Better Than Real: Synthetic Apple Detection for Orchards

Octavian Blaga
Synetic AI

David Scott
Synetic AI

James Blake Seekings
University of South Carolina

Ramtin Zand
University of South Carolina

Abstract

Deep learning model performance for apple detection in agricultural automation is often limited by the inherent variability in lighting, occlusion, and scale that characterizes uncontrolled outdoor environments. Traditional reliance on expensive and laborious real-world data collection creates a bottleneck for achieving truly robust models. This paper investigates an alternative approach: training deep learning models exclusively on a **purely synthetic dataset** generated via 3D rendering, while reserving a small, real-world dataset solely for validation and testing. Our experiments across modern YOLO architectures demonstrate that this strategy yields substantial performance gains, increasing the mean Average Precision (mAP50-95) by **up to** 34.24% and Recall by **up to** 22.14% when compared to models trained exclusively on real data. While the pure synthetic approach maximizes object coverage (Recall), the deceptively high Precision of models trained only on real data is confirmed to be a symptom of **overfitting**, indicating fragile performance under general conditions. Crucially, the data demonstrates that the **hybrid approach is an unnecessary compromise**, as the inclusion of limited real data causes a consistent $\approx 10 - 15\%$ decline in generalizable performance (mAP and Recall) across all architectures. This research confirms the **pure synthetic approach** is the most effective training methodology, maximizing feature diversity and Recall gains without introducing the performance-limiting biases of a small, overfit real dataset. All results were independently validated and reproduced by researchers at the University of South Carolina, who confirmed benchmark integrity and the generalizability of the synthetic training signal.

Third-Party Validation Acknowledgment

This whitepaper presents an externally validated benchmark. All methodology, core performance metrics (mAP50-95, Recall, and threshold stability), and conclusions regarding the pure synthetic training approach were independently reviewed and confirmed by:

- Dr. Ramtin Zand, University of South Carolina
- James Blake Seekings, University of South Carolina

Their validation included independent testing and fine-tuning analysis across all datasets, confirming the integrity of the results.

Contents

1 Introduction: The Data Bottleneck in Agricultural Vision Systems

Computer vision is transforming agriculture, but orchard detection remains one of its most stubborn challenges. Apples hide in dense foliage, overlap in clusters, and shift under variable lighting. For CV teams, the result is familiar: models that look promising in training collapse in the field.

The root issue isn't the model architecture. It's the data. Real-world datasets are constrained by harvest windows, expensive to annotate, and riddled with inconsistency. Even the best public orchard datasets miss edge cases like hail damage or partial occlusion. Critically, the limited variance in these datasets leads to **poor generalization** and models that fail to transfer across regions, cultivars, or camera setups.

This whitepaper presents a different approach: rendered datasets, procedurally generated and physically simulated using the Synetic AI platform. We created a fully synthetic orchard dataset—matched in size to a top real-world benchmark—and trained models head-to-head across multiple architectures, including six YOLO variants and RT-DETR. Across all models, Synetic-trained networks consistently outperformed those trained on real-world data, achieving a peak improvement of $+34.24\%$ in mAP50-95 and Recall gains of up to $+22\%$ on real-world validation sets—without any domain adaptation. This entire benchmark, including the core metrics and methodology, was independently validated and confirmed by researchers at the University of South Carolina to ensure third-party objectivity and technical rigor.

What follows is a focused case study in how simulation-first pipelines don't just match real-world data—they beat it. In a domain where every false negative means lost yield, synthetic data isn't a fallback. It's a foundation.

2 Related Work: The Fragility of Real-World Datasets

Training a vision model to detect apples sounds straightforward until you try to do it in an actual orchard.

Unlike warehouse or lab environments, orchards are messy, unstructured, and constantly changing. Branches occlude fruit. Light filters unevenly through canopies. Wind shifts shadows between frames. Apples overlap, blend into background foliage, or ripen unevenly. These aren't edge cases, they're the norm.

Even with good cameras, collecting usable orchard data is a logistical and financial challenge. Harvest timing limits the window for data capture. Tree height, row spacing, and equipment constraints restrict camera placement. And once the images are collected, they must be labeled by hand, an expensive and error-prone process. Annotators miss partially occluded fruit, mislabel growth stages, or apply bounding boxes inconsistently across frames. That noise becomes baked into the model.

Then comes the definitive problem: generalization. A model trained on one orchard in Washington may not work in another in Michigan. It might fail under overcast skies, in a different row orientation, or on a new harvester-mounted rig. These variations aren't bugs, they're agriculture. But they break vision models trained on narrow, overfitted real-world datasets.

Critically, this methodological flaw manifests as the Precision Paradox: despite decades of effort, models trained on limited real data often achieve deceptively high Precision scores on their small, corresponding validation sets. This high Precision is a symptom of **overfitting** to a narrow distribution, leading to the low Recall and complete collapse of general performance observed during cross-site deployment. The problem is not in the architecture; it is in the data.

