# Comparative Analysis of End-to-End Architectures versus Multi-Stage Pipelines for Agricultural Visual Recognition in 2026

## 1. Introduction: The Complexity of Pathological Plant Recognition

As the agricultural sector advances toward the widespread adoption of Agriculture 4.0, the role of computer vision has shifted from passive monitoring to active, real-time intervention. By 2026, the demand for automated systems capable of precise laser weeding, robotic harvesting, and targeted pesticide application has necessitated a rigorous re-evaluation of neural network architectures. The core challenge in this domain is no longer merely distinguishing a crop from a weed in ideal conditions; rather, it is the robust identification of specific plant species and distinct morphological parts (stems, leaves, fruits, panicles) under the chaotic visual conditions induced by disease.[1]

Pathological conditions introduce a layer of visual corruption that fundamentally challenges traditional feature extraction. A healthy tomato leaf possesses a predictable geometry and texture—serrated edges, a specific shade of green, and a planar structure. However, a leaf infected with *Tomato Yellow Leaf Curl Virus* (TYLCV) exhibits severe upward curling, chlorosis (yellowing), and size reduction, effectively altering the topological features that standard Convolutional Neural Networks (CNNs) rely upon for organ identification.[3] Similarly, fungal infections like *Botrytis* can cause fruits to shrivel, obscuring the curvature and specular reflections typically used by detectors to locate harvestable produce. When these morphological deformations are compounded by occlusion from dense canopies and variable field lighting, the task of identifying "a tomato stem" becomes a complex reasoning problem rather than a simple pattern matching task.[5]

This report provides an exhaustive analysis of the two dominant architectural paradigms vying for dominance in this space as of 2026: **Simpler End-to-End Models** and **Multi-Stage Pipelines**.

The "End-to-End" category has seen a renaissance with the release of **YOLO26**, **YOLOv12**, and **YOLOv11**, alongside specialized lightweight architectures like **ALNet** and **PMJDM**. These models strive to map raw pixel data directly to class labels and bounding boxes in a single forward pass, optimizing for the millisecond-level latency required by edge robotics.[7] They represent the philosophy of "efficiency through integration," attempting to learn disease-invariant representations of plant parts through massive supervised training and advanced architectural priors like attention mechanisms.[9]

Conversely, "Multi-Stage Pipelines" represent a composite approach, chaining together Foundation Models such as **Grounding DINO** (for open-set detection), **Segment Anything Model 2 (SAM-2)** (for zero-shot segmentation), and **Vision-Language Models (VLMs)** like **Qwen2.5-VL**, **Agri-R1**, or **BioCLIP 2** (for semantic reasoning). These systems prioritize interpretability and zero-shot generalization over raw speed, leveraging the emergent properties of large-scale pre-training to handle the "long tail" of rare diseases and morphological anomalies that supervised detectors miss.[3]

Through a detailed synthesis of benchmarks from **AgMMU**, **PlantDoc**, and **PhenoBench**, this report evaluates the trade-offs between these paradigms. It explores whether the raw inference speed of the latest YOLO variants can be reconciled with the semantic robustness of VLMs, and how industry leaders like Carbon Robotics are navigating this divide with proprietary "Large Plant Models".[12]

# 2. The Evolution of End-to-End Object Detectors (2024–2026)

The trajectory of single-stage detectors has been defined by a relentless pursuit of the Pareto frontier between accuracy and latency. In the context of diseased plant identification, this evolution has focused on three critical capabilities: detecting small, necrotic features; maintaining robustness against occlusion; and eliminating non-differentiable post-processing steps that introduce latency jitter in robotic control loops.

## 2.1 YOLOv11: The Architectural Workhorse of Precision Agriculture

Released in late 2024 and dominating the 2025 agricultural season, YOLOv11 established a new baseline for supervised plant disease detection. Unlike its general-purpose predecessors, YOLOv11 introduced specific architectural refinements that proved fortuitously effective for the high-frequency textural variations found in plant pathology.[9]

### 2.1.1 Architectural Refinements: C3K2 and C2PSA

The core innovation in YOLOv11 is the replacement of the C2f block (used in YOLOv8) with the **C3K2 (Cross-Stage Partial with Kernel 3)** block. This modification optimizes gradient flow through the network, allowing for deeper architectures that do not suffer from vanishing gradients—a critical feature when training on datasets like **PlantDoc**, where the visual difference between "Early Blight" and "Late Blight" may hinge on subtle textural nuances spanning only a few pixels.[9]

Furthermore, YOLOv11 integrates the **C2PSA (Cross-Stage Partial with Parallel Spatial Attention)** mechanism. In the context of identifying plant parts in diseased images, this attention mechanism allows the model to spatially "focus" on relevant anatomical features (e.g., the petiole or fruit calyx) while suppressing the high-frequency noise generated by background soil, mulch, or necrotic tissue debris. Benchmarking on a curated dataset of 50,000 leaf images demonstrated that this attention mechanism was directly responsible for a **3.4 percentage point increase in small-object recall** compared to YOLOv8.[9] This is particularly relevant for identifying early-stage fruiting bodies or emerging pests (like spider mites) that are often obscured by the larger, symptomatic lesions of a concurrent disease infection.

