhibiting minimal structural divergence from either adjacent stage. This phenological continuity introduces inherent label ambiguity, which constrains the establishment of discrete classification boundaries. Third, spectral discriminability is limited during this phase. Rapid canopy expansion and incomplete chlorophyll maturation produce Green–NIR–Red-Edge reflectance profiles that substantially overlap with both the preceding seedling stage (characterized by lower biomass but similar architecture) and the subsequent root enlargement stage (characterized by higher LAI and elevated NIR reflectance). As a result, model predictions demonstrate low confidence along transitional canopy boundaries, leading to spatial fragmentation and systematic confusion with adjacent stages. The low precision reflects commission errors caused by spectral leakage from neighboring classes, while the reduced recall indicates omission errors where true

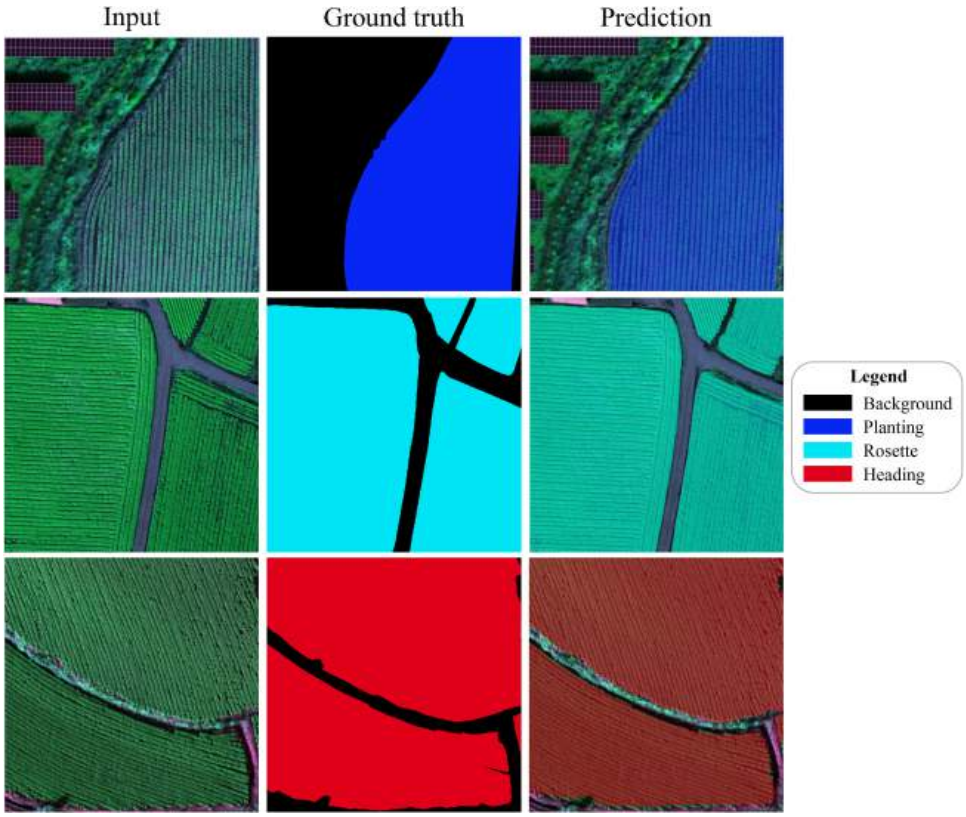

Fig. 8. Predicted outputs of the CropMap framework across the three distinct growth stages of cabbage.

leaf expansion pixels are misclassified into adjacent categories. These error patterns underscore the fundamental challenge of categorizing a continuous biological trajectory using single-date multispectral observations.

6.5. *Benchmarking CropMap against semantic segmentation baselines*

CropMap was systematically evaluated against four established semantic segmentation architectures: U-Net [9], DeepLabv3 [8], SegFormer [40], and MaskFormer [39]. All baseline models were implemented using the MMSegmentation framework [41] with identical training protocols, data splits, and augmentation policies to ensure comparability. Test set performance is presented in Table 9.

As indicated in Table 9, CropMap achieves superior results across all evaluated metrics. Conventional convolutional architectures, including U-Net and DeepLabv3, demonstrate limited effectiveness in distinguishing fine-grained growth-stage differences, with IoU scores of 0.4350 and 0.4682, respectively. Transformer-based models, such as MaskFormer and SegFormer, enhance performance by utilizing long-range contextual information, attaining IoUs of 0.5055 and 0.5124. CropMap outperforms these models, achieving the highest IoU (0.5382). These improvements suggest that hierarchical semantic constraints effectively reduce inter-class confusion among biologically similar and spatially adjacent stages.

