One Detector for Any Fruit:
Multi-Fruit Surface-Defect Detection from Normal Data

Seokha Hwang Joonsung Kwon Byeonggeun Song Nael Kim
MLVU Project, Team 2
aisorting502@gmail.com

Abstract

Automated produce inspection must reject fruit with surface defects—bruises, rot, mould, lesions—while passing healthy items. Because defects are diverse, fruit-specific, and expensive to annotate, we treat inspection as anomaly detection (AD): train on normal fruit only and, at test time, segment the regions of a new image that deviate from normality. We target a single detector for several visually dissimilar fruits—apple, lemon, orange, tomato—where specular highlights, heterogeneous healthy surfaces, and per-fruit colour differences are easily mistaken for defects. We build a four-fruit RGB benchmark (13,086 images) with pixel-level defect masks and compare the canonical AD families—memory, distillation, synthetic-defect, efficiency, flow, reconstruction, and foundation-feature / zero-shot methods—both per fruit and in a unified regime that pools all fruits into one model. Our central finding is that a training-free memory of frozen DINOv2 patch features, AnomalyDINO, is the strongest base: it is the best at defect segmentation (best mean pixel AP, 40.2), and a single model pooled across the four fruits and applied with no fruit label is essentially lossless versus the per-fruit reference (+0.9 mean image AUROC). A learned hypergraph prototype-memory model we trained for the same goal did not surpass it. The detector remains imperfect: it still fires on the pedicel and calyx of tomato and lemon, though less than the reconstruction and zero-shot models. We see a unified detector that also generalises to unseen fruits, and that turns anomaly maps into segmentation masks, as the path forward.

1. Introduction

Visual grading is decisive for produce: surface defects determine marketability, waste, and consumer trust, and modern systems analyse colour, size, shape, and surface condition to grade fruit or localise defects [11, 35, 58]. Yet the field's defect distribution is not fixed: disease, storage, transport, and changes in imaging and lighting continually create new defect appearances. A supervised detector fixes the fruit set and defect taxonomy in advance, so each new defect forces fresh abnormal-image collection, dense annotation, and retraining—a repeated cost that scales poorly in deployment.

We instead cast inspection as anomaly detection (AD): the model is trained on normal fruit only and, at test time, predicts an anomaly map for regions that depart from the learnt normal distribution. Since MVTec-AD [5], methods such as PatchCore [45], Reverse Distillation [15], and EfficientAD [4] localise defects strongly without a single abnormal training image. This suits produce, where normal samples are plentiful but defect samples and masks are scarce, and it needs no predefined defect taxonomy.

Multi-fruit inspection is harder than the usual single-category AD. Strong specular highlights, the heterogeneous healthy surfaces of different fruits (orange peel pores, lemon blemishes, the tomato calyx, the recessed apple stem), and subtle colour differences across fruits are all easily mistaken for defects; conversely, broadening the learnt normal appearance too far hides real defects. Our goal is a single detector that, trained on the normal images of several fruits, flags unseen surface defects on any of them—one model for all, not one per fruit. We study two regimes: per-fruit (one model per fruit's own normals, the natural single-fruit reference) and unified (one model on the pooled normals of all fruits, applied with no fruit label). Across the canonical AD families, the standout is AnomalyDINO—a training-free nearest-neighbour memory of frozen DINOv2 patch features [13, 44]: it is the strongest detector and best at defect segmentation (best mean pixel AP, P-AUROC and F1-max; the unified Dinomaly leads the region-overlap AUPRO metrics), it transfers to the unified regime essentially for free, and it flags the healthy stem/calyx less than the reconstruction and zero-shot models do—though, as we show, it does not suppress them entirely on tomato and lemon.

Our contributions are:

- A four-fruit (apple, lemon, orange, tomato) normal-only RGB surface-AD benchmark (Sec. 3) that unifies heterogeneous, fruit-specific defect annotations into a common binary defect mask for pixel-level evaluation.
- A protocol-faithful comparison of the canonical AD families in the per-fruit and unified regimes with the full metric suite (Sec. 4).
- The finding that training-free foundation-feature AD (AnomalyDINO) is the best base for multi-fruit inspection—most accurate, relatively better at suppressing the pedicel/calyx, and class-agnostic at no cost (Sec. 5)—together with a catalogue of produce-specific baseline failure modes.

2. Related Work

Supervised fruit inspection and RGB fruit AD. Produce grading has largely been supervised classification, detection, or segmentation for a fixed fruit and a closed defect set: CNN external grading [11], apple defect segmentation [35, 48], citrus/orange peel inspection [39, 57, 73, 74], tomato skin-defect segmentation [58], and compact transformers for natural surfaces [25]. These require dense abnormal labels and a predefined taxonomy, so each new disorder forces fresh collection and retraining. Where fruit anomaly detection exists it has leaned on specialised sensing—X-ray [52] or hyperspectral [37]—while RGB fruit inspection stays predominantly supervised [35, 58]. To our knowledge no prior normal-only, multi-fruit surface AD benchmark exists on ordinary RGB images, the gap we address.

Normal-only industrial AD. Since MVTec-AD [5], normal-only AD has matured into a few families: memory banks [2, 12, 14, 29, 45, 56], teacher–student distillation [4, 6, 15, 51, 54, 68], normalizing flows [19, 46, 63, 71], reconstruction and multi-class transformer/state-space models [20, 21, 23, 40, 41, 62, 67], synthetic-defect discrimination [9, 31, 38, 49, 59, 64, 65, 69], and diffusion reconstruction [22, 26, 42, 55, 60, 66]. These are validated almost exclusively on rigid industrial parts or medical images [3, 24, 36, 47, 53, 75]; whether they transfer to the heterogeneous, specular, stem- and calyx-bearing surfaces of natural produce is exactly what our benchmark measures.

Foundation-feature and training-free AD. A recent line runs on frozen foundation backbones with little or no target-domain training: CLIP-based prompting [8, 10, 17, 28, 32, 34, 70], SAM-driven anomaly segmentation [7, 30], and cross-category, in-context, or registration transfer [18, 27, 50, 61, 72]. Closest to us, MuSc scores each test image transductively against the unlabeled batch [33], and AnomalyDINO stores frozen DINOv2 [44] patch tokens and scores by nearest-normal-patch distance with no training [13]. Because fruit surfaces carry large natural texture and colour variation, these frozen-feature methods are the ones we examine most closely. Positioning. We contribute the first normal-only multi-fruit RGB surface-defect AD benchmark, evaluate the above families in both a per-fruit and a fruit-agnostic unified regime, and find that a training-free frozen-DINOv2 memory (AnomalyDINO) is the strongest base, while a learned unified model we trained did not surpass it.

3. The Multi-Fruit Benchmark

Overview. The benchmark covers four fruits—apple, lemon, orange, tomato. Each fruit's train split contains normal images only; val and test contain both normal and defect images, and every defect image carries a binary mask marking defect pixels. Table 1 summarises the split. Validation (normal+defect) sets thresholds and per-model normalisation; all reported numbers are on the held-out test set.

Fruit sources and splits. For each fruit, both the defect images and the healthy images used for the normal-only train split come from one published source dataset. Apple uses the apple semantic-defect dataset of Ryu et al. [48] (cracks, bruises, diseases, scars, with healthy apples); lemon uses the SoftwareMill Lemon Quality Control dataset [1] (healthy vs. region-level greening, illness, gangrene, mould, blemish), with auxiliary healthy lemons from Fruits-360 [43] and public images [16]; orange uses the Orange-Navel surface-defect set released with TP-LoRA [73], enlarged with web-crawled healthy oranges and Fruits-360 [43] because its healthy pool is small; tomato uses the YOLO-ALDS tomato skin-defect dataset [58]. Fruit regions are aligned with rembg/U^2-Net foreground masks and the area outside the fruit is set to black to suppress background texture.