### 2.1.2 Performance on Diseased Datasets

In multi-class health monitoring tasks (classifying regions as Healthy, Stressed, or Damaged), YOLOv11 demonstrated superior localization fidelity. Qualitative analysis reveals that while YOLOv8 often generated loose bounding boxes that included healthy background tissue—thereby diluting the classification signal—YOLOv11 produced tighter boxes that closely adhered to the boundaries of the necrotic lesions or the specific plant organ. On a standardized test set, YOLOv11 achieved a mean Average Precision (**mAP@0.5**) of **93.3%**, significantly outperforming the 92.0% achieved by YOLOv8, and maintained a real-time inference speed of **15 ms per image** on an NVIDIA RTX 3060.[9]

## 2.2 DMN-YOLO: Specialized Enhancements for Pathological Occlusion

While standard YOLOv11 offers robust performance, specific agricultural variants like **DMN-YOLO** have been developed to address the unique challenge of "dense lesions"—scenarios where a single leaf is covered in hundreds of small, overlapping disease spots (e.g., Apple Frogeye Leaf Spot).[5]

### 2.2.1 Multi-Scale Perception and Loss Functions

DMN-YOLO integrates a **Multi-branch Auxiliary Feature Pyramid Network (MAFPN)** to enhance the model's ability to perceive multi-scale lesions. Disease symptoms on a single plant can range from microscopic fungal spores to whole-leaf chlorosis. The MAFPN utilizes a **Superficial Assisted Fusion (SAF)** mechanism to preserve shallow semantic cues (texture, color) and an **Advanced Auxiliary Fusion (AAF)** module to enhance top-down gradient flow.[5]

Crucially, DMN-YOLO replaces the standard CIoU loss with a **Normalized Wasserstein Distance (NWD)** loss function. Traditional Intersection-over-Union (IoU) metrics fail when detecting very small, non-overlapping objects (like tiny necrotic spots). The NWD loss models the bounding boxes as 2D Gaussian distributions, allowing the network to measure the similarity between predicted and ground-truth boxes even when they do not strictly overlap. This modification resulted in a **5.0% improvement in mAP@0.5** for apple leaf diseases compared to the baseline YOLOv11n model, demonstrating the necessity of specialized loss functions for granular plant part identification.[5]

## 2.3 YOLOv12: The Attention-Centric Shift

YOLOv12 represents a divergent evolutionary path, moving away from pure CNN backbones toward an "Attention-Centric" design. While CNNs excel at local texture recognition, they often struggle with long-range dependencies—for example, associating a withered tip of a leaf with the specific stem it belongs to, across a cluttered canopy.[15]

YOLOv12 integrates transformer-like attention mechanisms directly into the backbone, theoretically allowing it to model the structural connectivity of the plant. However, benchmarks indicate a complex trade-off. While YOLOv12 theoretically offers higher accuracy for complex geometric reasoning, it often requires specific hardware acceleration (like FlashAttention) to match the inference speeds of its predecessors. In resource-constrained agricultural environments (e.g., drone-based scouting), the computational overhead of calculating global attention maps can render YOLOv12 less efficient than highly optimized CNNs like YOLOv11 or YOLO26.[16]

## 2.4 YOLO26: The Paradigm Shift to NMS-Free Inference

Released in January 2026, **YOLO26** marks the most significant architectural shift in the series, addressing the primary bottleneck for real-time agricultural robotics: **latency jitter** caused by Non-Maximum Suppression (NMS).[7]

### 2.4.1 The NMS Problem in Agriculture

In a typical high-density field scenario (e.g., a spinach field with weed pressure), a standard detector might predict thousands of overlapping bounding boxes. The NMS algorithm, which filters these duplicates, is a CPU-bound, heuristic process. Its execution time scales with the number of detections, making it non-deterministic. For a robotic weeder moving at 5 mph, a 50ms variance in processing time can result in the laser missing the weed and striking the crop.