The integration of the hierarchical framework introduces a minor trade-off in computational complexity and efficiency. CropMap exhibits the highest parameter footprint at 56.3 M, which constitutes an increase of approximately 4.4% over SegFormer (53.9 M) and 1.5% over DeepLabv3 (55.8 M). This modest increase in parameters results in an inference latency of 79.4 ms/patch. Although this execution time is slower than that of lighter convolutional architectures such as U-Net (48.3 ms/patch) and DeepLabv3 (62.7 ms/patch), CropMap remains highly competitive relative to other transformer-based approaches. Specifically, CropMap achieves a 23.2% reduction in latency compared to MaskFormer (103.5 ms/patch) while also delivering a substantial improvement in segmentation accuracy. These results indicate that CropMap effectively balances parameter density with structural constraint processing, supporting operational feasibility for large-scale field mapping workflows where accuracy is critical.

Fig. 10 presents a qualitative comparison of CropMap and transformer-based baselines across three representative field samples. CropMap generates masks that exhibit greater spatial coherence and closer alignment with the ground truth, characterized by cleaner boundaries and reduced noise. In contrast, SegFormer and MaskFormer often yield spatially fragmented predictions, particularly within seedling regions (magenta) and along phenological transition boundaries. White arrows indicate representative areas where these baselines produce spatially incoherent or misclassified clusters. Faint tile-boundary seams observed in the prediction maps result from sliding-window inference, in which the orthomosaic is processed as independent 480×480 pixel tiles. Although these edge discontinuities do not influence pixel-wise evaluation metrics, they may introduce minor visual artifacts in full-field deployments. Overall, these visual patterns support the quantitative metrics and demonstrate that CropMap's hierarchy-aware learning strategy reduces inter-class ambiguity and promotes spatially consistent predictions across subtle phenological transitions.

7. Discussion

CropMap advances UAV-based phenological mapping by addressing a more complex task than prior studies, which have primarily focused on single crop segmentation [12,15], single-crop RGB analysis [45], or lower temporal resolutions [46]. In contrast, CropMap enables simultaneous pixel-wise segmentation of six growth stages (three per crop) across multiple species, agroclimatic zones, and growing seasons using high-resolution (8 cm GSD) multi-spectral imagery. The observed performance gap between CropMap and general-purpose baselines such as