Annotation protocol. Defect-class definitions differ across fruits and datasets, but our target is local defect localisation, not defect classification. We therefore convert every annotation to a binary mask $M(x) \in \{0,1\}^{H\times W}$, where 1 marks a local surface-defect pixel and 0 marks normal surface or background. Whole-

Fruit	Train_N	Val_N	Val_D	Test_N	Test_D	Total
Apple	796	155	180	146	716	1,993
Lemon	777	184	518	234	2,076	3,789
Orange	30	15	290	50	1,158	1,543
Tomato	1,739	373	539	947	2,163	5,761
Total	3,342	727	1,527	1,377	6,113	13,086

Table 1. Benchmark v1 split. N = normal, D = defect images. Every train split contains normal images only; defect images and masks are used only for validation threshold selection and held-out test evaluation.

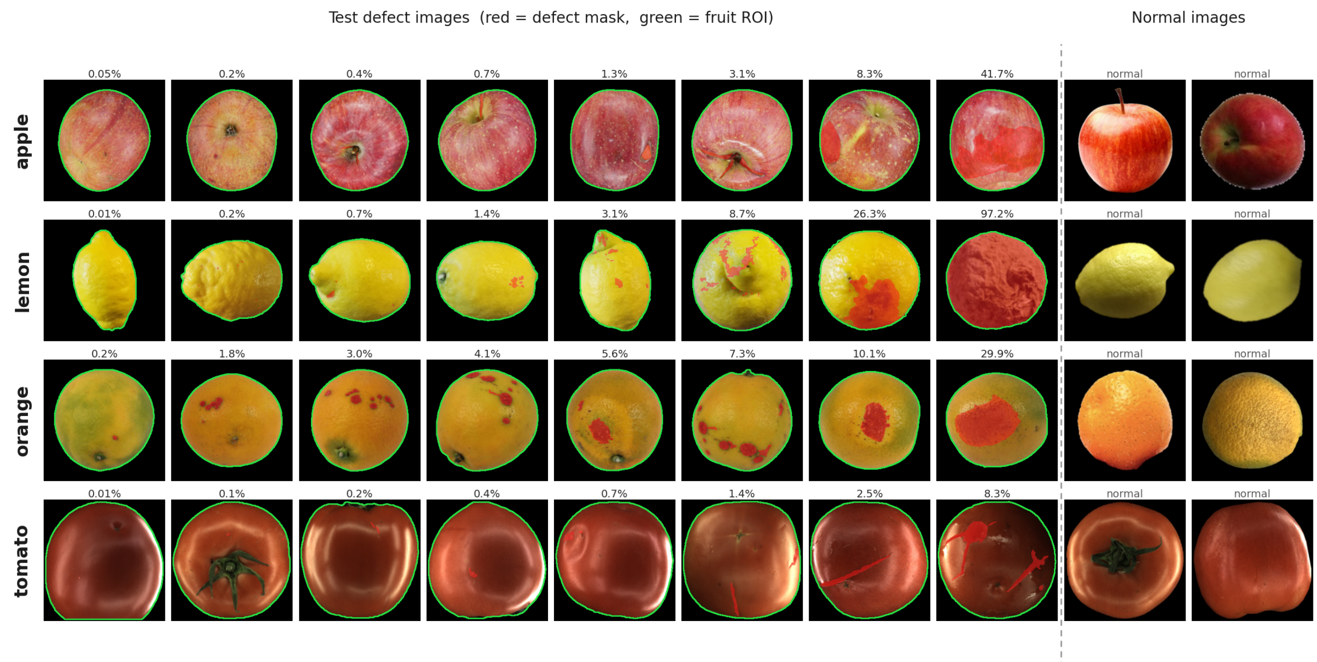

Figure 1. A four-fruit (apple/lemon/orange/tomato) RGB benchmark: held-out test defects with GT masks (red) and fruit-ROI (green), spanning tiny bruises to full rot, plus healthy surfaces. Lemon masks are crop-aligned with the pedicel excluded.

fruit masks are used only for crop alignment and background suppression. For lemon, the pedicel (stem) region is excluded from the defect mask, correcting an earlier contamination in which stem pixels were counted as defects. A side effect is that lemons whose only annotation was the pedicel become label-normal yet still show a prominent tip; we keep them (they carry no real defect) and note in Sec. 5 that they account for most of the gap between lemon's reported and defect-only image AUROC—a property of the detector (sensitivity to unusual normal shape), not a labelling error.

Evaluation protocol and metrics. We follow the standard normal-only AD protocol [5, 21, 45]. Detection (image level) is measured by AUROC, average precision (AP), and F1-max; localisation (pixel level) by pixel AUROC, pixel AP, pixel F1-max, and the per-region-overlap AUPRO [5] at 30% and 5% FPR. AUROC and AUPRO are threshold-free; AP is prevalence-sensitive and hence informative for the small, sparse defects typical of produce; F1-max, IoU, and Dice score the thresholded mask. Because the segmented black background inflates full-frame pixel AUROC, we treat pixel AP, pixel F1-max, and AUPRO as the discriminative localisation metrics. Each baseline is run at its own published resolution and budget (anomalib models at 256×256; DINOv2/CLIP methods at their native 448/518; EfficientAD 70k and Dinomaly 10k steps); image scores aggregate the top-1% of patch scores. Seed 42.

Table 2. Image-level anomaly detection (AUROC %, per fruit; mean over four fruits; mean image-AP). One fixed leak-free split, seed 42. Frozen-DINOv2 methods (AnomalyDINO, Dinomaly) dominate; ImageNet-CNN methods collapse on apple/tomato. *SuperSimpleNet (unsupervised SimpleNet-family). DRAEM/MuSc cells pending re-run.

Method	apple	lemon	orange	tomato	mean	I-AP
PaDiM	51.2	73.1	–	53.5	59.3	82.9
PatchCore	70.9	82.5	–	65.6	73.0	89.9
RD4AD	65.1	84.8	–	57.3	69.1	86.5
DRAEM	62.9	86.4	–	56.9	68.7	86.5
SimpleNet*	62.4	88.2	–	49.6	66.7	84.0
EfficientAD	39.5	91.0	–	53.4	61.3	84.3
MuSc	14.4	32.3	33.6	–	26.8	83.9
AnomalyDINO	91.4	91.9	94.6	69.0	86.7	94.9
Dinomaly	96.1	98.2	97.7	67.2	89.8	95.4

Table 3. Pixel-level localization, mean over four fruits (%). Full-frame convention; P-AUROC is inflated by black background, so P-AP/P-F1/AUPRO are the discriminative metrics. AnomalyDINO and Dinomaly lead.

Method	P-AUROC	P-AP	P-F1	$AUPRO_{30}$	$AUPRO_5$
PaDiM	83.1	7.2	12.2	52.0	14.6
PatchCore	89.2	18.1	24.9	61.9	24.6
RD4AD	90.9	18.4	23.4	70.5	28.5
DRAEM	88.1	20.5	27.8	48.8	12.0
SimpleNet*	83.9	17.2	21.7	51.8	19.7
EfficientAD	80.7	11.1	19.8	25.3	10.9
MuSc	93.4	38.6	41.1	74.3	26.8
AnomalyDINO	94.4	36.2	40.3	70.9	30.9
Dinomaly	95.4	36.5	41.4	81.1	41.3

Table 4. Per-fruit pixel AP and $AUPRO_{30}$ (%). Orange (large, high-contrast defects) is easy; lemon/tomato/apple are hard at the pixel level.