### 2.4.2 Native End-to-End Architecture

YOLO26 eliminates NMS entirely by adopting a **native end-to-end** design. It utilizes a "One-to-One" matching strategy during training (similar to DETR models but fully convolutional), ensuring that the network outputs exactly one high-confidence box per object. This results in deterministic latency, a critical requirement for control loops in autonomous machinery.[8]

### 2.4.3 MuSGD and Small Object Optimization

YOLO26 also introduces the **MuSGD Optimizer**, a hybrid optimization strategy inspired by Large Language Model training (specifically Moonshot AI's Kimi K2). This provides greater stability when training on massive, noisy agricultural datasets (such as the 150-million-image dataset used by Carbon Robotics). Furthermore, the integration of **Progressive Loss**

**(ProgLoss)** and **Self-Training Anchor Loss (STAL)** dynamically adjusts the loss weight for difficult examples. In field tests, this has shown specific improvements in detecting **emerging weeds** (cotyledon stage) and small plant organs (buds, flowers) that are often statistically overwhelmed by the larger foliage in standard loss functions.[18]

**Table 1: Comparative Benchmarks of End-to-End Architectures (2026 Baseline)**

| Metric | YOLO11n | YOLO11s | YOLO26n | YOLO26s |
|---|---|---|---|---|
| **mAP (val 50-95)** | 39.5 | 47.0 | **40.9** | **48.6** |
| **CPU Speed (ms)** | 56.1 | 90.0 | **38.9** | **87.2** |
| **Parameters (M)** | 2.6 | 9.4 | **2.4** | 9.5 |
| **FLOPs (B)** | 6.5 | 21.5 | **5.4** | **20.7** |
| **Post-Processing** | NMS Required | NMS Required | **None (End-to-End)** | **None (End-to-End)** |
| **Key Advantage** | Mature Ecosystem | Higher Accuracy | **Latency Stability** | **Small Object Rec.** |

Data synthesized from Ultralytics Technical Reports and Independent Benchmarks.[8]

# 3. Single-Pass Vision-Language Models (VLMs): The Reasoning Engines

While YOLO models excel at "localization" (where is the leaf?), they often lack "semantic depth" (why does the leaf look like that?). By 2026, **Vision-Language Models (VLMs)** have evolved to bridge this gap, offering the ability to perform **single-pass reasoning**—identifying a plant species, localizing its parts, and diagnosing its condition simultaneously based on a natural language prompt.

## 3.1 Qwen2.5-VL: Native Resolution and Zero-Shot Limits

**Qwen2.5-VL** represents the state-of-the-art in open-weight VLMs as of early 2026. Its defining feature is **Native Spatial Perception**, which allows the model to process images at their original resolution without the resizing or normalization steps that typically destroy fine-grained agricultural details (like spider mite webbing or early fungal spores).[20]

### 3.1.1 Zero-Shot Performance on PlantDoc

Benchmarking on the **PlantDoc** dataset reveals the strengths and critical limitations of general-purpose VLMs. In zero-shot settings (where the model is not fine-tuned on the specific dataset), **Qwen2.5-VL:7B** achieved a **fuzzy matching accuracy of 81.8%.**[22] This means the model could correctly identify the broad context (e.g., "This is a tomato leaf with a fungal infection").

However, the model struggled significantly with **exact species matching** (accuracy < 11.5%). Without specific prompting, Qwen2.5-VL often defaulted to generic terms like "Leaf" or "Plant" rather than specific taxonomic identifiers. It also exhibited hallucinations when prompted with lists of classes, occasionally inventing plausible-sounding but non-existent diseases. This behavior underscores that while general VLMs are powerful reasoning engines, they lack the rigid taxonomic grounding required for scientific plant identification without targeted fine-tuning.[22]

## 3.2 Agri-R1: Reinforcement Learning for Domain Specificity

To address the generalization limits of general-purpose models, the **Agri-R1** model was developed using **Reinforcement Learning (RL)** specifically tuned for agricultural logic. Unlike standard supervised fine-tuning, Agri-R1 utilizes a reward function that integrates domain-specific lexicons and fuzzy matching logic to penalize hallucinations and reward precise terminological usage.[23]

On the **CDDMBench** (Crop Disease Diagnosis Multimodal Benchmark), the 3B-parameter Agri-R1 model demonstrated a **+23.2% relative gain** in disease recognition accuracy compared to standard VLMs of similar size. This result is significant because it suggests that smaller, specialized models trained with domain-specific RL can outperform massive generalist models (like GPT-4o) on niche agricultural tasks, potentially offering a viable path for deploying reasoning capabilities on edge devices.[23]

