NOTE 2

Vision as a Foundation for World Representation in AI

Sentinel Agro Labs — Research Note

Abstract

Vision-based learning provides a direct interface to the physical world. This note summarizes why modern AI systems increasingly rely on self-supervised vision and video models to build internal representations of reality, and how these representations can support higher-level reasoning, prediction, and planning.

Motivation

Many real-world domains — including agriculture, climate monitoring, and robotics — involve:

- continuous environments,
- spatial relationships,
- temporal dynamics.

Purely symbolic or language-based systems struggle to model these properties reliably. Vision provides a natural substrate for learning structure and causality from data.

Key Developments in Vision Research

Recent advances emphasize **representation learning** rather than task-specific prediction:

- **Self-supervised learning** (e.g., DINOv2)
- **Predictive world models** (e.g., JEPA architectures)
- **Foundation vision models** (e.g., Segment Anything)

These models learn reusable features that generalize across tasks and environments.

World Models and Prediction

World models aim to:

- encode the current state of the environment,
- predict future states,
- support planning without direct action execution.

By learning in latent space, models can:

- reduce noise,
- abstract irrelevant detail,

- focus on structure and dynamics.

Relevance to Sentinel Agro Labs

Vision-based world representations are particularly relevant for:

- environmental sensing,
- agricultural monitoring,
- drone-based observation,
- long-term system understanding.

They form a natural bridge between raw sensor data and higher-level AI reasoning.

Future Integration with Language Models

Vision representations can be combined with language models by:

- summarizing latent states into structured descriptors,
- enabling language-based querying of visual context,
- supporting explainable decisions grounded in perception.

This hybrid approach aligns with current research trends in multimodal AI.

Status

This document reflects an **exploratory research perspective** based on publicly available literature. Implementation and experimentation will follow a staged research approach.

References (selected)

- DINOv2
- Segment Anything
- I-JEPA / V-JEPA