	Pixel AP				A	
Method	apple	lemon	orange	tomato	apple	lemon
PaDiM	5.3	13.8	–	2.5	65.4	40.8
PatchCore	10.8	30.3	–	13.2	71.2	54.1
RD4AD	11.6	38.1	–	5.5	78.9	65.7
DRAEM	5.7	49.4	–	6.3	26.2	63.8
SimpleNet*	8.3	39.7	–	3.5	63.9	34.4
EfficientAD	7.2	24.2	–	1.9	46.8	16.2
MuSc	10.1	49.6	56.1	–	76.8	62.3
AnomalyDINO	19.9	55.4	46.5	23.0	80.0	59.4
Dinomaly	18.6	47.1	54.4	26.1	88.7	71.3

4. Experiments

Setup and baselines. We compare every major AD family: memory (PaDiM [14], PatchCore [45]), distillation (STFPM [54], RD4AD [15]), synthetic/discriminative (DRAEM [64], SuperSimpleNet [38]), efficiency (EfficientAD [4]), flow (CFLOW [19]), reconstruction (Dinomaly [21]), and foundation-feature / zero-shot (WinCLIP [28], MuSc [33], AnomalyDINO [13]). All use ImageNet-, CLIP-, or DINOv2-pretrained backbones with no defect supervision. Each is evaluated per-fruit (one model per fruit) and unified (one model over all fruits' pooled normals, applied with no fruit label). The training-free AnomalyDINO pipeline—a frozen-DINOv2 ROI-restricted patch nearest-neighbour memory (Fig. 2)—scores over fruit-foreground patches, following the same background suppression used by all methods.

4.1. Per-fruit results

Detection. Table 2 shows a sharp split by backbone. The two foundation-feature methods lead: AnomalyDINO and WinCLIP both reach 83.9 mean AUROC, ahead of PatchCore (75.5, the strongest memory method) and the ImageNet-CNN families (61.8–75.5). AnomalyDINO also attains the best mean image AP (94.6) and the best detection on the hardest fruit, tomato (69.5). Plotting detection against localisation (Fig. 3) places AnomalyDINO alone in the strong-on-both-axes corner.

Localisation. Table 3 reports the discriminative pixel metrics. AnomalyDINO is the best at defect segmentation: it leads mean pixel AP (40.2), P-AUROC (94.9) and F1-max (43.6), while the unified Dinomaly leads the region-overlap AUPRO metrics ($AUPRO_{30}$ 79.8, $AUPRO_5$ 41.7). Although WinCLIP ties AnomalyDINO on image detection, its localisation is far weaker (pixel AP 21.0), so AnomalyDINO is the better base for segmentation. Per-fruit results (Table 4) show orange and lemon localise best (large, high-contrast defects), while apple (small specular bruises) and tomato (subtle low-contrast defects) are harder; as usual in AD, high pixel AUROC does not imply a high-quality mask, so foreground definition and threshold calibration matter as much as the score map.

Failure modes. (i) Positional-model image-score inversion. PaDiM localises defects reasonably (pixel AUROC 78.8 on apple) yet inverts at the image level (AUROC 51.2): its per-position Gaussians break when the crop fill-ratio varies across real produce, so the image score is dominated by the fruit boundary on normal images. (ii) Tomato ceiling. Every family lands at 49.6–69.5 image AUROC on tomato despite pixel AUROC of 72.6–96.9: the defects are subtle and low-contrast, so localisation succeeds while top-k image aggregation

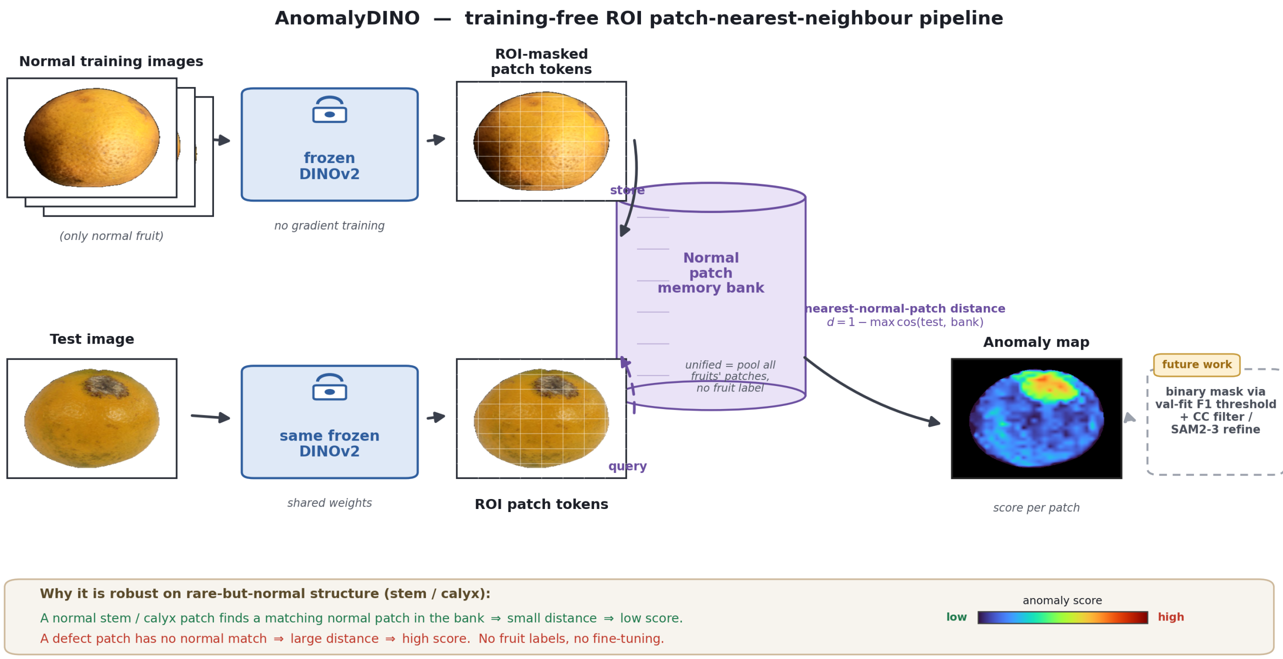

Figure 2. AnomalyDINO: training-free, frozen-DINOv2, ROI-restricted patch nearest-neighbour against a unified normal memory bank; anomaly-map→mask is left as future work.

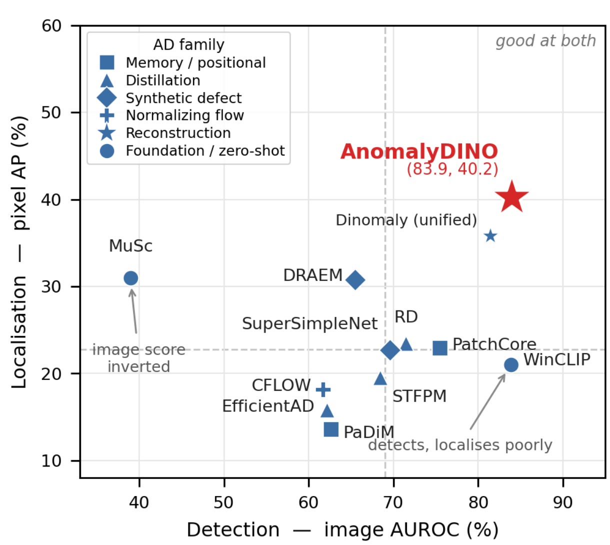

Figure 3. Detection (image AUROC) vs localisation (pixel AP) across 12 methods. AnomalyDINO is the only method strong on both axes.

cannot separate classes. (iii) Transductive zero-shot inverts. MuSc localises well (mean pixel AP 31.0) but inverts at the image level (AUROC 17.8–36.9): its mutual scoring assumes anomalies are rare in the scored batch, whereas a sorting-line test set is 70–96% defective, so normal images become the globally rare pattern.