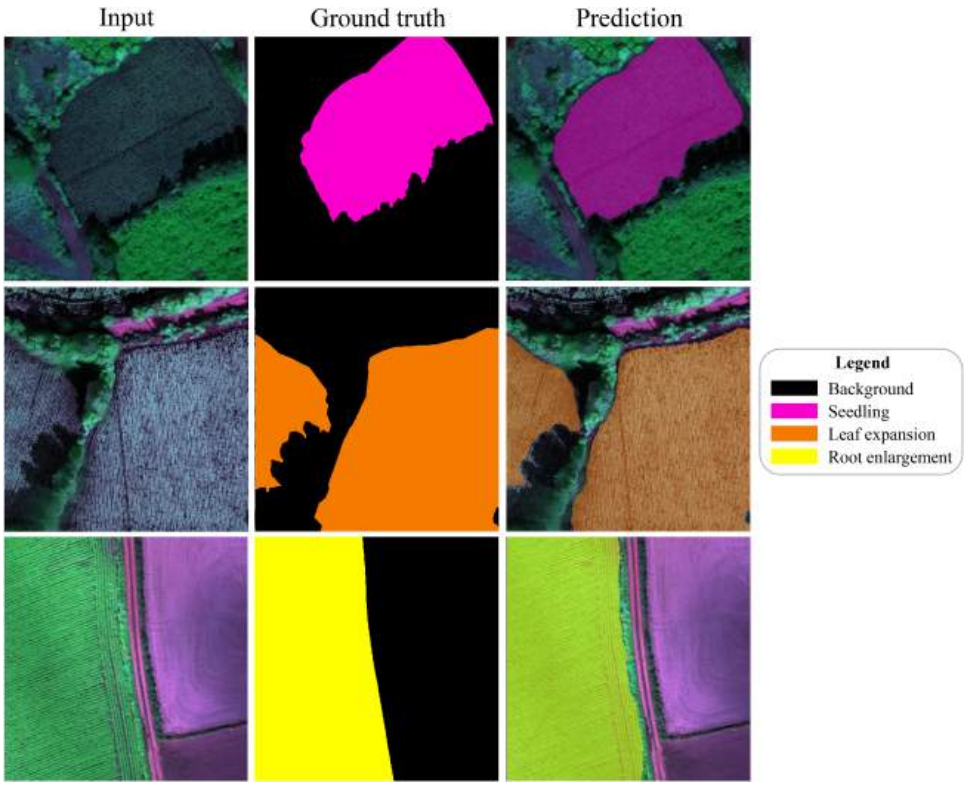

Fig. 9. Predicted outputs of the CropMap framework across the three distinct growth stages of radish.

Table 9
Quantitative comparison of segmentation metrics on the test set for CropMap and baseline architectures, excluding the background class. Results are reported as mean and standard deviation across three independent training runs with different random seeds.

Model	Params (M)	IoU	Precision	Recall	Inference time (ms/patch)
U-Net [9]	52.4	0.4350±0.018	0.6521±0.015	0.6234±0.017	48.3±2.1
DeepLabv3 [8]	55.8	0.4682±0.016	0.6844±0.013	0.6510±0.015	62.7±3.4
MaskFormer [39]	55.1	0.5055±0.014	0.7102±0.011	0.6821±0.013	103.5±5.2
SegFormer [40]	53.9	0.5124±0.013	0.7215±0.010	0.6940±0.012	71.8±3.8
CropMap (HSSN)	56.3	0.5382±0.011	0.7368±0.009	0.7021±0.010	79.4±4.1

SegFormer and MaskFormer reveals a key limitation: treating phenological development as discrete, independent categories often results in fragmented masks and boundary artifacts when models encounter the spectral and morphological continuity of adjacent stages. By explicitly encoding developmental progression through a hierarchical semantic structure, CropMap aligns decision boundaries with biological processes, reducing inter-class ambiguity and enforcing spatial coherence. This hierarchy-aware approach achieves an mIoU of 0.5382, which is competitive given the inherent spectral overlap of transitional phenology, and matches or exceeds performance reported for simpler binary or single-trait agricultural tasks. These results indicate that integrating physiological continuity priors into DL architectures provides a more robust and scalable solution for operational, field-scale crop monitoring.

The experimental results underscore the importance of advanced multispectral features and hierarchical spatial modeling for fine-grained phenological monitoring. The ablation study (Table 6) shows that standard RGB imagery is inadequate for distinguishing subtle growth phases, while the addition of the RE and N bands significantly improves class-wise IoU. Consistent with previous research [47,48], the RE channel demonstrates high sensitivity to changes in chlorophyll concentration and LAI, which is essential for segmenting the challenging leaf expansion stage. Combining these spectral signatures with strong background isolation (0.9128 IoU) enables the model to reduce environmental noise and concentrate on complex stage transitions. This approach effectively minimizes interference from plastic mulch or weeds, allowing the model to allocate its representational capacity to the biologically complex transitions between sequential growth stages.

The leaf expansion class constitutes the primary limitation of the model, achieving the lowest performance metrics (IoU: 0.3541) and lagging behind the rosette class by 0.2813. The F1-score for leaf expansion (0.5438) is the lowest of any crop stage and approximately 0.27 lower than that of the rosette stage (F1 = 0.8128), confirming that both precision and recall failures compound at this class. Two factors are the dominant causes of this limitation and are structurally difficult to resolve without targeted data collection and methodological advances.

- Class imbalance: leaf expansion comprises only 34,695 training patches (10.3% of the total), representing 3.1× fewer samples than the dominant rosette class (107,305 patches, 31.8%). Despite focal modulation in the Focal Tree-Min loss, which is designed to up-weight hard examples, the absolute scarcity of leaf expansion patches limits the amount of discriminative feature learning the model can achieve.
- Short phenological duration: the NIA guidelines define leaf expansion by a single morphological criterion, approximately 15 erect leaf stalks, corresponding to a developmental window of approximately