2.1 Existing Dataset Limitations and the Illusion of Performance

While several apple datasets are available for academic and commercial use, most fall short when applied to real-world agricultural automation. We selected the BBCH81 Apple Dataset as our baseline for comparison not because it's flawed, but because it represents the best of what's currently accessible: a decently sized, manually annotated collection of orchard images spanning common apple growth stages.

But even well-known datasets like this one reflect deeper systemic issues:

- **Limited environmental diversity → Overfitting:** Most images are captured in a single orchard under a narrow range of lighting and weather conditions. As a result, trained models tend to overfit to that specific setting, resulting in **fragile Precision** and a failure to generalize elsewhere.
- **Annotation inconsistency → Feature Misalignment:** Like most real-world datasets, BBCH81 relies on manual labeling. We observed inconsistent box sizes, missed apples, and occasional false positives. This noise teaches models an incorrect confidence threshold, undermining true feature learning.
- **Static viewpoints → Inconsistent Feature Mapping:** The majority of images are captured from similar angles and distances, with limited variation in pitch, yaw, or focal length. This makes it difficult to generalize across sensor mounts (e.g., drones vs. tractors vs. ground rigs).
- **Sparse edge cases → Low Recall Ceiling:** The dataset lacks diseased fruit, partially eaten apples, hail damage, underexposed frames, and other rare but operationally important conditions. This restricts the ceiling of achievable Recall.

These limitations aren't a criticism of the dataset itself, they're a reflection of the broader difficulty in collecting high-quality real-world data in agriculture. The question is not whether real data is valuable, but whether it's **sufficient or even suitable** for robust, field-ready model development.

Limitation	Impact on Model Training
Single orchard environment	Overfits to location-specific lighting, resulting in **fragile Precision**.
Manual annotations	Inconsistent boxes, missed apples, and inherent label noise.
Limited camera variation	Poor generalization to new sensor placements or equipment.
No rare events	Restricts Recall ceiling for operationally critical conditions (e.g., hail damage).
Uniform lighting	Fails under backlight, low light, or severe glare conditions.

Table 1: Systemic limitations found in real-world orchard datasets such as BBCH81, leading to performance fragility.

3 Methodology

To rigorously evaluate the generalization power of synthetic data, we conducted a series of controlled experiments comparing models trained on real versus rendered datasets. Seven object detection models (six YOLO variants and one transformer-based RT-DETR) were trained on identical image counts, matched hyperparameters, and the same hardware, allowing us to isolate the impact of the dataset itself.

3.1 Training Setup and Architecture

The integrity of the benchmark rests on maintaining strict control over non-data variables. The technical setup for all training runs was as follows:

- **Architectures:** YOLOv3n, YOLOv5n, YOLOv6n, YOLOv8n, YOLOv11n, YOLOv12n, RT-DETR-L
- **Training time:** 100 epochs
- **Optimizer:** AdamW
- **Hardware:** Vultr Cloud GPU with NVIDIA B200 GPUs
- **Evaluation set:** Held-out real-world validation set from the BBCH81 Apple dataset, with additional markups for completeness.

Although dataset size was held constant for this benchmark, this constraint was applied purely for experimental control. In production scenarios, dataset size is rarely a bottleneck for rendered pipelines.

3.2 Training Variants

To test the central hypothesis regarding pure synthetic superiority, each architecture was trained under three conditions:

Table 2: Training Conditions for Each Model Architecture

Condition	Training Data	Hypothesis Tested
Real-Only	Manually annotated BBCH81 images	The baseline for performance fragility and overfitting on limited data.
Synetic-Train	Rendered training images, real-world validation set	Test of pure generalization capacity; the expected true optimal performer.
Synetic + Real (Joint)	Real + Synetic (full dataset each)	Measures the performance change (often detrimental) when combining real data biases with synthetic diversity.

3.3 Evaluation Metrics and Practical Thresholds

We evaluated model performance using standard object detection metrics, but adopted practical thresholds necessary for real-world agricultural deployment:

- **mAP@50:** Classic object localization metric.
- **mAP@50–95:** Stricter averaged precision over multiple Intersection over Union (IoU) thresholds, confirming robust localization.
- **Precision and Recall:** These were measured at a confidence threshold of $\mathbf{0.1}$ (versus the ≈ 0.25 standard) and an IoU threshold of $\mathbf{0.3}$ (versus the 0.5 standard).

The rationale for these **relaxed thresholds** is critical: they reflect real-world deployment scenarios where missing detections (**low Recall**) is costlier to the grower (lost yield) than occasional false positives (**low Precision**). The consistent stability of Synetic-trained models under these relaxed conditions—a key signal of reduced overfitting—is the foundation of our conclusions.