## 3.3 SCOLD: Soft-Target Contrastive Learning

**SCOLD (Soft-target COntrastive learning for Leaf Disease identification)** addresses a specific problem in agricultural data: label noise. Crowdsourced datasets (like PlantVillage or iNaturalist) often contain ambiguous or incorrect labels. Standard contrastive learning (like CLIP) can overfit to these noisy labels, leading to brittle performance.[25]

SCOLD employs **task-agnostic pretraining** with **contextual soft targets**, smoothing the labels to prevent overconfidence. Developed on a corpus of 186,000 image-caption pairs covering **97 unique concepts**, SCOLD outperforms OpenAI-CLIP-L and BioCLIP on fine-grained leaf disease tasks. Its architecture is specifically optimized to map the visual features of a "diseased leaf" to the textual concept of the disease, making it highly effective for retrieval tasks (e.g., "Find all images of maize leaves with bacterial blight").[26]

### 3.4 BioCLIP 2: Hierarchical Taxonomic Embeddings

**BioCLIP 2** represents the pinnacle of "taxonomically aware" vision models. Trained on the **TreeOfLife-200M** dataset (comprising over 200 million biology images), it learns hierarchical embeddings that mirror the biological tree of life.[28]

Crucially for organ identification, BioCLIP 2 exhibits an emergent property known as **intra-species variation separation**. This means the model learns to cluster images of the same species together while maintaining distinct sub-clusters for different morphological forms (e.g., juvenile leaves vs. mature fruits vs. flowers). This allows the model to correctly identify a "withered, brown tomato stem" as belonging to *Solanum lycopersicum* because it maps the visual embedding to the persistent taxonomic concept, rather than relying solely on the green texture of a healthy plant. In zero-shot benchmarks, BioCLIP 2 outperforms general CLIP models by approximately **18%** on species classification tasks.[28]

## 4. Multi-Task Learning (MTL): The Middle Ground

Between the extremes of pure object detectors and massive VLMs lies the domain of **Multi-Task Learning (MTL) CNNs**. These models attempt to predict multiple attributes (Species, Organ, Disease) simultaneously using a shared backbone, offering a balance of efficiency and semantic richness.

### 4.1 PMJDM: Joint Detection Model

The **PlantDisease Multi-task Joint Detection Model (PMJDM)** exemplifies this approach. It utilizes a shared feature extractor (based on **ConvNeXt** or **EfficientNet**) that feeds into multiple task-specific heads:

1. **Species Classification Head:** Identifies the crop type (e.g., Potato, Tomato).
2. **Disease Detection Head:** Localizes the lesion and identifies the pathogen.
3. **Spatial Consistency Module:** Uses Conditional Random Fields (CRF) to ensure that the detected disease is spatially consistent with the plant part (e.g., preventing the model from detecting "root rot" on a leaf).[30]

Evaluated on a dataset of 26,073 images, PMJDM achieved a **61.83% mAP50**, surpassing traditional two-stage detectors like Faster R-CNN. The shared representation allows the model to leverage correlations between tasks—for instance, knowing the species helps narrow down the probability space for potential diseases (e.g., *Late Blight* is common in potatoes but not in corn), improving overall accuracy without the computational cost of running two separate models.[30]

### 4.2 ALNet: Lightweight Efficiency

**ALNet (Attentive and Lightweight Network)** focuses on extreme efficiency for edge deployment. Using a custom hybrid block design inspired by **ShuffleNet** and **EfficientNet**, it achieves high accuracy (99.78% on grapevine datasets) with only **0.17 million parameters**—approximately 18 times smaller than SqueezeNet. While it lacks the dense localization capabilities of YOLO, its ability to perform rapid classification makes it an ideal "pre-filter" in a cascaded system, determining if a frame contains a plant of interest before triggering a heavier detector.[32]

## 5. Multi-Stage Pipelines: The Composite Approach

While end-to-end models prioritize speed, **Multi-Stage Pipelines** prioritize precision and modularity. By chaining together state-of-the-art Foundation Models, these systems can perform tasks that are currently impossible for a single regression-based network, such as zero-shot segmentation of novel plant parts.

