An Evaluation Template for Broccoli Head Segmentation

Stratification and Robustness in the Context of Deployment Use Cases

Pieter van den Haspel[1]
Supervisors: dr.ir. Cynthia Liem[1]**, dr.ir. Jeroen Wildenbeest**[2]
[1]EEMCS, Delft University of Technology, The Netherlands
[2]Inholland University of Applied Sciences, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Pieter van den Haspel
Final project course: CSE3000 Research Project
Thesis committee: dr.ir. Cynthia Liem, dr.ir. Jeroen Wildenbeest

***Abstract*—Broccoli head segmentation is the upstream step of a sizing pipeline that feeds harvest decisions, growth models, and yield forecasts, but the broccoli vision literature reports each on a different metric stack without confidence intervals, stratification, or robustness testing. This work identifies which evaluation variables actually differentiate broccoli head segmentation models under two deployment use cases: end-of-life-cycle harvest decisions and whole-cycle growth monitoring.**

Five recent broccoli segmentation models are evaluated with cluster-bootstrap confidence intervals on two publicly available broccoli image datasets, under stratification by leaf occlusion, head size, and time of day; under two controlled photometric and motion-blur perturbations calibrated against published in-deployment image-sharpness anchors; and under a noise-injection sweep on a published downstream growth-prediction model.

Stratification by occlusion and by head size each flip the ranking between two of the broccoli segmentation models tested; no single architecture is operationally optimal across the growth curve. The two controlled perturbations agree on the qualitative robustness tiers but not on the strict rank order; a cross-camera evaluation produces no usable ordering at all. Upstream sizing-label noise propagates linearly into downstream growth-prediction MAE with a small coefficient, so cleaning sizing labels beyond the most accurate currently published sizing pipeline buys at most one to four hours of biological head growth, well under one harvest-decision cycle. The implication is that broccoli vision evaluation needs deployment-context-aware reporting rather than a single headline number, and sizing-pipeline accuracy is not the operational lever that moves downstream growth-prediction performance.

I. Introduction

Modern fresh-market broccoli production relies on plant-level decisions: when each head has reached harvestable size, which plants to cut on a given pass, and how each plant's growth trajectory contributes to the field's overall yield estimate. Within a single field, broccoli heads do not mature uniformly. The maximum fraction of harvestable heads in a single cutting depends on cultivar but typically ranges between 30 and 90%, with most cultivars staying below 50% [9]: a substantial proportion of any field-wide cut either is not yet at marketable size or is past its prime. Reducing this waste requires plant-by-plant harvest decisions rather than field-wide passes. The current operational practice is for harvest workers to visually assess and individually select each head between May and November [9], which is labour-intensive and increasingly hard to staff as seasonal-labour availability declines.

A proposed alternative is automated plant-level monitoring as part of *precision agriculture*: a camera mounted on a moving platform captures each row, an algorithm locates and sizes every head, and the resulting estimates feed downstream models that predict yield, schedule harvest, and trigger field interventions. This is currently an experimental setup rather than the standard practice, but the operational case for it is established. A Japanese drone-imaging study reports that a one- to two-day deviation in predicted optimal harvest date materially affects food-loss and farmer profit [14], and the Dutch climate agreement targets 50% of agricultural land managed under precision-agriculture methods by 2030 [11].

The accuracy of every downstream decision rests on a single upstream step: getting the size of each head right. A growing body of work develops broccoli sizing pipelines using deep instance-segmentation methods, but their evaluation conventions are inconsistent and have not been tested under realistic field conditions. Of the seven core broccoli vision papers [2], [6], [8], [9], [14], [16], [17] no two report the same metric stack, none publishes confidence intervals on its headline numbers, and no paper systematically characterises model behaviour under realistic field perturbations. Beyond the choice of which metrics to report, the headline numbers themselves are coarse: papers report a single aggregate value per metric with no breakdown by occlusion, head size, time of day, lighting, or any other condition under which a model might behave differently. A reader is therefore left with two layers of opacity at once (which metrics to compare, and within a single metric whether the headline number actually applies to the deployment regime they care about) and cannot tell whether reported improvements are real, robust, or actionable.