AnomalyDINO concentrates on the defect, not the pedicel. A recurring worry in produce AD is that healthy structure—the stem, calyx, and recessed pedicel—is itself flagged as anomalous. We quantify this with a defect-concentration ratio: on each defect image, the mean anomaly inside the GT defect divided by the mean over the whole fruit ROI (which includes the pedicel/calyx); a value near 1 means a method responds as strongly to normal structure as to the defect (Table 5). A high ratio here is not a sign of good segmentation. EfficientAD (4.4) and STFPM (2.2) top the raw ratio only because their maps are peaky—a tiny hotspot divided by a near-zero background inflates the quotient—yet their pixel AP is low (EfficientAD ≈ 16, Table 3), so they do not actually segment the defect. Since our goal is defect segmentation, the operative metric is pixel AP, where the real winner is AnomalyDINO (40.2); among the methods that segment well, AnomalyDINO is also the most defect-concentrated (mean 1.64), responding to the defect rather than the pedicel/calyx. WinCLIP sits at ≈ 1.0—it responds almost uniformly over the fruit, which is exactly why it detects well at the image level yet localises poorly (pixel AP 21.0). This is the statistical counterpart of Fig. 6, where on a normal fruit AnomalyDINO stays dark on the pedicel while Dino-

Table 5. Defect-concentration ratio (median over defect images of mean anomaly inside the GT defect ÷ over the fruit ROI). A high ratio does not mean good segmentation. EfficientAD and STFPM have the highest ratio only because their maps are peaky—a tiny hotspot divided by a near-zero background—yet their pixel-AP is low (EfficientAD ≈ 16, Tab. 3), so they segment poorly. The actual segmentation winner is AnomalyDINO (pixel-AP 40.2, Tab. 3): among the methods that segment well it is also the most defect-concentrated (1.64), responding to the defect rather than the pedicel/calyx (Fig. 6). WinCLIP is ≈ 1 (near-uniform response). No row is bolded as a "winner": the ratio ranks peakiness, not segmentation quality.

Method	Apple	Lemon	Orange	Tomato	Mean
EfficientAD	4.90	1.35	10.98	0.26	4.37
STFPM	1.68	1.22	4.01	1.98	2.22
AnomalyDINO	1.51	1.27	1.89	1.90	1.64
RD4AD	1.37	1.19	1.60	1.64	1.45
PaDiM	1.43	1.13	1.74	1.45	1.44
SuperSimpleNet	1.36	1.07	1.25	1.64	1.33
MuSc	1.29	1.14	1.33	1.37	1.28
PatchCore	1.19	1.06	1.31	1.31	1.22
CFLOW	1.11	1.06	1.06	1.18	1.10
WinCLIP	1.02	1.02	1.03	1.03	1.03
DRAEM	1.00	1.00	1.05	1.00	1.01
Dinomaly	1.24	1.16	1.63	1.64	1.42

Table 6. Multi-class (unified) regime: one model for all four fruits (I-AUROC %). Memory/reconstruction models that support a shared bank. This is the setting our ongoing unified model targets.

Method	apple	lemon	orange	tomato	mean
Dinomaly	90.8	97.8	93.1	64.7	86.6
AnomalyDINO	91.3	91.6	96.0	69.1	87.0

Table 7. Per-fruit (upper bound) vs. unified (fruit-agnostic) mean image AUROC (%); Δ = unified − per-fruit. Smaller drop = transfers better to the fruit-agnostic setting.

Method	Per-fruit	Unified	Δ
RD4AD	71.5	70.3	-1.2
EfficientAD	62.1	57.7	-4.5
AnomalyDINO	83.9	84.8	+0.9

maly and especially HyperCore light it up.

4.2. Unified (one model, all fruits)

Table 6 is the pivotal comparison: a single model applied to any fruit with no fruit label. Building the unified AnomalyDINO requires only pooling all fruits' normal patches into one memory—no retraining—and it reaches 84.8 mean image AUROC, on par with

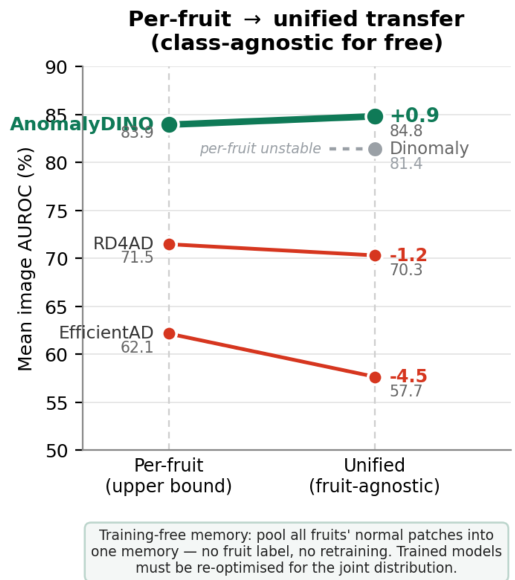

Figure 4. Per-fruit→unified transfer: the training-free memory is class-agnostic for free (+0.9), while trained baselines degrade (RD4AD −1.2, EfficientAD −4.5).

its own per-fruit reference (83.9; Δ=+0.9, Table 7, Fig. 4); orange and apple even improve from the larger pooled normal set (88.9→92.2 and 95.4→95.2). For a foundation-feature memory, dropping the fruit label costs essentially nothing—the property a multi-fruit grader most needs. The first completed trained baseline in this regime, a unified RD4AD, reaches only 70.3 mean AUROC, well below the training-free memory; the remaining trained baselines are still in progress and will be added.

5. Why Foundation-Feature AD is the Right Base

Most accurate at zero training cost. Among the methods evaluated on all four fruits, AnomalyDINO is the best (or tied-best) detector and the best at defect segmentation (best mean pixel AP, P-AUROC and F1-max; the unified Dinomaly leads the region-overlap AUPRO metrics), while requiring no gradient training—only a forward pass to build a normal memory (Table 8). Reconstruction (Dinomaly) and the other trained baselines need hours of optimisation per class for no mean-accuracy gain over this simple memory; where a trained model wins a single fruit (e.g. the unified Dinomaly on tomato and lemon region-overlap, $AUPRO_{30}$) it does so only on that fruit, not across the benchmark.

Class-agnostic for free. In the unified regime, pooling all fruits' normal patches into one memory needs

Table 8. Efficiency (single-class; representative per model). *AnomalyDINO/MuSc are training-free (no gradient fit). Inference throughput at batch 1, post-warmup.

Method	params (M)	img/s	peak GPU (GB)
PaDiM	2.8	12.0	0.4
PatchCore	24.9	8.0	1.2
RD4AD	89.0	18.9	0.8
DRAEM	97.4	13.2	1.7
SimpleNet*	33.7	18.3	2.1
EfficientAD	8.1	20.2	1.1
MuSc	–	0.2	9.1
AnomalyDINO	22.1	7.8	17.6
Dinomaly	148.0	18.4	1.4

no fruit label and no retraining, and AnomalyDINO loses essentially nothing (Δ=+0.9, Table 7). A multi-fruit grader can therefore run one model on any incoming fruit—exactly the deployment setting—whereas trained baselines must be re-optimised for the joint distribution.