### 5.1 The Anatomy of a 2026 Pipeline

The standard composite pipeline for high-precision phenotyping typically consists of three distinct stages:

1. **Open-Set Detection (Grounding DINO):** The system receives a text prompt (e.g., "Find all tomato fruits"). Grounding DINO, an open-set detector, returns bounding boxes for these objects. This allows the system to find novel objects (e.g., a specific weed species) without any retraining.[3]
2. **Zero-Shot Segmentation (SAM-2):** The bounding boxes from the detector are used as "prompts" for the **Segment Anything Model 2 (SAM-2)**. SAM-2 generates pixel-perfect masks for the objects, effectively separating the plant part from the background. This step is crucial for diseased images, as it allows downstream classifiers to analyze only the plant tissue, ignoring soil or mulch that might confuse the model.[34]
3. **Visual Question Answering (VLM/Classifier):** The segmented image regions are passed to a VLM (like **BLIP-2** or **BioCLIP 2**) or a specialized classifier to identify the species and disease state.

### 5.2 Case Study: TLDVLM (Tomato Leaf Disease VLM)

The **TLDVLM** pipeline serves as a prime example of this architecture. By integrating **Grounding DINO** for detection and **SAM-2** for segmentation, the pipeline ensures that the visual input to the final classifier (a fine-tuned **BLIP-2** with Low-Rank Adaptation) is strictly limited to the leaf tissue.

- **Performance:** This approach achieved an accuracy of **97.27%** on tomato disease identification, significantly outperforming single-pass equivalents that struggled with background noise and variable lighting conditions.[3]

- **Explainability:** Unlike a YOLO model that outputs a class ID, the VLM component can provide a natural language explanation of the diagnosis (e.g., "The leaf shows concentric rings characteristic of Early Blight"), which is invaluable for human-in-the-loop agronomy.

### 5.3 The Cost of Modularity

The primary trade-off for this precision is computational cost and latency.

- **Latency:** Running three separate large-scale neural networks sequentially introduces significant latency. While a YOLO26 model might run at >100 FPS on a GPU, a full Grounding DINO + SAM-2 + VLM pipeline often operates at **<5 FPS**. This renders it unsuitable for real-time actuation (e.g., spraying while driving) but acceptable for "stop-and-stare" scouting robots.[18]
- **Hardware Requirements:** These pipelines typically require substantial VRAM (often >24GB for the full pipeline), necessitating server-grade GPUs or high-end edge devices like the **NVIDIA Jetson Orin AGX**, whereas YOLO models can run comfortably on a **Jetson Nano** or **Raspberry Pi**.[8]

## 6. Benchmarking and Performance Analysis

To provide a concrete comparison, we synthesize data from key 2026 benchmarks, including **AgMMU** (Agricultural Multimodal Understanding), **PlantDoc** (Object Detection in the Wild), and **PhenoBench** (Crop/Weed Segmentation).

### 6.1 Accuracy vs. Task Complexity

The **AgMMU** benchmark reveals a stark divergence in performance based on task type.

**Table 2: Performance Comparison on AgMMU Classification Tasks**

| Model Architecture | Model Type | Disease ID Accuracy | Pest/Damage ID Accuracy | Species ID Accuracy |
|---|---|---|---|---|
| **YOLO11 (Supervised)** | Single-Stage | **>90%** | **High** | **High** |
| **Gemini-3 Pro** | VLM (Multi-Choice) | ~62% | Low (<40%) | 74% |
| **GPT-4o** | VLM (Open-Ended) | ~45% | ~30% | ~50% |

| | | | | |
|---|---|---|---|---|
| **Agri-R1** | Specialized VLM | +23% vs Baseline | Moderate | High |

Data synthesized from AgMMU Benchmark Results.[37]

**Key Insight:** Supervised, task-specific models like **YOLO11** (and by extension YOLO26) consistently outperform zero-shot Foundation Models in raw accuracy. The gap is most pronounced in **Pest/Damage Identification**, a task that relies on detecting subtle, localized visual cues (e.g., a small bite mark or a tiny insect) that general-purpose VLMs often miss in favor of broader semantic scene descriptions. VLMs perform best when the task is constrained to **Multiple-Choice Question Answering (MCQA)**, where the output space is limited.[11]

### 6.2 Inference Efficiency

For deployment, the metric of choice is often **FPS per Watt**.

- **YOLO26n:** ~39ms latency on CPU. Highly efficient, enabling deployment on battery-powered edge devices.
- **YOLO11n:** ~56ms latency on CPU. Robust but slightly less efficient than the NMS-free YOLO26.
- **Composite Pipeline:** >500ms latency. Requires GPU acceleration. The energy cost per inference is orders of magnitude higher, limiting its use to high-value phenotyping or intermittent scouting rather than continuous field monitoring.[8]

### 6.3 Disease-Invariance and Generalization

While YOLO models win on speed and supervised accuracy, Multi-Stage Pipelines win on **robustness to distribution shifts**.