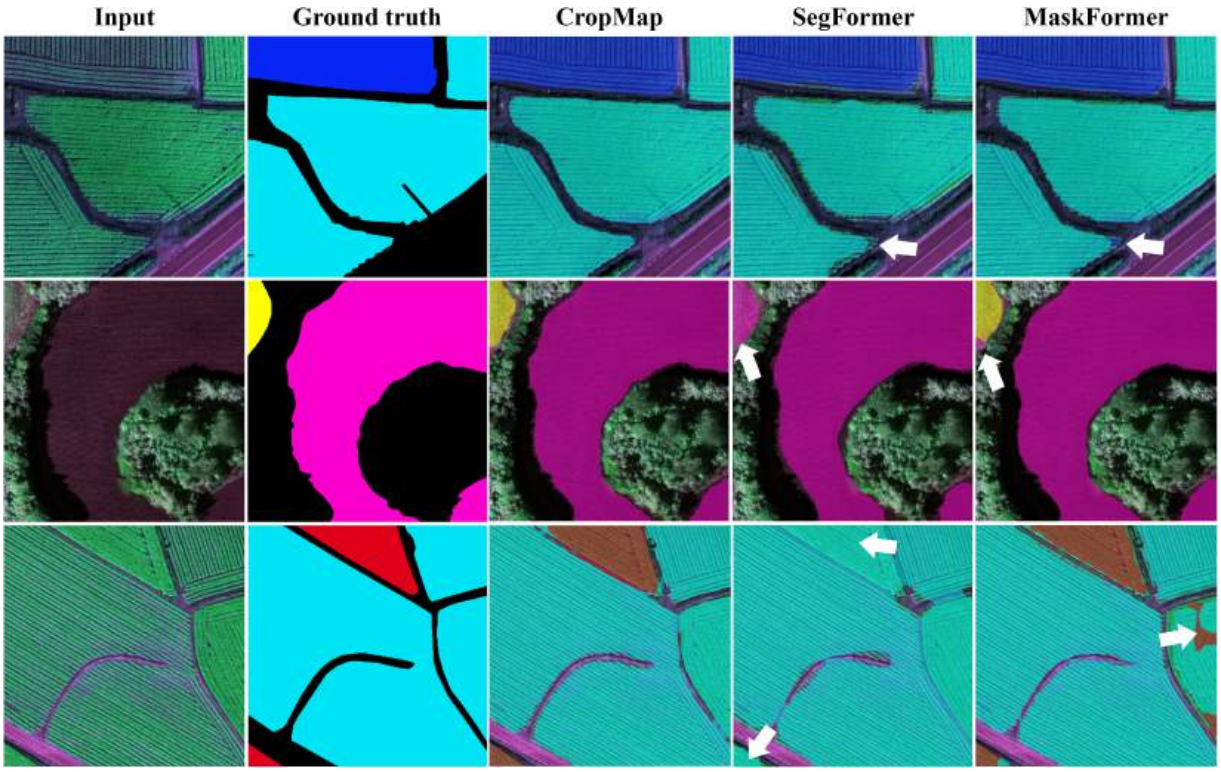

Fig. 10. Qualitative comparison of CropMap (HSSN) and transformer-based baselines. Columns from left to right: Input (pseudo-RGB), Ground Truth, CropMap, SegFormer, and MaskFormer. Note: White arrows indicate boundary errors and artifacts present in the baseline models. Square-shaped tile seams are due to independent 480×480 patch-based inference. Seedling (magenta), planting (blue), rosette (cyan), heading (red), root enlargement (yellow), and background (black) are color-coded.

7–10 days under typical autumn conditions. With bi-weekly UAV revisit frequency, some fields may be captured only once or missed entirely during this window before transitioning to root enlargement. This means the collected leaf expansion patches underrepresent the true phenological diversity of the stage, and many patches may capture transitional states between seedling and leaf expansion rather than the clearly defined stage center.

As a result, the reported mIoU of 0.5376 should be interpreted as reflecting strong performance on background separation (IoU = 0.9128) and well-represented vegetative stages (rosette IoU = 0.6354), rather than indicating uniform readiness for operational deployment across all six growth stages.

In addition, several limitations persist despite these advances. First, the dataset encompasses only a single growing season (2022), which restricts the assessment of inter-annual variability in phenological trajectories, spectral responses, and agroclimatic conditions. As a result, model robustness may decline under atypical weather patterns or altered planting schedules. Second, the geographic coverage is limited to South Korean production zones, where local cultivars and management practices may not adequately represent broader East Asian or global contexts, potentially limiting cross-regional applicability. Third, the phenological taxonomy is reduced to three stages per crop; although this is operationally efficient, it may omit transitional substages that are essential for precision interventions such as thinning or targeted fertilization. Fourth, the framework was developed solely using high-resolution (8 cm GSD) UAV imagery, leaving its applicability to regional-scale deployment with lower-resolution satellite platforms (e.g., PlanetScope at 3 m, Sentinel-2 at 10 m) untested. Finally, the evaluation was based exclusively on internal splits of the NIA dataset, which may not accurately represent the framework's performance under substantial domain shifts.