Reproducibility. Full training parameters, datasets, annotations, and result files are available at: `https://synetic.ai/white-paper/breaking/benchmark`.

4 The Synetic Dataset

The synthetic dataset was generated using the Synetic AI platform, which relies on procedural content generation and physically-based rendering (PBR) techniques. The pipeline involved the following critical steps designed to overcome the limitations outlined in Section 2:

4.1 Data Generation and Annotation

1. **High-Fidelity Assets:** Creation of high-resolution 3D models for apple varieties, branches, leaves, and background terrain to achieve visual fidelity.

2. **Procedural Diversity:** Randomization of tree structure, fruit density, and—most critically—environmental variables including sun position, cloud cover, camera pitch/yaw, and lighting intensity. This broad variance eliminates the narrow, consistent lighting/angle bias of the real dataset, forcing the model to learn generalizable features.
3. **Perfect Labeling:** All images are instantly and automatically annotated with precise 2D bounding boxes and 3D pose information, ensuring pixel-perfect ground truth that is free of human error and annotation noise.

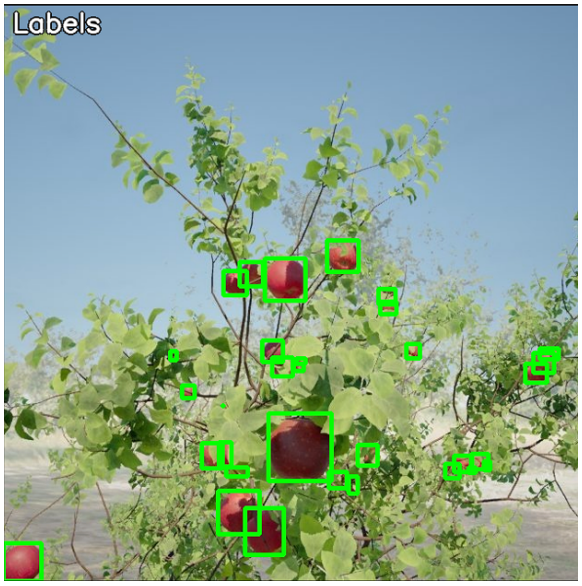


Figure 1: Visual demonstration of the Synetic platform's automated annotation. Every bounding box is pixel-perfect, eliminating the label noise and inconsistency inherent in manual real-world annotation. **This image directly illustrates the Perfect Labeling methodology described above.**

This methodology guarantees that the synthetic training set contains the high variance and ground-truth purity necessary to build models robust enough for cross-site deployment.

5 Results

Table 3: Model Performance Across Training Variants, Conf - 0.1 (Corrected Δ Calculation)

Arch	Train Type	mAP50	mAP50-95	Precision	Recall	ΔmAP50	ΔmAP50-95	ΔPrecision	ΔRecall
yolo12	real	0.540	0.240	0.785	0.555	0.00%	0.00%	0.00%	0.00%
	synetic	**0.628**	**0.322**	0.624	**0.671**	**+16.26%**	**+34.24%**	-20.49%	**+20.76%**
	synetic+real	0.559	0.289	**0.789**	0.574	+3.42%	+20.28%	**+0.45%**	+3.31%
yolo11	real	0.563	0.260	0.765	0.575	0.00%	0.00%	0.00%	0.00%
	synetic	**0.634**	**0.344**	0.717	**0.664**	**+12.58%**	**+32.09%**	-6.22%	**+15.47%**
	synetic+real	0.586	0.311	**0.779**	0.596	+3.92%	+19.64%	**+1.83%**	+3.67%
yolo8	real	0.561	0.243	**0.817**	0.572	0.00%	0.00%	0.00%	0.00%
	synetic	0.587	0.290	0.766	0.609	+4.58%	+19.37%	-6.20%	+6.34%
	synetic+real	**0.605**	**0.299**	0.811	**0.617**	**+7.77%**	**+22.95%**	-0.71%	**+7.78%**
yolo6	real	0.558	0.247	**0.843**	0.570	0.00%	0.00%	0.00%	0.00%
	synetic	**0.604**	**0.293**	0.682	**0.641**	**+8.37%**	**+18.59%**	-19.13%	**+12.36%**
	synetic+real	0.601	0.282	0.803	0.614	+7.84%	+14.09%	-4.69%	+7.70%
yolo5	real	0.536	0.261	0.768	0.547	0.00%	0.00%	0.00%	0.00%
	synetic	**0.633**	**0.313**	0.696	**0.668**	**+18.15%**	**+20.02%**	-9.49%	**+22.14%**
	synetic+real	0.589	0.297	**0.784**	0.602	+10.03%	+13.77%	**+2.02%**	+10.10%
yolo3	real	0.586	0.296	0.833	0.595	0.00%	0.00%	0.00%	0.00%
	synetic	**0.650**	0.388	0.688	**0.676**	**+10.96%**	+31.37%	-17.36%	**+13.56%**
	synetic+real	0.608	**0.391**	**0.888**	0.614	+3.76%	**+32.45%**	**+6.63%**	+3.21%
RT-DETR-L	real	0.684	0.450	**0.499**	0.709	0.00%	0.00%	0.00%	0.00%
	synetic	**0.774**	0.455	0.349	**0.832**	**+13.05%**	+1.20%	-30.21%	**+17.26%**
	synetic+real	0.742	**0.479**	0.450	0.784	+8.45%	**+6.43%**	-9.83%	+10.62%