- **YOLO Limitation:** A YOLO model trained on green leaves will likely fail to detect a completely brown, withered leaf unless that specific morphology was heavily represented in the training set (via augmentations like Mosaic or MixUp).
- **VLM/BioCLIP Strength:** Foundation models leverage semantic concepts. BioCLIP 2 maps a "withered tomato leaf" to the same taxonomic embedding space as a "healthy tomato leaf" because it understands the biological entity *Solanum lycopersicum* across its various visual states. This makes pipelines using BioCLIP 2 significantly more robust to "weird" or anomalous disease presentations that were not present in the training data.[38]

## 7. Industry Application: The Carbon Robotics Strategy

The theoretical trade-offs discussed above are being navigated in real-time by industry leaders. **Carbon Robotics**, creators of the LaserWeeder, exemplifies the state-of-the-art industrial application of these technologies as of 2026.

### 7.1 The Large Plant Model (LPM)

Carbon Robotics utilizes a proprietary **Large Plant Model (LPM)** trained on over **150 million labeled plants** collected from 175+ field robots operating globally.

- **Architecture:** While the exact architecture is proprietary, the requirement for millisecond-level latency (to fire lasers at moving weeds) strongly suggests a highly optimized **single-stage architecture** (likely a custom variant of YOLO or a streamlined Transformer) rather than a slow multi-stage pipeline.[12]
- **The Data Flywheel:** The true "secret sauce" is not the model architecture but the **Data Flywheel**. Every robot acts as an edge node, collecting "failure cases" (e.g., a crop misidentified as a weed) and feeding them back to the central server. This allows the model to learn "disease-invariant" recognition through brute-force exposure to millions of variations of diseased and healthy plants, rather than relying on the zero-shot generalization capabilities of a VLM.[13]
- **Plant Profiles:** The system allows farmers to fine-tune the model on the fly using **Plant Profiles**, where 2-3 photos of the specific crop/weed in the current field are used to adapt the model. This effectively implements **Few-Shot Learning** at the edge, bridging the gap between general pre-training and local field conditions.[12]

## 8. Strategic Recommendations for Model Selection

Based on the comparative analysis, we propose the following decision framework for engineers and researchers selecting a model for plant identification in 2026:

### 8.1 Scenario A: Real-Time Robotic Actuation

- **Task:** Robotic weeding, precision spraying, fruit harvesting.
- **Requirement:** Deterministic latency (<50ms), high reliability.
- **Recommended Model: YOLO26 (End-to-End)**.
- **Rationale:** The NMS-free architecture guarantees consistent inference times, eliminating the jitter that causes actuation errors. The MuSGD optimizer and specific loss functions (ProgLoss) ensure high recall for small, obscured plant parts.
- **Implementation Strategy:** Fine-tune YOLO26 on a diverse, augmented dataset (using Mosaic/MixUp) that explicitly includes diseased and deformed plant examples to build robustness.[9]

### 8.2 Scenario B: Diagnostic Scouting & Phenotyping

- **Task:** Disease diagnosis, yield estimation, new pathogen discovery.
- **Requirement:** High semantic understanding, explainability, zero-shot adaptability.
- **Recommended Model: Multi-Stage Pipeline (Grounding DINO + SAM-2 + BioCLIP 2)**.
- **Rationale:** The modularity allows for the identification of novel diseases or plant parts

without retraining. SAM-2 provides the precise segmentation needed for accurate symptom analysis, and BioCLIP 2 offers robust taxonomic identification even for withered/deformed organs.
- **Implementation Strategy:** Deploy on a platform with sufficient compute (e.g., a tractor-mounted GPU server or cloud-connected drone) where latency is less critical. Use the VLM to generate natural language reports for agronomists.[29]

### 8.3 Scenario C: Resource-Constrained IoT Monitoring

- **Task:** Static sensor nodes, low-cost drone swarms.
- **Requirement:** Extreme energy efficiency, low memory footprint.
- **Recommended Model: ALNet** or **Quantized YOLO26n**.
- **Rationale:** ALNet offers sufficient accuracy for classification tasks with a minimal parameter count (0.17M), maximizing battery life. For detection, a quantized (INT8) version of YOLO26n provides the best trade-off between speed and detection capability on hardware like the Raspberry Pi.[32]

## 9. Conclusion

The landscape of agricultural computer vision in 2026 is defined by a clear functional segmentation. **End-to-End Models** like **YOLO26** have conquered the domain of *perception-for-action*, solving the latency and reliability problems that plagued earlier robotic systems. They are the "muscle" of the autonomous farm, enabling high-speed, reflexive interventions. Conversely, **Multi-Stage Pipelines** and **VLMs** have become the "brain," offering the reasoning capabilities, explainability, and zero-shot adaptability required for high-level diagnosis and decision support.

