ICPR 2026
ORAL PRESENTATION

DSI-YOLO:

A Physics-Aware Framework for Citrus Detection in Unstructured Orchard Environments

Robust citrus detection in all-weather, occlusion-heavy, and unstructured environments using physics-aware deep learning.

PHYSICS-AWARE
Modeling real-world radar signals

ROBUST & STABLE
Reliable across weather and scene variations

ACCURATE DETECTION
High precision in dense, occluded orchards

EFFICIENT & GENERAL
Lightweight and generalizes to diverse environments

Why Citrus Detection is Difficult?

Unstructured orchard environments impose three severe physical challenges on visual perception.

Spectral Homogeneity	Green citrus shares highly similar appearance with surrounding leaves, causing blurred boundaries and weak foreground-background contrast.	• Green fruits exhibit similar spectral characteristics to surrounding foliage. • RGB images lack discriminative features in spectral dimension.
Topological Entanglement	Dense fruit clustering and mutual occlusion make adjacent instances overlap, leading to feature confusion and missed detections.	• Dense occlusion and contact among fruits result in topological entanglement. • Standard detectors struggle to separate clustered fruits.
Dynamic Scale Drift	Camera-target distance changes during robotic motion, resulting in large scale variation and unstable detection responses.	• Significant scale variation exists in natural orchard scenes. • Fixed-scale detectors cannot adapt to multi-scale object sizes.

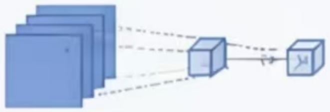
Limitations of Existing Detectors

ICPR 2026 ORAL PRESENTATION

Conventional detectors suffer from three inherent inductive bias failures in frequency,topology, and scale dimensions, leading to suboptimal performance in citrus orchards.

Existing Detectors

CNN-based Detectors



Global Feature Aggregation

Fixed Receptive Fields

Inherent Limitations

1 Lack of Frequency Discriminability

RGB convolution struggles to decouple texture from color.

2 Isotropic Aggregation Limits

Full-map interaction introduces feature assimilation.

3 Static Geometric Mismatch

Fixed receptive fields cannot adapt to dynamic scale drift.

Consequence in Citrus Orchards

1 Feature Collapse

Spectrally similar fruits and leaves become indistinguishable.

→ **Consequence:** The model mistakes foliage for fruit or misses fruits entirely, leading to low detection recall.

Key Issues:

- Green fruits exhibit similar spectral characteristics to surrounding foliage.
- RGB images lack discriminative features in spectral dimension.

2 Feature Assimilation

Adjacent fruits influence each other and are merged into one prediction.

→ **Consequence:** Multiple fruits are detected as a single object, causing inaccurate counting and localization.

Key Issues:

- Dense occlusion and contact among fruits result in topological entanglement.
- Standard detectors struggle to separate clustered fruits.

3 Scale Mismatch

Objects at different distances are poorly detected due to non-adaptive receptive fields.

→ **Consequence:** Small or distant fruits are easily missed, reducing detection robustness.

Key Issues:

- Significant scale variation exists in natural orchard scenes.
- Fixed-scale detectors cannot adapt to multi-scale object sizes.