8. Conclusion and future works

This study introduced and validated CropMap, a hierarchical semantic segmentation framework developed to address the challenges of fine-grained crop growth-stage mapping. Integration of Hierarchical Semantic Segmentation Networks (HSSN) with multispectral UAV imagery demonstrated that organizing labels according to biological development substantially reduces inter-class confusion between adjacent phenological stages. Experimental results confirmed that a three-band composite (Green, Near-Infrared, and Red-Edge) provides the best-performing spectral configuration among those evaluated for phenological monitoring. This performance exceeds that of traditional convolutional neural network (CNN)-based architectures and contemporary transformer models (e.g., SegFormer, MaskFormer), which often display fragmentation and boundary artifacts when processing the subtle, spatially overlapping transitions typical of early developmental phases such as seedling establishment and leaf expansion.

Future work will address the existing limitations through multi-year and cross-regional data collection encompassing greater environmental variability. In addition, efforts will focus on refining the phenological taxonomy for substage-level decision-making and systematically assessing cross-sensor domain adaptation strategies to bridge the gap between UAV-scale precision and satellite-based operational monitoring.

CRediT authorship contribution statement

L. Minh Dang: Writing – review & editing, Writing – original draft, Methodology; **Sufyan Danish:** Visualization, Data curation; **Kyungbok Min:** Visualization, Formal analysis; **Gul E. Arzu:** Data curation, Conceptualization; **Lilia Tightiz:** Visualization, Validation; **Han Yong Park:** Methodology, Funding acquisition; **Hyoung-Kyu Song:** Supervi-

sion, Funding acquisition; **Hyeonjoon Moon:** Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research was supported by the "Regional Innovation System & Education (RISE)" through the Seoul RISE Center, funded by the Ministry of Education (MOE) and the Seoul Metropolitan Government. (2026-RISE-01-019-04) and by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03038540) and by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries(IPET) through Technology Commercialization Support Program, funded by Ministry of Agriculture, Food and Rural Affairs(MAFRA)(RS-2025-02218444).

Data availability

Data will be made available on request.