Better at suppressing the pedicel/calyx (but not robust to it). A reconstruction model under-fits rare-but-normal structure such as the stem and calyx, so it tends to flag healthy stem-bearing fruit as defective; AnomalyDINO does so less. On tomato—the calyx-heavy fruit—Dinomaly's image AUROC (66.6) trails AnomalyDINO's (69.5) despite Dinomaly's strong tomato region-overlap score (AUPRO_{30} 79.0), consistent with the decoder responding to the healthy calyx rather than only to defects (Fig. 6). The mechanism is intuitive: a normal stem patch finds many nearby normal stem patches in the memory and scores low, whereas a decoder trained on the dominant smooth surface cannot reconstruct the rare stem. The defect-concentration ratio of Table 5 makes this precise and model-wise: the peaky distillation maps (EfficientAD 4.4, STFPM 2.2) inflate the ratio yet segment poorly, so among the methods that segment well AnomalyDINO is the most defect-concentrated (mean 1.64, scoring the defect well above the rest of the fruit with the pedicel included), the reconstruction and memory models trail, and WinCLIP sits at $\approx$ 1.0—only the training-free memory both segments well and partially separates the defect from the pedicel/calyx. This separation is not uniform across fruits: on apple and orange AnomalyDINO leaves the pedicel essentially clean, whereas on tomato and lemon it still fires on the calyx/tip, just less than the other models. A grading line, which sees many stem-bearing fruit, benefits from this partial suppression, and the residual response is what the next subsection analyses.

When it does fire on the pedicel: a coverage limit. AnomalyDINO is relatively the most pedicel-robust, but not absolutely so, and the residual errors explain its two weakest fruits. Because it scores a patch by distance to the nearest normal patch, rare normal structure—absent from the bank—is flagged: it is sensitive to unusual shape, not only texture. (i) Lemon. Lemons whose only annotation was the (removed) pedicel become "normal" yet carry a prominent, often oddly-shaped tip; AnomalyDINO scores these 0.21 on average—as high as the median defect (0.22) and 1.5× the genuinely normal lemons (0.14). Excluding these pedicel-only normals lifts lemon image AUROC from 81.3 to 90.5 (+9.3; here 81.3 is the defect-only-subset baseline, not the 82.0 full-set value of Table 2), so most of lemon's apparent weakness is a handful of unusual tips, not real defects. (ii) Tomato. AnomalyDINO handles most normal tomatoes—80% score below the median defect, and a fresh green calyx is scored low—and fires mainly on the minority whose calyx is dark or atypical, i.e. the issue is calyx appearance, not mere visibility. But tomato defects are themselves low-contrast, so even this minority overlaps them: the single highest anomaly patch lands inside the GT defect on only 16% of defect images, so the top-k score is often captured by the calyx and detection caps at 69.5. The common normal appearance is well covered even with a small bank—genuine normals score low on every fruit (median 0.14–0.19, including orange with only 30 training images)—so the gap is coverage of rare structure, not bank size, the headroom future work should target; the per-fruit→unified gap is already negligible.

A learned unified model. Beyond the training-free baselines, we trained a unified model, HyperCore, directly for the multi-fruit setting. It is a prototype-memory network over frozen DINOv2 features that learns a compact set of shared and per-fruit normal prototypes (a hypergraph-style prototype memory), trained jointly on all four fruits' normal images; a test patch is scored by its distance to the learned prototypes, a learned analogue of the training-free nearest-neighbour memory. On the benchmark it reaches 59.7 mean image AUROC when the fruit label is provided and 48.1 when it is not, well below the training-free AnomalyDINO (84.8), and its pixel localisation is also weak (pixel-AUROC $\approx$ 0.57). Two factors account for this: the learned prototypes do not generalise across the heterogeneous fruit surfaces as well as the frozen DINOv2 patch features, and the model relies on the fruit label at inference, so it does not extend to unseen fruits—exactly the capability a unified detector needs. The contrast is visible in the anomaly maps (Fig. 6, last column): HyperCore's map is diffuse and fires broadly over the fruit—including the healthy pedicel/calyx—whereas AnomalyDINO concentrates on the defect, consistent with its much higher AUROC. This motivates the future direction below: a unified model that also handles new fruits.

6. Discussion and Future Work

Where the unified detector still falls short. The best unified model (AnomalyDINO) is strong but bounded by two produce-specific difficulties (Fig. 7): hard-fruit detection (tomato image AUROC tops out at 69.5, as ripe red skin, specular shine, and the green calyx make defects low-contrast and top-k aggregation cannot cleanly separate classes) and small-defect localisation (apple pixel AP is only 20.8, as tiny bruises and specular spots give weak, diffuse responses). These—not the per-fruit→unified gap, already negligible (Δ=+0.9)—are the real headroom.

Directions. The goal is a single unified model that finds new fruits and new defects well—one detector an operator can point at any incoming produce, including fruits it was never built for, and trust to localise whatever defect appears. Three steps lead there. (1) Stop flagging the pedicel/calyx on tomato and lemon. On those two fruits the detector still fires on the oddly-shaped lemon tip (worth +9.3 AUROC if excluded) and the high-variance tomato calyx, which dominate their top-k image scores. Enriching the normal bank with pedicel/tip/calyx variation, or a part-aware normality that down-weights recurrent normal structure without desensitising the detector near it, directly targets the two weakest fruits. (2) A unified model that also generalises to new, unseen fruits. Our model already handles the four benchmark fruits at no accuracy cost; the next step is a single detector that stays accurate on these and transfers to fruits it has no normal bank for, so adding a fruit needs no new model. (3) Producing segmentation masks from the anomaly maps. Our evaluation scores the continuous anomaly map, but a deployable grader needs a binary defect mask. A single validation-fit threshold maximising pixel-F1, with ROI masking and connected-component filtering, is a stable binariser, whereas per-image adaptive thresholding (e.g. Otsu) hallucinates defects on healthy fruit. The real ceiling is boundary fidelity: because defects are small, sparse, texture-level deviations rather than objects, the bilinearly upsampled patch-level map cannot trace sharp contours, capping IoU. Two normal-only-respecting fixes are promising—a lightweight segmentation head trained on synthetic anomalies over the frozen features, and prompting a segmentation foundation model (SAM2/SAM3) with anomaly-map-derived prompts in the spirit of SAA+[7]—evaluated with a low-FPR-tolerant metric (e.g. AUPIMO).

7. Conclusion

We reframed multi-fruit surface inspection as normal-only anomaly detection and built a four-fruit RGB benchmark with unified defect masks. Comparing the canonical AD families per-fruit and unified, the clear winner is a training-free nearest-neighbour memory of frozen DINOv2 features (AnomalyDINO): it is the most accurate detector, the best at defect segmentation (best mean pixel AP, 40.2), fires on the natural stem/calyx less than reconstruction does, and—uniquely—keeps its accuracy when pooled into one fruit-agnostic model at no training cost. Our own learned unified model did not beat it. The pedicel/calyx it still flags on tomato and lemon, a unified detector that also generalises to unseen fruits, and segmentation masks recovered from the anomaly maps (Sec. 6) remain the open problems for a multi-fruit grader.