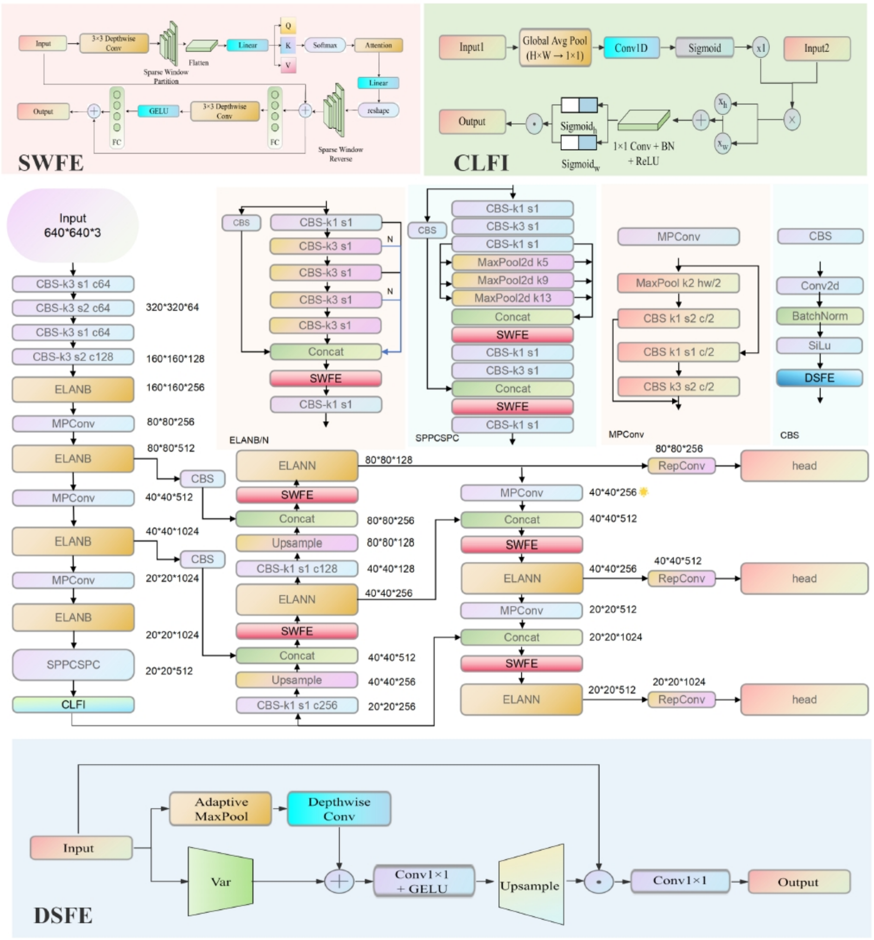
These limitations motivate the design of **DSI-YOLO**,which embeds physical priors to overcome the above challenges.

DSI-YOLO: Physics-Aware Detection Framework

How does DSI-YOLO solve orchard-specific physical challenges?

Input Image → **Physics-Aware Feature Interaction** → Detection Head → **Robust Citrus Detection**

Physics-Aware Feature Interaction:
- **DSFE** — Frequency
- **SWFE** — Topology
- **CLFI** — Scale



Frequency-aware texture reconstruction
Solves spectral homogeneity

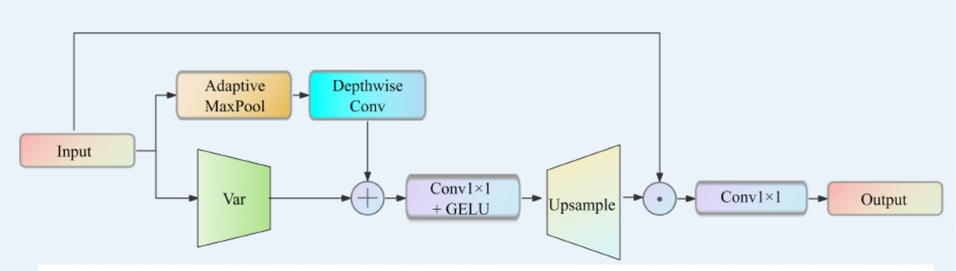
Sparse-window topology decoupling
Solves topological entanglement

Dynamic receptive field calibration
Solves dynamic scale drift

DSI-YOLO embeds orchard **physical priors** into feature interaction instead of directly modifying detector structure.

Dual-Statistical Feature Enhancer (DSFE)

How does DSFE recover fruit boundaries under spectral camouflage?



What is Spectral Homogeneity?

- Green citrus shares highly similar appearance with surrounding leaves, causing **blurred boundaries** and **weak foreground-background contrast.**

Why is it a Problem?

Fruit ≈ Leaf

↓

Weak boundary perception

How We Address it

Input Feature

↓

Dual Streams

Low-frequency + **High-frequency**

↓

Adaptive Fusion — Low-frequency semantics + high-frequency texture statistics

↓

Boundary Reconstruction

DSFE separates **semantic structure** and **texture statistics** to recover camouflaged fruit boundaries.

Sparse Window Feature Extractor (SWFE)

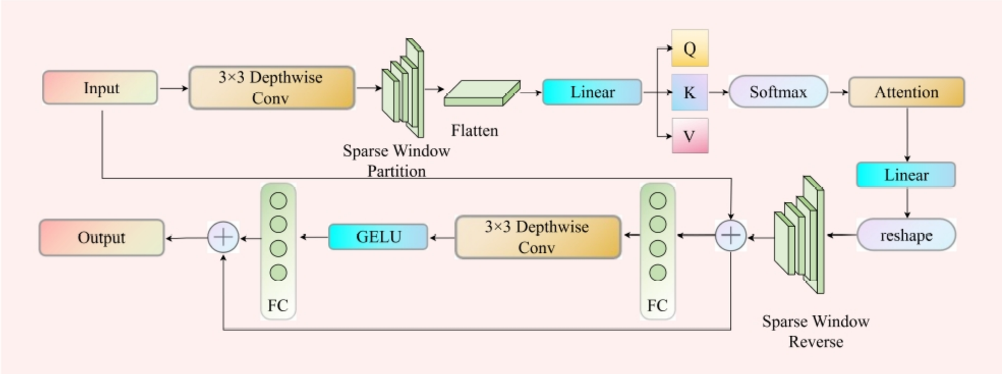
Dense fruit clustering causes feature assimilation between adjacent instances.