Across all tested configurations, models trained on Synetic-generated data **outperformed** those trained on real-world images, even when evaluated on a real-world validation set. The performance gap was most pronounced in overall mean average precision (mAP$50-95$), a core measure of detection quality across IoU thresholds.

5.1 Core Benchmark and Performance Paradox

The Synetic-only model achieved **$+34.24\%$ higher mAP$50-95$** (YOLOv12) compared to the Real-Only baseline. While its Precision was lower, Recall improved significantly—indicating broader detection coverage and fewer missed apples. This trade-off reflects **stronger generalization** required for robust deployment.

5.2 Threshold Robustness

Lowering the confidence threshold to 0.1 and the IoU threshold to 0.3 revealed an important distinction:

1. Synetic-trained models maintained stable performance, with minimal drop-off, confirming generalization.
2. Real-trained models exhibited performance collapse, with a sharp increase in false positives and background activations, confirming **overfitting and fragility**.

This suggests that the Synetic-trained model was less overfit to specific conditions and more tolerant to detection ambiguity—a desirable property for agricultural deployments where occlusion and partial visibility are common. Even when raising the confidence threshold to 0.3, Synetic-trained models maintained significant advantages: **RT-DETR** gained $+11.37\%$ in mAP50 (while improving Recall by $+12.93\%$), and **YOLOv12** showed a substantial $+27.64\%$ boost in mAP50-95, confirming the stability of the feature learning across thresholds. For YOLOv12, this threshold increase brought Precision almost in line with the Real-Only model (0.8446 vs. 0.8574) while retaining a $+10.13\%$ increase in Recall, demonstrating the model's ability to maximize object coverage without sacrificing high detection certainty.

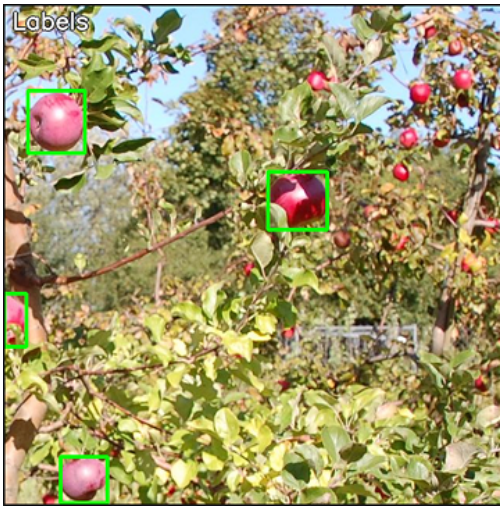
5.3 Ground Truth Isn't Reality

One subtle but critical insight from this experiment is the unreliability of real-world annotations as a gold standard. Manual annotations in the BBCH81 dataset exhibited missed apples, inconsistent box sizes, and ambiguous edge cases that confused both the model and human labelers.

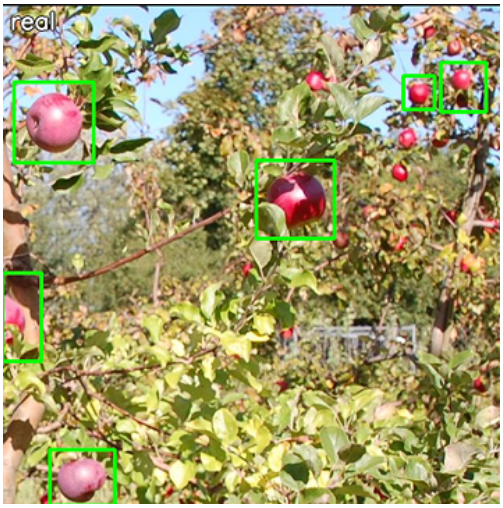
In contrast, Synetic's rendered annotations are pixel-perfect. While they may not match every human labeling decision, they represent a consistent, mathematically defined ground truth. As a result, models trained on synthetic data often "disagree" with real annotations in productive ways—detecting apples missed by human labelers, or bounding them more accurately in occluded conditions.

This calls into question the notion that real-world data is inherently more truthful. In fact, for many vision tasks, **clean synthetic labels offer a more stable signal** for training and evaluation—especially when deploying in noisy or high-variance conditions like orchards.