For the specific challenge of identifying plant species and parts in diseased images, the "best" model is context-dependent. If the goal is to *act* on the plant (spray it, pick it), the speed and small-object precision of **YOLO26** are unrivaled. If the goal is to *understand* the plant (diagnose the pathogen, map the spread), the semantic depth of a **SAM-2 + BioCLIP 2** pipeline is indispensable. The future of this field likely lies in **distillation**: using the massive, slow, and intelligent pipelines to auto-label data, which is then used to train the fast, efficient end-to-end models, creating a virtuous cycle of intelligence and speed that drives the next generation of precision agriculture.

### Alıntılanan çalışmalar

1. Recent advances in plant disease detection: challenges and ..., erişim tarihi Şubat 11, 2026, https://pmc.ncbi.nlm.nih.gov/articles/PMC12570820/
2. End-to-end deep learning pipeline for real-time Bragg peak segmentation: from training to large-scale deployment - Frontiers, erişim tarihi Şubat 11, 2026, https://www.frontiersin.org/journals/high-performance-computing/articles/10.3389/fhpcp.2025.1536471/full

3. Visual-language transformer-based tomato leaf disease detection ..., erişim tarihi Şubat 11, 2026, https://pmc.ncbi.nlm.nih.gov/articles/PMC12560288/
4. Multi-class image predictions of the proposed models after the test process. - ResearchGate, erişim tarihi Şubat 11, 2026, https://www.researchgate.net/figure/Multi-class-image-predictions-of-the-proposed-models-after-the-test-process_fig7_383553932
5. DMN-YOLO: A Robust YOLOv11 Model for Detecting Apple Leaf Diseases in Complex Field Conditions - MDPI, erişim tarihi Şubat 11, 2026, https://www.mdpi.com/2077-0472/15/11/1138
6. Diseases Detection of Occlusion and Overlapping Tomato Leaves Based on Deep Learning, erişim tarihi Şubat 11, 2026, https://pmc.ncbi.nlm.nih.gov/articles/PMC8702556/
7. YOLO11 vs. YOLO26: The Evolution of Real-Time Object Detection - Ultralytics YOLO Docs, erişim tarihi Şubat 11, 2026, https://docs.ultralytics.com/%5C/compare/yolo11-vs-yolo26/
8. YOLO26: Real-Time Object Detection Model for Edge AI [2026] - Tictag, erişim tarihi Şubat 11, 2026, https://www.tictag.io/blog/yolo26-real-time-object-detection-model-for-edge-ai-2026
9. YOLO-based deep learning framework for real-time multi-class plant ..., erişim tarihi Şubat 11, 2026, https://pmc.ncbi.nlm.nih.gov/articles/PMC12764837/
10. Tomato Plant Disease Detection using YOLO11 - bvicam, erişim tarihi Şubat 11, 2026, https://bvicam.in/INDIACom/news/INDIACom%202025%20Proceedings/Main/papers/1444.pdf
11. [2512.15977] Are vision-language models ready to zero-shot replace supervised classification models in agriculture? - arXiv, erişim tarihi Şubat 11, 2026, https://arxiv.org/abs/2512.15977
12. Carbon AI | Innovate Your Agriculture — Carbon Robotics, erişim tarihi Şubat 11, 2026, https://carbonrobotics.com/carbon-ai
13. Carbon Robotics Unveils World's First Large Plant Model Trained On 150 Million Plants, erişim tarihi Şubat 11, 2026, https://quantumzeitgeist.com/carbon-robotics-ai-plant-model/
14. Cotton pest and disease diagnosis via YOLOv11-based deep learning and knowledge graphs: a real-time voice-enabled edge solution - Frontiers, erişim tarihi Şubat 11, 2026, https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2025.1671755/full
15. YOLOv12: State-of-the-Art Object Detection Model, erişim tarihi Şubat 11, 2026, https://yolov12.com/
16. YOLO26 Compared To Other Models: Not as Good as YOLO12 | by Zain Shariff - Medium, erişim tarihi Şubat 11, 2026, https://medium.com/@zainshariff6506/yolo26-not-as-good-as-yolo12-04682a9ee2cb
17. GTDR-YOLOv12: Optimizing YOLO for Efficient and Accurate Weed Detection in

Agriculture, erişim tarihi Şubat 11, 2026, https://www.mdpi.com/2073-4395/15/8/1824