References

[1] Food and Nations, Global agriculture towards 2050, 2026 https://www.fao.org/fileadmin/user_upload/lon/HLEF2050_Global_Agriculture.pdf, accessed 2026-01-03.
[2] J. Kim, H. Park, B. Moon, S. Kim, Effect of fermentation conditions on functional quality of napa cabbage kimchi, Foods 14 (16) (2025) 2826.
[3] A. Bouguettaya, H. Zarzour, A. Kechida, A.M. Taberkit, Deep learning techniques to classify agricultural crops through UAV imagery: a review, Neural Comput. Appl. 34 (12) (2022) 9511–9536.
[4] P. Sadeghi-Tehran, K. Sabermanesh, N. Virlet, M.J. Hawkesford, Automated method to determine two critical growth stages of wheat: heading and flowering, Front. Plant Sci. 8 (2017) 252.
[5] S. Rasti, C.J. Bleakley, G.C.M. Silvestre, N.M. Holden, D. Langton, G.M.P. O'Hare, Crop growth stage estimation prior to canopy closure using deep learning algorithms, Neural Comput. Appl. 33 (5) (2021) 1733–1743.
[6] H. Wang, Y. Li, Y. Zhang, J. Shang, G. Li, L. Dinh-Tien, L.M. Dang, H.-K. Song, H. Moon, Drone-based high-precision object detection in remote sensing with attention-guided feature fusion, Tsinghua Sci. Technol. 31 (2) (2026) 1263–1281.
[7] L.M. Dang, A.S. Sagar, N.D. Bui, L.V. Nguyen, T.-H. Nguyen, Attention-guided marine debris detection with an enhanced transformer framework using drone imagery, Process Saf. Environ. Prot. 197 (2025) 107089.
[8] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, (2017). arXiv preprint arXiv:1706.05587
[9] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-assisted Intervention, Springer, 2015, pp. 234–241.
[10] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[11] N. Lüling, D. Reiser, J. Straub, A. Stana, H.W. Griepentrog, Fruit volume and leaf-area determination of cabbage by a neural-network-based instance segmentation for different growth stages, Sensors 23 (1) (2022) 129.
[12] Y. Yokoyama, T. Matsui, T.S.T. Tanaka, An instance segmentation dataset of cabbages over the whole growing season for UAV imagery, Data Br. 55 (2024) 110699.
[13] S.T. Arab, A. Takezaki, M. Kogoshi, Y. Nakano, S. Kikuchi, K. Tanaka, K. Hayashi, Integrating UAV-derived diameter estimations and machine learning for precision cabbage yield mapping, Sensors 25 (18) (2025) 5652.
[14] Z. Ye, K. Yang, Y. Lin, S. Guo, Y. Sun, X. Chen, R. Lai, H. Zhang, A comparison between pixel-based deep learning and object-based image analysis (OBIA) for individual detection of cabbage plants based on UAV visible-light images, Comput. Electron. Agric. 209 (2023) 107822.
[15] X. Yue, K. Qi, X. Na, Y. Zhang, Y. Liu, C. Liu, Improved YOLOv8-Seg network for instance segmentation of healthy and diseased tomato plants in the growth stage, Agriculture 13 (8) (2023) 1643.
[16] L. Li, T. Zhou, W. Wang, J. Li, Y. Yang, Deep hierarchical semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1246–1257.
[17] W. Wang, T. Zhou, S. Qi, J. Shen, S.-C. Zhu, Hierarchical human semantic parsing with comprehensive part-relation modeling, IEEE Trans. Pattern Anal. Mach. Intell. 44 (7) (2021) 3508–3522.
[18] W. Wang, H. Zhu, J. Dai, Y. Pang, J. Shen, L. Shao, Hierarchical human parsing with typed part-relation reasoning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8929–8939.
[19] K. Gong, Y. Gao, X. Liang, X. Shen, M. Wang, L. Lin, Graphonomy: universal human parsing via graph transfer learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7450–7459.
[20] A.R. Allu, S. Mesapam, Impact of remote sensing data fusion on agriculture applications: a review, Eur. J. Agron. 164 (2025) 127478.
[21] M. Saki, R. Keshavarz, D. Franklin, M. Abolhasan, J. Lipman, N. Shariati, A data-driven review of remote sensing-based data fusion in precision agriculture from foundational to transformer-based techniques, IEEE Access 13(2025) 166188–166209.
[22] W.H. Maes, K. Steppe, Perspectives for remote sensing with unmanned aerial vehicles in precision agriculture, Trends Plant Sci. 24 (2) (2019) 152–164.
[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[24] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al., Swin transformer v2: scaling up capacity and resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12009–12019.
[25] L.T. Ramos, A.D. Sappa, Multispectral semantic segmentation for land cover classification: an overview, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 17 (2024) 14295–14336.
[26] X. Hu, S. Chen, D. Zhang, Domain adaptation in agricultural image analysis: a comprehensive review from shallow models to deep learning, (2025) arXiv e-prints arXiv–2506.