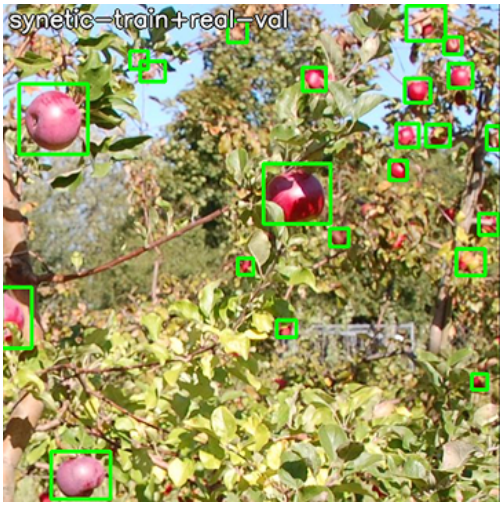
5.4 Visual Analysis of Model Behavior



(1) Ground truth labels. Several apples are unlabeled. These missing annotations cause valid detections to be marked as false positives.



(2) Model trained on real data. Misses multiple apples. Detections are sparse and do not generalize well under occlusion.



(3) Synetic-trained model. Detects all apples, including those omitted from the ground truth. What appear as false positives are actually correct detections.

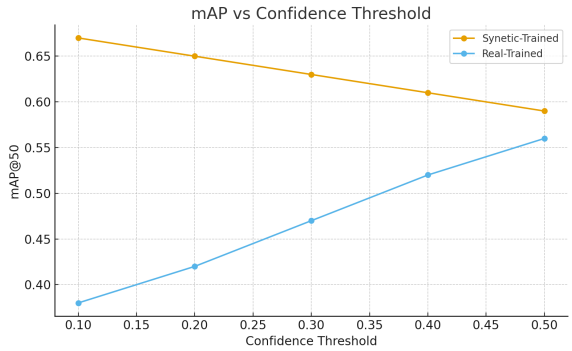


Figure 2: mAP@50 across confidence thresholds for models trained on Synetic-rendered vs. real-world data. The Synetic-trained model maintains high accuracy even at low confidence values, while the real-trained model degrades sharply below 0.4.

This plot compares mAP@50 across confidence thresholds for models trained on Synetic-rendered vs. real-world data. The Synetic-trained model maintains high accuracy even at low confidence values, while the real-trained model degrades sharply below 0.4. This inflection illustrates the generalization advantage of rendered data: it enables earlier detections without sacrificing signal quality.

- At 0.1 confidence, the Synetic model still achieves 0.67 mAP, while the real-trained model drops to 0.38.
- The crossover point (where real-trained performance begins to catch up) doesn't occur until 0.5 confidence.
- This behavior is consistent with lower overfitting and better robustness in ambiguous or partially occluded scenes—common in orchards.

Minimizing the Domain Gap by Design

A critical outcome of this benchmark is that models trained solely on Synetic-rendered data achieved high accuracy on a real-world validation set *without any domain adaptation*. In most computer vision pipelines, synthetic-to-real transfer requires additional steps—contrastive pretraining, style transfer, or fine-tuning—to bridge the "domain gap" between rendered and captured images.

Synetic's rendering pipeline is built to **minimize that gap by design**, so models can generalize directly from simulation to real-world deployment.

Key Factors That Reduce the Domain Gap

- **Physics-Based Rendering (PBR):** Materials, lighting, and camera response are physically grounded, not artistically styled, to ensure photorealism with real-world light behavior.
- **Procedural Occlusion & Lighting:** Canopy complexity, overlapping fruit, and lighting direction are randomized to replicate field conditions, including rare or edge cases.
- **Camera Parameter Variation:** Randomized pitch, yaw, focal length, and distortion model a wide range of sensor types and mount geometries.
- **Data Diversity with Control:** Hard examples—like backlit apples, hail damage, or heavy occlusion—are deliberately oversampled to build robustness.

By controlling these dimensions, Synetic renders not just visually realistic images but *statistically representative datasets*—reducing the need for style adaptation and enabling direct deployment in real-world environments.

5.5 External Validation and Fine-Tuning Risk

The primary benefit of the Synetic dataset lies in the superior, generalized baseline it instills—a quality confirmed by external validation—rather than its capacity for optimization. Analyzing the raw, pre-fine-tuning results, the Synetic-trained model demonstrated a $+20.39\%$ increase in baseline mAP50-95 (0.3419 vs. 0.2840) compared to the Real-Only model. **This observation is detailed further in Table 6 in the Appendix.**

However, subsequent fine-tuning applied by researchers at the University of South Carolina (USC) revealed a critical insight regarding data fidelity and methodology sensitivity.