This paper closes that gap by addressing the main research question: *what evaluation metrics and benchmarks can be established for a more nuanced understanding of the success of broccoli head segmentation techniques?* Four sub-questions decompose this: **RQ1.** What evaluation metrics are currently used across the broccoli vision literature, and what reporting structure best supports cross-paper comparison? **RQ2.** For what variables does stratification differentiate broccoli head segmentation models more clearly than aggregate metrics do? **RQ3.** What scene-level variables can be used to evaluate the robustness of broccoli head segmentation models, and how can these variables be applied systematically to existing test sets? **RQ4.** How does upstream sizing-MAE propagate into downstream growth-prediction MAE in an operational pipeline?

We address RQ1 with a structured survey of seven empirical broccoli vision papers, distilled into a MoSCoW-style priority grid parameterised by two deployment use cases. RQ2 and RQ3 are tested empirically on two publicly released broccoli image datasets: *Sexbierum* [2] and *LAR Broccoli* [6]. The dataset names refer to the cultivar and cropping-system combinations rather than to the geographic locations themselves. We evaluate five publicly available segmentation models with cluster-bootstrap confidence intervals clustered on plant identity, controlled perturbation along two scene variables (motion blur, wet-and-sunny) calibrated to in-deployment image-sharpness anchors from [6], and a cross-camera comparison on LAR's four cameras. RQ4 retrains the downstream growth-prediction MLP of [9] (training data referred to as *Verdonk* below, under the same naming convention) under a sweep of injected upstream sizing-label noise to measure how an additional millimetre of upstream sizing MAE propagates into downstream prediction MAE.

II. Background and Related Work

A. Broccoli farming and the harvest-timing problem

Broccoli (*Brassica oleracea* var. *italica*) for the Dutch fresh market is harvested manually across roughly 2,500 hectares each season, with selective per-plant harvesting needed because single-pass mechanical cuts leave much of the field unmarketable [3], [9], [13], [15]; Greenports Nederland lists innovation as a priority theme [5].

B. Existing broccoli vision pipelines

Seven empirical broccoli vision papers form the comparison set for this work. [2] train Mask R-CNN and ORCNN on a tractor-mounted RGB-D capture, reporting visible- and amodal-mask IoU and per-head diameter MAE stratified by leaf occlusion; ORCNN is explicitly trained to be occlusion-robust by predicting the head's amodal extent, while Mask R-CNN predicts only the visible part. [6] release the LAR Broccoli dataset captured with three cameras at three platform speeds and report mAP at COCO IoU thresholds; their per-camera Laplacian-variance measurement is the only quantitative anchor for in-deployment imaging sharpness in the literature. [9] uses a YOLOv8n head detector to feed a multi-layer perceptron that predicts diameter from weather features, reporting MAE under 10-fold cross-validation. [8] apply semantic segmentation for maturity identification, [17] combine segmentation with a yield-grading classifier, [14] use drone imagery with a custom mid-IoU and a profit/income model, and [16] compare drone-based segmentation models per growth stage with a yield RMSE of 1.81 cm. Detailed per-paper metric stacks appear in Appendix A.

The seven papers fall into two deployment use cases: end-of-life-cycle harvest decisions on mature heads, served by [2], [6], [8], [17]; and whole-cycle growth monitoring, served by [9], [14], [16].

C. Definitions: amodal and visible masks, MAE and IoU

Two ground-truth conventions for broccoli head segmentation exist in the literature. A *visible* mask annotates only the pixels of the head that are not occluded by leaves; an *amodal* mask annotates the full extent of the head as if no occlusion existed (Fig. 1). The convention chosen at training time determines the inference target: a visible-trained model outlines what it sees, an amodal-trained model extrapolates beyond visible boundaries [4]. Of the seven core broccoli papers, only [2] publishes both conventions side by side.

We rely on two metrics throughout the rest of the paper. The *intersection-over-union* (IoU) of a predicted mask P and a ground-truth (GT) mask G is $|P \cap G|/|P \cup G|$, a dimensionless segmentation-quality score bounded in $[0, 1]$. The *mean absolute error* (MAE) of a sizing pipeline is the mean over predicted heads of $|\hat{d}_i - d_i|$, the absolute deviation between the predicted diameter $\hat{d}_i$ and its ground-truth value d_i. MAE is reported in millimetres for upstream sizing and in centimetres for Mateijsen's downstream growth-prediction MLP, in line with each paper's units.