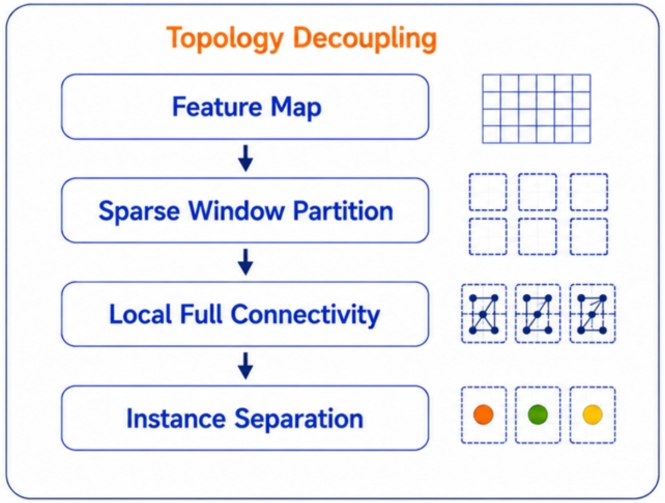
Global Interaction

Captures long-range dependencies among all fruit instances.

Feature Assimilation

Aggregates and refines features through global context.





SWFE **isolates local fruit regions** and **prevents cross-instance interference.**

Channel-Local Feature Interaction (CLFI)

Dynamic scale drift causes unstable perception when camera-target distance changes.

Dynamic Scale Drift

Changes in camera-target **distance** during robotic motion cause large variations in object scale, leading to **unstable detection**.

Near field

Near view: Objects are closer to the camera.

Robot motion

• • •

Far field

Far view: Objects are farther from the camera.

Camera-target distance changes

As the robot moves, the distance between the camera and target changes continoously, causing dynamic scale drift.

Near fruit

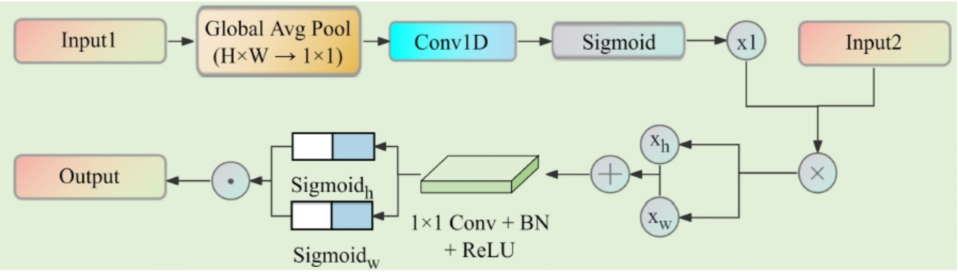
↓

Large scale response: Objects appear larger with more detail, resulting in stronger signals.

Far fruit

↓

Small scale response: Objects appear smaller with less detail. resulting in weaker signals.



DSI-YOLO

Dynamically **adapts** to **scale variations** for more robust detection.

Channel-aware Calibration — Calibrate feature responses across different channels.

\+

Spatial Refinement — Refine spatial details to enhance localization.

↓

Adaptive Scale Response — Dynamically adjust responses to different target scales.

CLFI dynamically **adjusts receptive fields** according to feature semantics and **target scale variations**, effectively handling **dynamic scale drift** in unstructured orchard environments.

Experimental Setup

Dataset

CitDet Benchmark

We evaluate DSI-YOLO on the CitDet dataset, a high-resolution benchmark specifically designed for in-field citrus detection.

CitDet contains 579 images with 44,233 annotated citrus instances.

The images are captured in unstructured orchard environments using a mobile robotic platform under natural conditions.

Detailed statistics of the dataset are summarized as follows.