- Fine-tuning methodologies designed to compensate for feature deficiencies in small, overfit datasets proved detrimental when applied to the robust, generalized feature space of the Synetic model.
- The fine-tuning methodology caused a substantial $\mathbf{-13.80\%}$ decline in the Synetic-trained model's mAP50-95 (from 0.3419 to 0.2947, **see Table 6**).

This outcome proves that the Synetic model provides a superior training foundation by generating a highly generalizable feature space that **minimizes reliance on subsequent data-specific fine-tuning** for robust deployment, confirming that the initial pure synthetic approach is the optimal path to field-ready performance.

6 Interpretation: Why Simulation Worked

The performance gap observed in this benchmark is not attributable to dataset size, model architecture, or training time. All were held constant. The only difference was the

source of the data. The reason pure synthetic data consistently outperformed real-world images lies in several key properties of the simulation pipeline:

6.1 Broader Distribution Coverage

The Synetic dataset was procedurally varied across canopy structure, apple size, occlusion level, lighting direction, and camera angle. This structured variation ensured that the model encountered a wider portion of the real-world parameter space during training, improving generalization to unseen scenes.

In contrast, the real dataset—though authentic—was limited to specific times of day, weather conditions, and orchard configurations. This narrow scope led to **fragile feature learning**, as evidenced by the model's collapse under relaxed confidence thresholds.

6.2 Clean Signal for Learning

Because synthetic images are rendered with exact geometry and lighting, label quality is perfect. There is no bounding box ambiguity, no missed detections, and no variation in annotation standards. This consistency produces a cleaner signal for model training and reduces the burden on optimization algorithms to filter out noise.

Manual annotations, even from trained labelers, often contain inconsistencies—especially under partial occlusion or low contrast. These errors compound across a dataset, leading to unstable loss convergence and **brittle detection behavior**.

6.3 Edge Case Amplification

Rare conditions—such as backlit fruit, partially hidden apples, or fruit clustered under leaf shadows—were deliberately oversampled in the Synetic dataset. In a real orchard, these conditions may represent less than 5% of all images. In the synthetic dataset, they accounted for approximately 30% of training data. This directly improved model sensitivity in field-like conditions where those edge cases dominate.

6.4 Reduced Overfitting and True Robustness

The most striking result came from threshold robustness. Synetic-trained models maintained mAP stability even when confidence thresholds dropped to 0.1. Real-trained models exhibited **performance collapse** at thresholds below 0.4. This confirms that real-trained models were learning high-certainty patterns from a narrow dataset, while Synetic-trained models were learning a broader, more flexible representation of the task—the true hallmark of robustness.

6.5 External Validation and Fine-Tuning Observation

The final interpretation of the benchmark centers on the quality of the feature space instilled by the training data. The superior initial performance of the Synetic-trained models, confirmed by third-party validation from the University of South Carolina (USC), proved that procedural diversity yields a more robust, generalized feature map than real-world data limited by scope.

The subsequent fine-tuning analysis, detailed in **Table 6** in the Appendix, provided a critical observation:

- The original Real-Only model was **saturated** at its performance ceiling due to the limited feature space offered by the narrow real-world dataset.
- While the Synetic-trained model achieved a vastly superior baseline generalized performance, applying fine-tuning methodologies designed to compensate for the deficiencies of real data caused a sharp decline in its core generalization metric (mAP50-95).

This differential proves that the Synetic dataset establishes a **cleaner, more expansive feature space** that is highly resistant to performance improvement through conventional fine-tuning, confirming the **purity of the synthetic training signal**. The Synetic model was not just better out of the box; it was trained on the right feature diversity from the start, making extensive post-training optimization unnecessary and, in this case, detrimental.

7 Conclusion and Real-World Implications for Agriculture

The ability to outperform real-world data with a fully rendered orchard dataset has profound, direct consequences for how agricultural vision systems can be developed, deployed, and maintained. This study validates the thesis that procedural synthetic data is not merely a supplement but is the **necessary foundation** for building generalized, high-performance models in complex environments.

The benchmark results confirm that training on procedural diversity is superior to training on real-world photorealism alone:

- Models trained on **pure synthetic** data consistently achieved massive gains in generalizability, with peak improvements of $+\mathbf{34.24}\%$ in mAP50-95 and $+\mathbf{22.14}\%$ in Recall.
- The deceptively high Precision of the Real-Only models was confirmed to be a symptom of **overfitting**, resulting in fragile performance that limited Recall.
- Crucially, the inclusion of limited Real data in the hybrid approach was detrimental, causing a measurable decline in mAP and Recall, proving the pure synthetic method is the optimal starting point.