[27] H. Wang, Y. Yao, J. Liu, X. Zhang, Y. Zhao, S. Li, Z. Liu, X. Zhang, Y. Zeng, Unsupervised cross-regional and cross-year adaptation by climate indicator discrepancy for crop classification, J. Remote Sens. 5 (2025) 0439.
[28] S. Peng, L. Zhang, R. Xie, Y. Qu, CSTN: A cross-region crop mapping method integrating self-training and contrastive domain adaptation, Int. J. Appl. Earth Obs. Geoinf. 136 (2025) 104379.
[29] T. Qiu, C. Song, J.S. Clark, B. Seyednasrollah, N. Rathnayaka, J. Li, Understanding the continuous phenological development at daily time step with a Bayesian hierarchical space-time model: impacts of climate change and extreme weather events, Remote Sens. Environ. 247 (2020) 111956.
[30] R.d.S. Torres, M. Hasegawa, S. Tabbone, J. Almeida, J.A. dos Santos, B. Alberton, L.P.C. Morellato, Shape-based time series analysis for remote phenology studies, in: 2013 IEEE International Geoscience and Remote Sensing Symposium-IGARSS, Ieee, 2013, pp. 3598–3601.
[31] W. Liu, M. Mõttus, J.-P. Gastellu-Etchegorry, H. Fang, J. Atherton, Seasonal and vertical variation in canopy structure and leaf spectral properties determine the canopy reflectance of a rice field, Agric. for. Meteorol. 355 (2024) 110132.
[32] C. Szymon, K. Szuniewicz, K. Kowalczyk, A. Dumalski, M. Ogrodniczak, Ł. Zieleniewicz, Assessment of accuracy in unmanned aerial vehicle (uav) pose estimation with the real-time kinematic (rtk) method on the example of dji matrice 300 rtk, Sensors 23 (4) (2023) 2092.
[33] S. Zsebő, L. Bede, G. Kukorelli, I.M. Kulmány, G. Milics, D. Stencinger, G. Teschner, Z. Varga, V. Vona, A.J. Kovács, Yield prediction using NDVI values from GreenSeeker and MicaSense cameras at different stages of winter wheat phenology, Drones 8 (3) (2024) 88.
[34] D. Tyagi, V. Mishra, H. Verma, Agisoft metashape, and Pix4Dmapper UAV photogrammetry software, Proc. UASG 2021: Wings 4 Sustain.: Unmanned Aer. Syst. Geomat. (2023) 121.
[35] Y. Kang, Q. Meng, M. Liu, Y. Zou, X. Wang, Crop classification based on red edge features analysis of GF-6 WFV data, Sensors 21 (13) (2021) 4328.
[36] S. El-Hendawy, Y.H. Dewir, S. Elsayed, U. Schmidhalter, K. Al-Gaadi, E. Tola, Y. Refay, M.U. Tahir, W.M. Hassan, Combining hyperspectral reflectance indices and multivariate analysis to estimate different units of chlorophyll content of spring wheat under salinity conditions, Plants 11 (3) (2022) 456.
[37] M. Dang, H. Wang, Y. Li, T.-H. Nguyen, L. Tightiz, N. Xuan-Mung, T.N. Nguyen, Computer vision for plant disease recognition: a comprehensive review, Bot. Rev. 90 (3) (2024) 251–311.
[38] D.-W. Kim, G. Jang, H.-J. Kim, Development of CNN-based semantic segmentation algorithm for crop classification of Korean major upland crops using NIA AI HUB, IEEE Access 13 (2025) 8425–8438.
[39] B. Cheng, A. Schwing, A. Kirillov, Per-pixel classification is not all you need for semantic segmentation, Adv. Neural Inf. Process. Syst. 34 (2021) 17864–17875.
[40] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 12077–12090.
[41] M. Contributors, MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. 2020, 2023, ().
[42] G.-H. Kwak, N.-W. Park, Impact of texture information on crop classification with machine learning and UAV images, Appl. Sci. 9 (4) (2019) 643.
[43] J.D. Stamford, S. Vialet-Chabrand, I. Cameron, T. Lawson, Development of an accurate low cost NDVI imaging system for assessing plant health, Plant Methods 19 (1) (2023) 9.
[44] M. Liu, Y. Zhan, J. Li, Y. Kang, X. Sun, X. Gu, X. Wei, C. Wang, L. Li, H. Gao, et al., Validation of red-edge vegetation indices in vegetation classification in tropical monsoon region-a case study in Wenchang, Hainan, China, Remote Sens. 16 (11) (2024) 1865.
[45] N. Alam, A.S. Sagar, L.M. Dang, W. Zhang, H.Y. Park, M. Hyeonjoon, Deep learning based radish and leaf segmentation for phenotype trait measurement, Signal Image

Video Process. 19 (1) (2025) 178.
[46] H. Zheng, X. Zhou, J. He, X. Yao, T. Cheng, Y. Zhu, W. Cao, Y. Tian, Early season detection of rice plants using RGB, NIR-GB and multispectral images from unmanned aerial vehicle (UAV), Comput. Electron. Agric. 169 (2020) 105223.
[47] Y. Sun, Q. Qin, Y. Zhang, H. Ren, G. Han, Z. Zhang, T. Zhang, B. Wang, A leaf chlorophyll vegetation index with reduced LAI effect based on sentinel-2 multispectral red-edge information, Comput. Electron. Agric. 236 (2025) 110500.
[48] S. Gao, K. Yan, J. Liu, J. Pu, D. Zou, J. Qi, X. Mu, G. Yan, Assessment of remote-sensed vegetation indices for estimating forest chlorophyll concentration, Ecol. Indic. 162 (2024) 112001.