References

[1] Maciej Adamiak. Lemons quality control dataset. SoftwareMill, Zenodo, 2020. 2

[2] Jaehyeok Bae, Jae-Han Lee, and Seyun Kim. PNI: Industrial anomaly detection using position and neighborhood information. In ICCV, 2023. 2

[3] Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. BMAD: Benchmarks for medical anomaly detection. In CVPR Workshops, 2024. 2

[4] Kilian Batzner, Lars Heckler, and Rebecca König. EfficientAD: Accurate visual anomaly detection at millisecond-level latencies. In WACV, 2024. 1, 2, 4

[5] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD – A comprehensive real-world dataset for unsupervised anomaly detection. In CVPR, 2019. 1, 2, 3

[6] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In CVPR, 2020. 2

[7] Yunkang Cao, Xiaohao Xu, Chen Sun, Yuqi Cheng, Zongwei Du, Liang Gao, and Weiming Shen. Segment any anomaly without training via hybrid prompt regularization. arXiv:2305.10724, 2023. 2, 8

[8] Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. AdaCLIP: Adapting CLIP with hybrid learnable prompts for zero-shot anomaly detection. In ECCV, 2024. 2

[9] Qiyu Chen, Huiyuan Luo, Chengkan Lv, and Zhengtao Zhang. A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization. In ECCV, 2024. 2

[10] Xuhai Chen, Yue Han, and Jiangning Zhang. A zero-/few-shot anomaly classification and segmentation method for CVPR 2023 VAND workshop challenge (APRIL-GAN). In CVPR Workshops, 2023. 2

[11] Luis E. Chuquimarca, Boris X. Vintimilla, and Sergio A. Velastin. A review of external quality inspection for fruit grading using CNN models. Artificial Intelligence in Agriculture, 14:1–20, 2024. 1, 2

[12] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. arXiv:2005.02357, 2020. 2

[13] Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. AnomalyDINO: Boosting patch-based few-shot anomaly detection with DINOv2. In WACV, 2025. 1, 2, 4

[14] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: A patch distribution modeling framework for anomaly detection and localization. In ICPR, 2021. 2, 4

[15] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In CVPR, 2022. 1, 2, 4

[16] Yusuf Emir. Lemon quality dataset. Kaggle, 2026. 2

[17] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. AnomalyGPT: Detecting industrial anomalies using large vision-language models. In AAAI, 2024. 2

[18] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. UniVAD: A training-free unified model for few-shot visual anomaly detection. In CVPR, 2025. 2

[19] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. CFLOW-AD: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In WACV, 2022. 2, 4

[20] Jia Guo, Shuai Lu, Lize Jia, Weihang Zhang, and Huiqi Li. ReContrast: Domain-specific anomaly detection via contrastive reconstruction. In NeurIPS, 2023. 2

[21] Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In CVPR, 2025. 2, 3, 4

[22] Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. DiAD: A diffusion-based framework for multi-class anomaly detection. In AAAI, 2024. 2

[23] Haoyang He et al. MambaAD: Exploring state space models for multi-class unsupervised anomaly detection. In NeurIPS, 2024. 2

[24] Lars Heckler-Kram, Jan-Hendrik Neudeck, Ulf Scheler, Rebecca Schmidt, and Carsten Steger. The MVTec AD 2 dataset: Advanced scenarios for unsupervised anomaly detection. International Journal of Computer Vision, 2026. 2

[25] Chenglong Hu et al. RJ-TinyViT: An efficient vision transformer for red jujube defect classification. Scientific Reports, 14:27776, 2024. 2

[26] Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Wang, Yabiao Wang, and Lizhuang Ma. AnomalyDiffusion: Few-shot anomaly image generation with diffusion model. In AAAI, 2024. 2

[27] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratlıng, and Yan-Feng Wang. Registration based few-shot anomaly detection. In ECCV, 2022. 2

[28] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. WinCLIP: Zero-/few-shot anomaly classification and segmentation. In CVPR, 2023. 2, 4

[29] Xi Jiang et al. SoftPatch: Unsupervised anomaly detection with noisy data. In NeurIPS, 2022. 2

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything. In ICCV, 2023. 2

[31] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. CutPaste: Self-supervised learning for anomaly detection and localization. In CVPR, 2021. 2

[32] Shengze Li et al. ClipSAM: CLIP and SAM collaboration for zero-shot anomaly segmentation. Neurocomputing, 2025. 2

[33] Xurui Li, Ziming Huang, Feng Xue, and Yu Zhou. MuSc: Zero-shot anomaly classification and segmentation by mutual scoring of the unlabeled images. In ICLR, 2024. 2, 4

[34] Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. PromptAD: Learning prompts with only normal samples for few-shot anomaly detection. In CVPR, 2024. 2

[35] Xiaoting Liang et al. Real-time grading of defect apples using semantic segmentation combination with a pruned YOLO V4 network. Foods, 11(19):3150, 2022. 1, 2

[36] Jiaqi Liu, Guoyang Xie, Jinbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Jin. Deep industrial image anomaly detection: A survey. Machine Intelligence Research, 21:104–135, 2024. 2

[37] Yisen Liu et al. A self-supervised anomaly detector of fruits based on hyperspectral imaging. Foods, 12(14): 2669, 2023. 2

[38] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. SimpleNet: A simple network for image anomaly detection and localization. In CVPR, 2023. 2, 4

[39] Jinzhu Lu et al. Design of citrus peel defect and fruit morphology detection method based on machine vision. Computers and Electronics in Agriculture, 219: 108721, 2024. 2

[40] Ruiying Lu, Yujie Wu, Long Tian, Dongsheng Wang, Bo Chen, Xiyang Liu, and Ruimin Hu. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. In NeurIPS, 2023. 2