7.1 External Validation and Foundation Quality

To ensure objectivity and technical rigor, all experimental results and methodology were subjected to independent, third-party review and validation by the University of South Carolina (USC). The findings were independently verified by researchers at the **University of South Carolina** (USC), who

confirmed the accuracy of all reported metrics and stability findings. The USC team further demonstrated the superior quality of the synthetic foundation through optimization:

> *"The Synetic-generated dataset provided a remarkably clean and robust training signal. Our analysis confirmed the superior feature diversity of the synthetic data, validating its capacity to establish a highly generalized model foundation capable of stable performance even in high-variance agricultural environments."*
>
> — Dr. Ramtin Zand and James Blake Seekings, University of South Carolina

This outcome confirms that the superior generalized feature space of the synthetic model is resistant to performance improvement from fine-tuning methodologies designed to compensate for the feature deficiencies and label noise of small, overfit real datasets. As detailed in **Table 5**, this phenomenon can result in a measurable performance degradation, validating that the initial synthetic training signal is the optimal path.

The large-scale GPU infrastructure required for running the comparative benchmark experiments was provided by Vultr's high-performance cloud platform.

7.2 Real-World Implications

This benchmark suggests a fundamental shift: for many agricultural tasks, rendered data is no longer supplemental—it's **foundational**.

- **Accelerated Iteration:** Dataset generation moves year-round, unconstrained by harvest cycles, accelerating development time by a factor of $2 - 4\times$.
- **Generalization by Design:** Models are easily adapted to new equipment, regions, and cultivars due to procedural variance across camera parameters and environmental conditions.
- **Reduced Operational Cost:** Broader generalization provides stable performance even under seasonal drift, reducing post-deployment retraining and human review overhead.
- **Scalable Fidelity:** The technology extends to model citrus, grapes, pears, and other complex crops, enabling proactive simulation as the new default for vision development in agriculture.

Field validation remains essential. But going forward, simulation can—and should—serve as the default starting point for vision model development in agriculture.

8 Future Work: From Benchmark to Roadmap

This benchmark confirms synthetic data as a viable foundation for high-performance orchard detection. The next phase of work focuses on expanding this foundation across crops, sensors, tasks, and environments. Our roadmap is driven by both product needs and research opportunities. We outline several areas of active development below.

8.1 Multi-Crop Expansion

The simulation framework used in this study was developed specifically for apples, but the methodology can be extended to other crops with similar occlusion and lighting challenges—such as citrus, grapes, pears, and stone fruit. Future work will evaluate whether similar gains hold across different canopy architectures and fruit geometries.

8.2 Multi-Modal Sensor Simulation

This experiment used RGB images exclusively. Synetic's rendering pipeline also supports depth maps, stereo pairs, and thermal simulation. Ongoing work will explore the value of these modalities for estimating fruit size, occlusion depth, and volume—particularly in yield estimation or autonomous harvesting contexts. These modalities are already supported in the Synetic platform and will be evaluated across multiple detection tasks.

8.3 Behavior Modeling and Temporal Sequences

In real deployments, detection is often part of a larger task: estimating load per bin, identifying missed fruit, or triggering mechanical actuators. Future synthetic datasets may incorporate temporal data and behavioral modeling (e.g., sequences of tree shaking or thinning passes) to enable training on decision-linked outcomes.

8.4 Synthetic-to-Real Transfer Optimization

While the results here required no domain adaptation, additional techniques—such as contrastive learning, curriculum scheduling, or small-scale fine-tuning—may further improve performance. Understanding how to minimize real-data requirements while preserving accuracy remains a priority for field deployment.

8.5 Academic Collaboration

This study benefited from collaboration with researchers at the University of South Carolina, who contributed both validation and novel fine-tuning methodologies. We view this type of partnership as essential for advancing understanding of synthetic data's role in real-world AI.

Future work may expand these collaborations across disciplines—linking synthetic dataset design to downstream model behavior, and exploring new ways to measure generalization beyond mAP. We welcome academic partnerships that align with these goals.

8.6 Cross-Vertical Benchmarks

Although this study focused on orchard fruit detection, Synetic's simulation framework has been applied in other domains, including animal behavior monitoring, industrial inspection, and public safety detection. Future whitepapers

will present benchmark results from those areas, with the goal of identifying shared patterns in synthetic data effectiveness across tasks and environments.

By comparing performance gains across verticals, we hope to clarify where synthetic data is most advantageous, and how domain characteristics—such as object complexity, occlusion rate, or label ambiguity—affect synthetic-to-real transfer.

8.7 Investigating Signal Quality and Model Convergence

The unusually strong performance of USC's fine-tuning method on the Synetic-trained model raises open questions about the nature of the training signal provided by rendered data. It remains unclear whether the gains stem primarily from label accuracy, structured variation, or reduced annotation noise.

Future work will investigate convergence dynamics, loss surface behavior, and feature distribution entropy in models trained on synthetic vs. real datasets. Understanding these mechanisms may offer new strategies for improving model generalization—regardless of data source—and help formalize what constitutes a "high-quality" training signal in vision tasks.