- **CitDet Benchmark**
- **579** high-resolution images
- **44,233** annotated instances
- **2448 × 3264** image resolution
- Average object size **< 0.1%** of image area (**~50 × 50 pixels**)
- Official split: train / val / test = **6 : 2 : 2**

Implementation Details

Hardware:	NVIDIA GeForce RTX 4090 (24 GB VRAM)
Framework:	PyTorch 2.4.0 + CUDA 12.4
Optimizer:	SGD
LR Schedule:	Cosine Annealing
Epochs:	1000
Batch Size:	4
Input Resolution:	640 × 640 (Multi-scale)
Warmup Epochs:	3
Data Augmentation:	Mosaic + Mixup
Label Assignment:	OTA

Training Objective / Loss Function

Total Loss = Box Loss + Obj Loss + Cls Loss

$\lambda_{box} = 0.05, \ \lambda_{obj} = 0.7, \ \lambda_{cls} = 0.3$

Evaluation Metrics

- **Precision (P)**
- **Recall (R)**
- **mAP@0.5 (AP_{50})**
- **mAP@0.5:0.95**

Precision and Recall evaluate detection reliability, while mAP measures localization and overall detection quality.

Qualitative Comparison

- Fig. 5. Qualitative comparison on CitDet: (a) Ground Truth; (b) Faster R-CNN; (c) YOLOv5; (d) YOLOv7 (Baseline); (e) DSI-YOLO (Ours).
- **Yellow circles** indicate targets missed by the baseline but recalled by our method (with GT). **Purple circles** highlight valid fruits detected by our method but missed by both manual annotation and the baseline **(Unlabeled TPs)**.
  Our method demonstrates superior robustness to dense occlusion and spectral camouflage compared to the YOLOv7 baseline.

(a) GT | (b) Faster R-CNN | (c) YOLOv5 | (d) YOLOv7 (Baseline) | (e) DSI-YOLO (Ours)

Key Observations

Improved recall of missed fruits
DSI-YOLO recovers targets missed by baseline detectors (yellow circles).

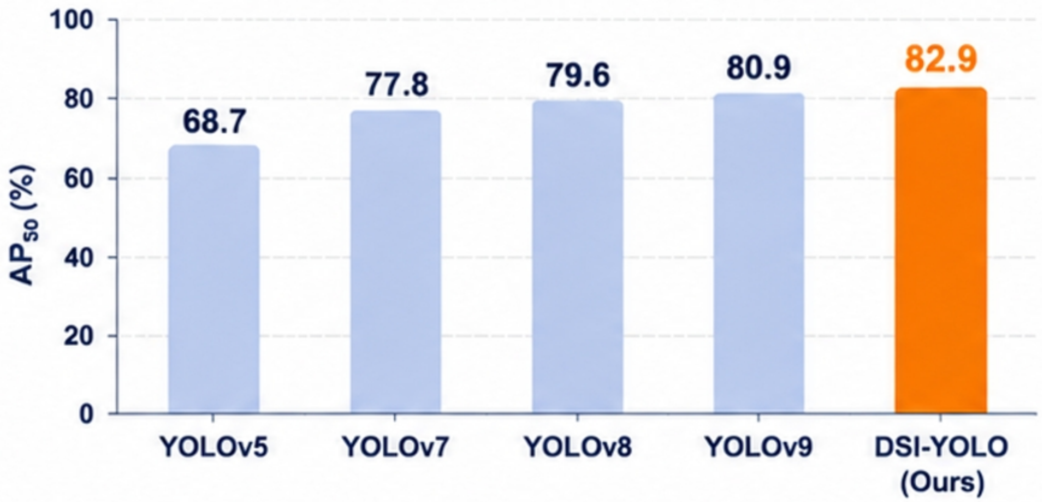
Better detection under dense occlusion
DSI-YOLO maintains object separation in clustered fruit regions.

Robustness against spectral camouflage
DSI-YOLO detects fruits with similar appearance to surrounding foliage.

Comparison with State-of-the-Art Methods

DSI-YOLO achieves superior performance on the CitDet benchmark.

AP_{50} (%) Comparison



Performance Comparison on CitDet

Model	Parameters	$AP_{50:95}$ (%)	AP_{50} (%)
Faster R-CNN [12]	41.3 M	22.0	51.5
YOLOv5m [13]	21.1 M	34.8	70.0
YOLOv7 [11]	36.5 M	38.4	77.8
YOLOv8m [19]	25.9 M	37.8	76.9
YOLOv9c [20]	25.3 M	38.1	77.5
YOLOv10m [21]	15.4 M	36.5	75.4
YOLOv11m [22]	20.1 M	37.2	76.5
YOLOv12m [23]	20.2 M	36.9	76.1
YOLOv13m [24]	18.0 M	36.1	75.2
DSI-YOLO-v7 (Ours)	38.2 M	**42.7**	**82.9**

DSI-YOLO obtains the **best AP_{50}** of 82.9%,
outperforming YOLOv7 by +5.1 percentage points.