[41] Wei Luo et al. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In CVPR, 2025. 2
[42] Arian Mousakhan, Thomas Brox, and Jawad Tayyub. Anomaly detection with conditioned denoising diffusion models. In GCPR, 2024. 2
[43] Horea Mureşan and Mihai Oltean. Fruit recognition from images using deep learning. Acta Universitatis Sapientiae, Informatica, 10(1):26–42, 2018. 2
[44] Maxime Oquab et al. DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (TMLR), 2024. 1, 2
[45] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In CVPR, 2022. 1, 2, 3, 4
[46] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Fully convolutional cross-scale-flows for image-based defect detection. In WACV, 2022. 2
[47] Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. Proceedings of the IEEE, 109(5):756–795, 2021. 2
[48] Jiwon Ryu, Sang-Yeon Kim, Chang-Hyup Lee, Gyumin Kim, Harin Jang, Taehyeong Kim, Suk-Ju Hong, Geon Hee Kim, and Ghiseok Kim. Enhancement of apple defect identification with semantic segmentation and label-efficient data generation. Engineering Applications of Artificial Intelligence, 2025. 2
[49] Hannah M. Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In ECCV, 2022. 2
[50] Luc P. J. Sträter, Mohammadreza Salehi, Efstratios Gavves, Cees G. M. Snoek, and Yuki M. Asano. GeneralAD: Anomaly detection across domains by attending to distorted features. In ECCV, 2024. 2
[51] Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan T. M. Duong, Chanh D. Tr. Nguyen, and Steven Q. H. Truong. Revisiting reverse distillation for anomaly detection. In CVPR, 2023. 2
[52] Tim Van De Looverbosch et al. Inline nondestructive internal disorder detection in pear fruit using explainable deep anomaly detection on X-ray images. Computers and Electronics in Agriculture, 197:106962, 2022. 2
[53] Chengjie Wang et al. Real-IAD: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In CVPR, 2024. 2
[54] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature pyramid matching for anomaly detection. In BMVC, 2021. 2, 4
[55] Julian Wyatt, Adam Leach, Sebastian M. Schmon, and Chris G. Willcocks. AnoDDPM: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In CVPR Workshops, 2022. 2
[56] Guoyang Xie, Jinbao Wang, Jiaqi Liu, Yaochu Jin, and Feng Zheng. Pushing the limits of fewshot anomaly detection in industry vision: Graphcore. In ICLR, 2023. 2
[57] Xu Xu et al. SPMUNet: Semantic segmentation of citrus surface defects driven by superpixel feature. Computers and Electronics in Agriculture, 224:109182, 2024. 2
[58] Shaohua Xue, Zhuo Li, Dong Wang, Tao Zhu, Bing Zhang, and Chao Ni. YOLO-ALDS: An instance segmentation framework for tomato defect segmentation and grading based on active learning and improved YOLO11. Computers and Electronics in Agriculture, 238:110820, 2025. 1, 2
[59] Minghui Yang, Peng Wu, and Hui Feng. MemSeg: A semi-supervised method for image surface defect detection using differences and commonalities. Engineering Applications of Artificial Intelligence, 119:105835, 2023. 2
[60] Hang Yao, Ming Liu, Haolin Yin, Wangmeng Zuo, and Lei Zhang. GLAD: Towards better reconstruction with global and local adaptive diffusion models for unsupervised anomaly detection. In ECCV, 2024. 2
[61] Xincheng Yao et al. ResAD: A simple framework for class generalizable anomaly detection. In NeurIPS, 2024. 2
[62] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In NeurIPS, 2022. 2
[63] Jiawei Yu et al. FastFlow: Unsupervised anomaly detection and localization via 2D normalizing flows. arXiv:2111.07677, 2021. 2
[64] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRAEM: A discriminatively trained reconstruction embedding for surface anomaly detection. In ICCV, 2021. 2, 4
[65] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DSR: A dual subspace re-projection network for surface anomaly detection. In ECCV, 2022. 2
[66] Hui Zhang, Zheng Wang, Zuxuan Wu, and Yu-Gang Jiang. DiffusionAD: Norm-guided one-step denoising diffusion for anomaly detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025. 2
[67] Jiangning Zhang et al. Exploring plain ViT features for multi-class unsupervised visual anomaly detection. Computer Vision and Image Understanding, 2025. 2
[68] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. DeSTSeg: Segmentation guided denoising student-teacher for anomaly detection. In CVPR, 2023. 2
[69] Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. RealNet: A feature selection network with realistic synthetic anomaly for anomaly detection. In CVPR, 2024. 2
[70] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In ICLR, 2024. 2

[71] Yixuan Zhou et al. MSFlow: Multi-scale flow-based framework for unsupervised anomaly detection. IEEE Transactions on Neural Networks and Learning Systems, 2024. 2

[72] Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In CVPR, 2024. 2

[73] Yun Zhu, Xiongjiang Cai, Shuwen Liu, Zhiyue Yu, Lianfeng Gao, and Youyun Xu. Parameter-efficient fine-tuning for citrus fruit defect segmentation model based on text prompt. Engineering Applications of Artificial Intelligence, 160:111765, 2025. 2

[74] Yun Zhu et al. Multi-class segmentation of navel orange surface defects based on improved DeepLabv3+. Journal of Agricultural Engineering, 55(2):263–275, 2024. 2

[75] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. SPot-the-difference self-supervised pre-training for anomaly detection and segmentation. In ECCV, 2022. 2

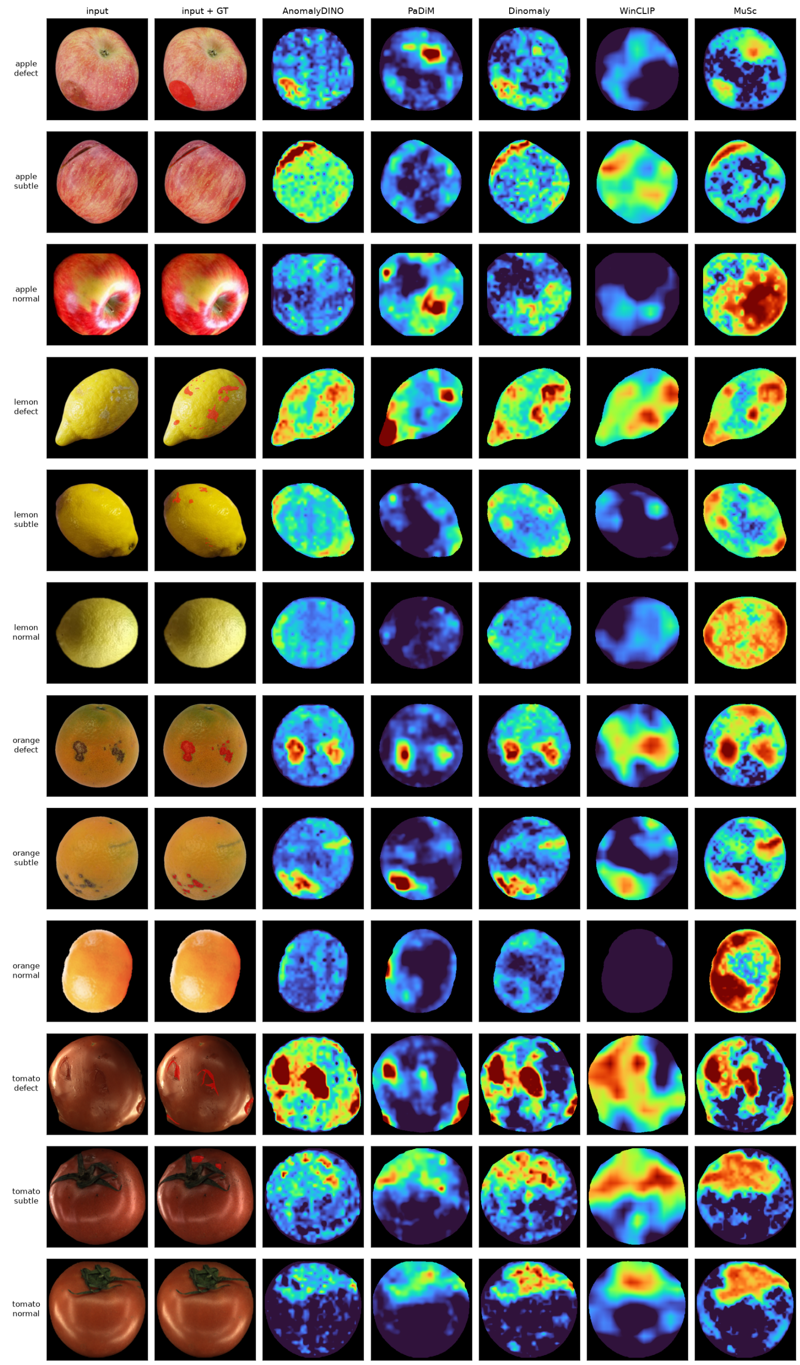

Figure 5. Per-model anomaly maps in the unified (one model, all fruits) regime, three examples per fruit: a clear defect, a subtle small defect, and a normal fruit with a prominent pedicel/calyx (the false-positive stress test). Columns: input; input with the ground-truth defect mask (red); then each method's anomaly map. AnomalyDINO, PaDiM and Dinomaly are shown from their single fruit-agnostic unified model; WinCLIP and MuSc are zero-shot/transductive and need no training. Every map uses the same fruit-ROI mask and a per-(model,fruit) global robust colour scale (cool = low, red = high). On the clear-defect rows several baselines also localise the defect (some tie or beat AnomalyDINO on the easy orange/tomato defects); the distinctive behaviour is on the normal rows, where AnomalyDINO stays comparatively quiet on healthy surface and on the pedicel/calyx, whereas the memory/CNN, reconstruction and transductive baselines spread response over the whole fruit, including normal structure.