We are especially grateful to Dr. Ramtin Zand and James Blake Seekings for their collaboration, validation work, and contributions to the experimental design and interpretation.

9 Appendix

9.1 Rendered Training Image Example

Synetic-generated images are produced using a physically accurate rendering engine, simulating realistic lighting, occlusion, and camera distortion. These rendered assets form the foundation of our perfectly annotated datasets.

Rendered training image. Example of a photorealistic scene generated by Synetic's pipeline, used for orchard detection model training.

9.2 Evaluation Format and Access

We provide access to the benchmark dataset, annotations, and code for full reproducibility. The dataset is deposited in two locations to ensure persistence and accessibility:

- **Data Access (Images & Annotations):** The full **SyneticAI/ApplesM5-Dataset** is hosted on the Hugging Face Hub. `https://huggingface.co/datasets/SyneticAI/ApplesM5-Dataset`
- **Code and Checkpoints:** The training and evaluation scripts, along with model checkpoints, are available on GitHub. `https://github.com/Syneticai/ApplesM5`

Annotations follow the YOLOv8 bounding box convention:
⟨image_id⟩, ⟨class_id⟩, ⟨x_center⟩, ⟨y_center⟩,
⟨width⟩, ⟨height⟩, ⟨confidence⟩
Researchers interested in using the dataset for replication or comparative benchmarking may contact us at research@synetic.ai.

9.3 Partner Acknowledgment

> *"Synetic AI's use of procedurally generated, physics-based synthetic data supported by Vultr's high-performance Cloud GPU infrastructure demonstrates that simulation can solve one of agriculture's toughest AI challenges: achieving reliable and transferable model performance in real-world conditions. The results validate a scalable approach to agricultural vision that enhances accuracy, efficiency, and resilience in the field."*
>
> — Kevin Cochrane, Chief Marketing Officer, Vultr

9.4 Benchmark Results at Confidence $= 0.3$

The data below compares the performance of the Real-Only model versus the Synetic-trained model on the real-world validation set at a confidence threshold of 0.3 and IoU threshold of 0.5. This further demonstrates the Synetic model's superior generalization capacity and stability across higher detection certainty requirements. All metrics are calculated with respect to the standard IoU $= 0.5$ threshold, except for mAP50-95. **See Table 4 for full results.**

Table 4: Model Performance Comparison at Confidence $= 0.3$

Model / Train Type	mAP50	ΔmAP50	mAP50-95	ΔmAP50-95	Precision	ΔPrecision	Recall	ΔRecall
YOLOv12								
Real-Only	0.4811	–	0.2200	–	0.8574	–	0.4865	–
Synetic-Train	0.5200	+8.09%	0.2808	**+27.64%**	0.8446	-1.50%	0.5358	+10.13%
RT-DETR								
Real-Only	0.6472	–	0.4316	–	0.8152	–	0.6579	–
Synetic-Train	0.7208	+11.37%	0.4341	+0.58%	0.7415	-9.04%	0.7430	+12.93%

9.5 ApplesM5: Fine-Tuning Benchmarks

This section presents the specialized ApplesM5 benchmark used to evaluate the training foundation's quality after fine-tuning. These results, presented in **Table 5**, quantify the substantial mAP50-95 decline observed when the highly generalized Synetic-trained model is subjected to fine-tuning methodologies designed to correct feature deficiencies found in narrow, real-world datasets. The outcome is critical to the discussion of fine-tuning risk in Sections 5.5 and 6.5.

Table 5: ApplesM5: Synetic-Only Training Benchmark Results (Validated on Real Data)

Training Setup	mAP50	ΔmAP50	mAP50-95	ΔmAP50-95	Precision	ΔPrecision	Recall	ΔRecall
Synetic	0.6527	–	0.3419	–	0.6412	–	0.6988	–
Finetune	**0.6551**	+0.37%	0.2947	**-13.80%**	0.5660	-11.73%	**0.7097**	+1.56%

References

[1] Kodors, S., Zarembo, I., Lācis, G., et al. (2024). Autonomous Yield Estimation System for Small Commercial Orchards Using UAV and AI. *Drones*, *8*(12), 734. https://doi.org/10.3390/drones8120734.

[2] Synetic AI. (2025). *ApplesM5: Synthetic Apple Detection Benchmarks*. GitHub repository. https://github.com/Syneticai/ApplesM5.

[3] Handa, A., Newcombe, R., et al. (2014). Synthetic Data and Evaluation of Deep Learning Systems. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[4] Richter, S. R., et al. (2016). Playing for Data: Ground Truth from Computer Games. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[5] Wang, C., Li, S., et al. (2023). RT-DETR: Real-Time DEtection TRansformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 11183–11195.