AP_{50} **82.9%**	Gain over YOLOv7 **+5.1 pp**	$mAP_{50:95}$ **42.7%**	Gain over YOLOv7 **+4.1 pp**	Recall **79.2%**	Gain over YOLOv7 **+6.9 pp**

DSI-YOLO consistently outperforms state-of-the-art YOLO variants across all metrics,
demonstrating **strong robustness** in complex orchard environments.

Ablation Study

Module-wise Comparison

Same training settings; each module is toggled independently.

Model	AP_{50} (%)
Baseline	77.8
+DSFE	81.0
+SWFE	81.0
+CLFI	80.9
All	82.9

DSFE

Boundary Enhancement

Suppresses spectral camouflage and improves target boundary discriminability.

SWFE

Cluster Separation

Decouples densely entangled fruits and reduces feature assimilation.

CLFI

Scale Stability

Calibrates dynamic scale drift across near-field and far-field targets.

Overall Best: **82.9 AP_{50}**

Single-module gain: **up to 81.0**

Complementary design

Three modules provide complementary improvements from frequency, topology and scale perspectives.

Conclusion

- We propose **DSI-YOLO**, a physics-aware detector for unstructured orchards that integrates three complementary cues: frequency-domain texture, local topological geometry, and cross-scale feature consistency.
- With **DSFE, SWFE**, and **CLFI**, it explicitly targets Spectral Homogeneity, Topological Entanglement, and Dynamic Scale Drift.
- On CitDet, DSI-YOLO improves **AP_{50}** to **82.9% (+5.1pp** vs. YOLOv7) and **Recall** to **79.2% (+6.9pp)**, and ablations confirm the three modules are complementary in reducing both false positives and false negatives.

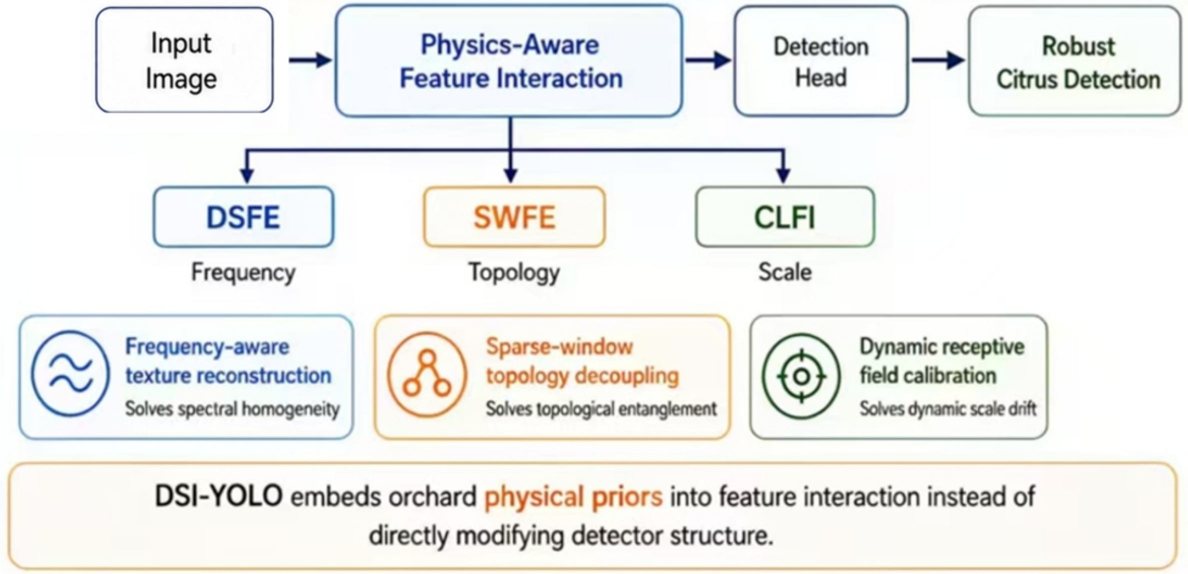


Fig. 1 Overview of the DSI-YOLO Physics-Aware Detection Framework.

- Our current system relies on RGB inputs; future work will incorporate multi-modal sensing (e.g., NIR/depth) and unsupervised domain adaptation to further improve robustness and generalization.

AP_{50}
82.9%
(+5.1pp vs. YOLOv7)

Recall
79.2%
(+6.9pp vs. YOLOv7)

Three modules are complementary

RGB inputs → multi-modal sensing (e.g., NIR/depth)