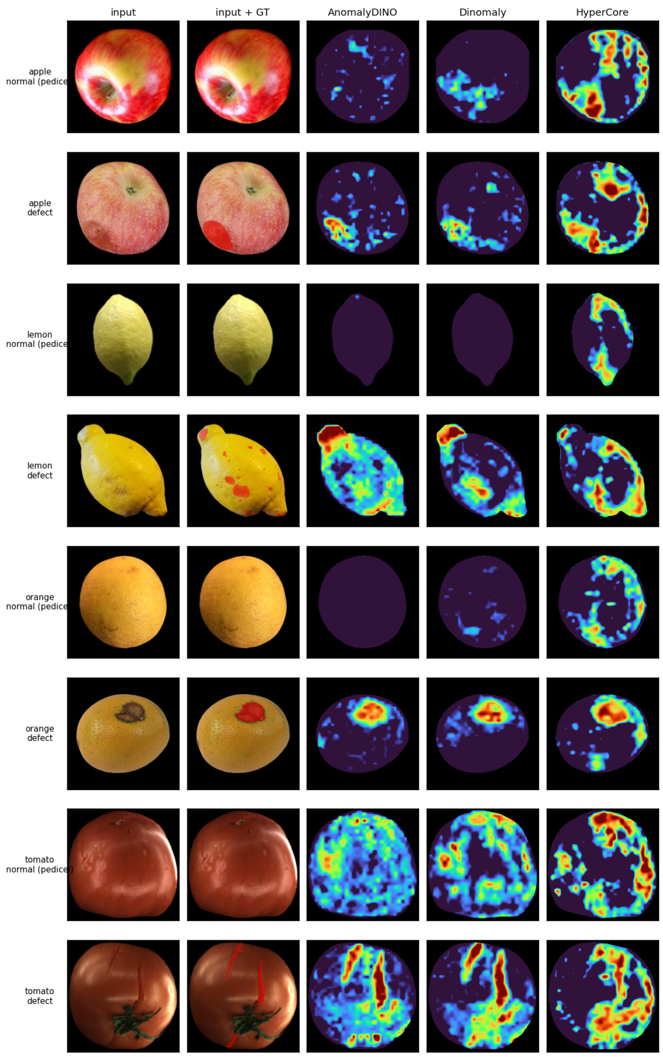

Figure 6. Pedicel/calyx response across all four fruits, comparing the training-free memory (AnomalyDINO) against a reconstruction model (Dinomaly) and our learned prototype model (HyperCore). For each fruit: a normal fruit with a prominent pedicel/stem (top) and a defect fruit (bottom). Columns: input; input+GT; AnomalyDINO; Dinomaly; HyperCore (same ROI; per-(model,fruit) global colour scale). On the normal rows AnomalyDINO stays dark on the apple/orange pedicel and on a typical lemon tip and tomato calyx while Dinomaly and especially HyperCore light them up; on tomato and lemon, however, an atypical/oddly-shaped calyx or tip is one AnomalyDINO does also respond to (consistent with the analysis in Sec. 5). On the defect rows AnomalyDINO concentrates on the defect. The learned/reconstruction models flag the pedicel as anomalous and the training-free memory fires on it less—this partial suppression, and HyperCore's diffuse over-response, mirror their detection accuracy.

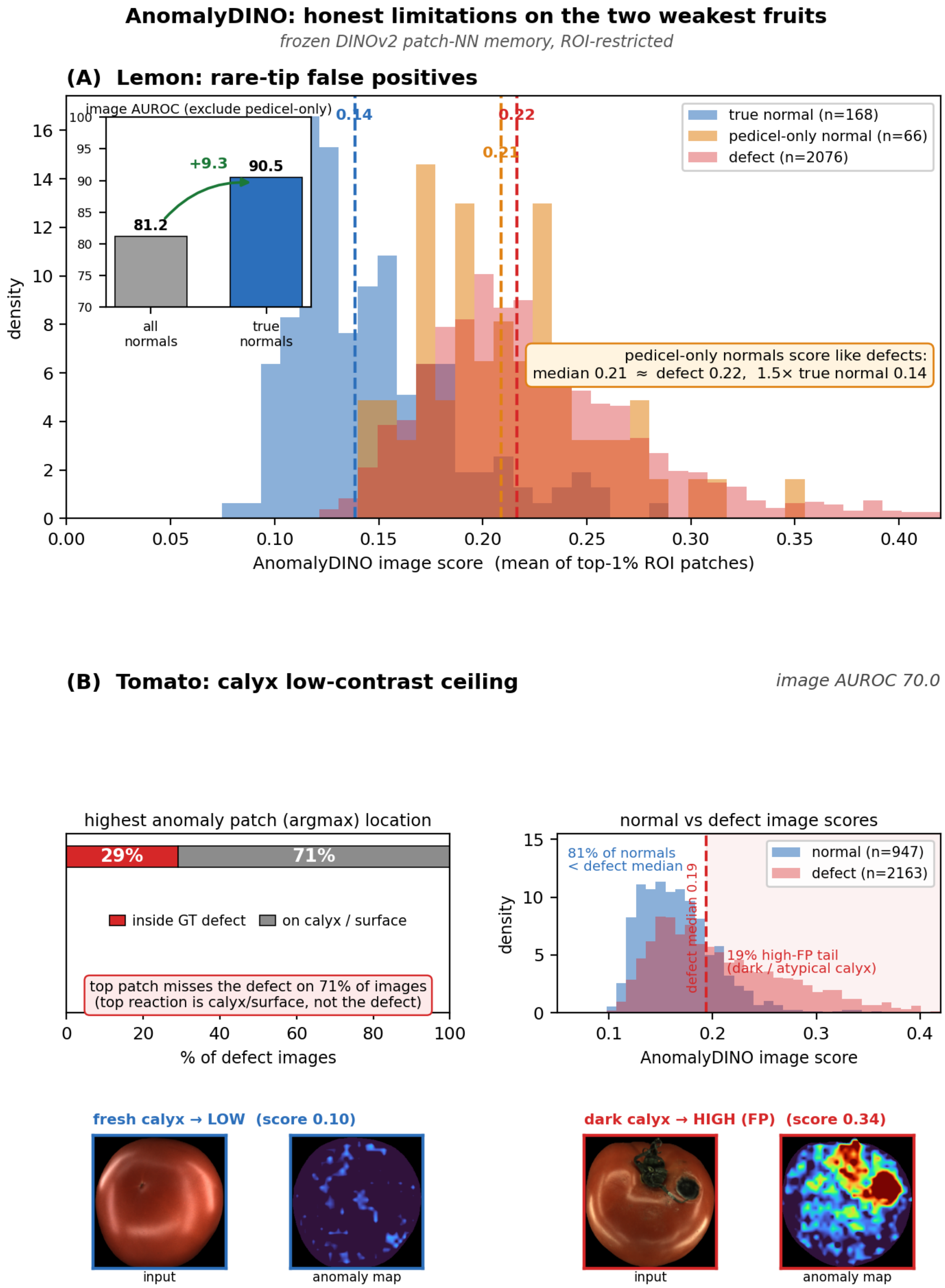

Figure 7. Failure modes: lemon rare-tip false positives (excluding pedicel-only normals lifts AUROC +9.3) and the tomato low-contrast calyx ceiling.