18. YOLO26 vs. YOLO11: A New Era of End-to-End Vision AI, erişim tarihi Şubat 11, 2026, https://docs.ultralytics.com/compare/yolo26-vs-yolo11/
19. Ultralytics YOLO26 - Ultralytics YOLO Docs, erişim tarihi Şubat 11, 2026, https://docs.ultralytics.com/models/yolo26/
20. Qwen2.5 VL! Qwen2.5 VL! Qwen2.5 VL! | Qwen, erişim tarihi Şubat 11, 2026, https://qwenlm.github.io/blog/qwen2.5-vl/
21. Qwen2. 5-VL Technical Report, erişim tarihi Şubat 11, 2026, https://arxiv.org/abs/2502.13923
22. (PDF) Zero-Shot Plant Disease Recognition Using Open Large Vision-Language Models, erişim tarihi Şubat 11, 2026, https://www.researchgate.net/publication/399486552_Zero-Shot_Plant_Disease_Recognition_Using_Open_Large_Vision-Language_Models
23. [2601.04672] Agri-R1: Empowering Generalizable Agricultural Reasoning in Vision-Language Models with Reinforcement Learning - arXiv, erişim tarihi Şubat 11, 2026, https://www.arxiv.org/abs/2601.04672
24. arxiv.org, erişim tarihi Şubat 11, 2026, https://arxiv.org/html/2601.04672v1
25. [2505.07019] A Vision-Language Foundation Model for Leaf Disease Identification - arXiv, erişim tarihi Şubat 11, 2026, https://arxiv.org/abs/2505.07019
26. A Vision-Language Foundation Model for Leaf Disease Identification - arXiv, erişim tarihi Şubat 11, 2026, https://arxiv.org/html/2505.07019v1
27. enalis/scold · Hugging Face, erişim tarihi Şubat 11, 2026, https://huggingface.co/enalis/scold
28. BioCLIP 2: Emergent Properties from Scaling Hierarchical ..., erişim tarihi Şubat 11, 2026, https://imageomics.github.io/bioclip-2/
29. imageomics/bioclip-2 - Hugging Face, erişim tarihi Şubat 11, 2026, https://huggingface.co/imageomics/bioclip-2
30. PMJDM: a multi-task joint detection model for plant disease identification - PMC, erişim tarihi Şubat 11, 2026, https://pmc.ncbi.nlm.nih.gov/articles/PMC12137281/
31. PMJDM: a multi-task joint detection model for plant disease identification - Frontiers, erişim tarihi Şubat 11, 2026, https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2025.1599671/full
32. An attention-augmented lightweight convolutional framework for fine-grained plant leaf disease classification - Frontiers, erişim tarihi Şubat 11, 2026, https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2026.1762956/full
33. Zero-shot instance segmentation for plant phenotyping in vertical farming with foundation models and VC-NMS - PMC, erişim tarihi Şubat 11, 2026, https://pmc.ncbi.nlm.nih.gov/articles/PMC12086161/
34. erişim tarihi Şubat 11, 2026, https://docs.ultralytics.com/models/sam-2/#:~:text=While%20SAM%202%20excels%20in,deployment%20in%20resource%2Dconstrained%20environments.
35. YOLO-SAM AgriScan: A Unified Framework for Ripe Strawberry Detection and

Segmentation with Few-Shot and Zero-Shot Learning - PubMed, erişim tarihi Şubat 11, 2026, https://pubmed.ncbi.nlm.nih.gov/41471673/

36. YOLO26 vs. EfficientDet: The New Standard in Object Detection - Ultralytics YOLO Docs, erişim tarihi Şubat 11, 2026, https://docs.ultralytics.com/compare/yolo26-vs-efficientdet/
37. AgMMU: A Comprehensive Agricultural Multimodal Understanding Benchmark - arXiv, erişim tarihi Şubat 11, 2026, https://arxiv.org/html/2504.10568v2
38. Zero-Shot Semantic Segmentation for Robots in Agriculture, erişim tarihi Şubat 11, 2026, https://www.ipb.uni-bonn.de/pdfs/chong2025iros.pdf
39. Carbon Robotics Launches Large Plant Model AI for Real-Time Weed Detection - UBOS, erişim tarihi Şubat 11, 2026, https://ubos.tech/news/carbon-robotics-launches-large-plant-model-ai-for-real%E2%80%91time-weed-detection/
40. How to train Ultralytics YOLO models to detect animals in the wild, erişim tarihi Şubat 11, 2026, https://www.ultralytics.com/blog/how-to-train-ultralytics-yolo-models-to-detect-animals-in-the-wild