Fig. 1. Visible-mask versus amodal-mask annotation conventions on the same broccoli head. *Left:* visible-mask covers only pixels of the head that are not occluded by leaves. *Right:* amodal-mask covers the entire head including the leaf-covered regions.

III. Methodology

This section states one hypothesis per research sub-question and describes the empirical procedure used to test it. Statistical aggregation is described once in Section III-E because it is shared across all four protocols.

A. RQ1: literature inventory and MoSCoW design

H1. *Metric reporting across the broccoli vision literature is heterogeneous and omits stratification, confidence intervals, and robustness instrumentation, so the body of work does not currently support a single deployment-aware reporting unit. A use-case-parameterised grid based on the surveyed metrics can be assembled, but its operational relevance depends on the deployment case rather than on any single metric being universally appropriate.*

We survey the seven empirical broccoli papers in scope [2], [6], [8], [9], [14], [16], [17]. For each paper we extract every distinct metric reported on model predictions and the rationale the paper gives for choosing it, producing a list of which metrics appear and in how many of the seven papers. We then combine that list with the rationales given by the papers, with the recommendations from the broader evaluation-methodology literature (in particular [12]), and with the empirical findings from the three sequential sub-questions of this paper to build a priority list of metrics by deployment use case. The result is a MoSCoW grid in which each (metric, use case) cell is classified MUST / SHOULD / COULD according to its operational relevance for the case, with cells outside scope explicitly marked rather than reported as failures.

B. RQ2: stratified evaluation with cluster bootstrap

H2. *Stratification by realistic evaluation variables (leaf occlusion, head size, and time of day) exposes condition-dependent behaviour of broccoli head segmentation models that aggregate metrics conceal, and at least one of these axes produces a model-ranking change that is statistically distinguishable from zero under paired bootstrap.*

We evaluate three diameter-estimation pipelines on the Sexbierum dataset: Mask R-CNN and ORCNN with mask-based

sizing [2], and Mateijsen's YOLOv8n [9] with bounding-box sizing. The first two are tested on the 345-image held-out split (57 plants) [2]. This is the same split that paper publishes as a held-out test set, so the published Mask R-CNN and ORCNN checkpoints are not contaminated with these images and the absolute MAE values report generalisation rather than training error. The YOLOv8n is tested on the full 1613-image set (250 plants) [9], since it was trained on a different field and is therefore out-of-domain on Sexbierum regardless of split membership. Per-image diameter MAE is computed against the caliper-measured ground truth from [2] and stratified along three axes: leaf occlusion (10 visible-pixel-ratio bins following [2]), head size (small < 100 mm, medium 100–150 mm, large > 150 mm), and time of day (morning < 11 h, midday 11–13 h, afternoon ≥ 14 h). Mask-IoU evaluation adds Christofi's amodal-trained YOLOv26s-seg (a broccoli-trained-from-scratch segmentation model included here under direct collaboration with the author and not re-distributed by this paper, see Section VIII) and a COCO-pretrained YOLOv26s-seg baseline, each scored against the GT convention it was trained for.

C. RQ3: controlled perturbation and cross-camera characterisation

H3. *All three scene-level variables we evaluate (motion blur, lighting plus moisture, and camera) produce consistent differences in robustness between the broccoli vision models tested, and the orderings they produce agree across the three variables.*

The three scene-level variables were chosen because each maps to a documented field-deployment challenge for agricultural computer vision. *Motion blur* is the operative variable for any moving-platform broccoli capture (especially drone-based deployment, an explicitly identified future direction for plant-level monitoring [14], [16]) and is the only field-deployment variable measured quantitatively in the broccoli vision literature, via the per-camera Laplacian-variance anchors of [6]. *Wet-and-sunny* reflects a documented illumination challenge in which specular reflection from wet plant surfaces under direct sunlight degrades camera-based perception in the field [10]; the synthetic perturbation itself follows the photometric-corruption tradition of [7]'s brightness corruption with broccoli-mask-localised highlight blobs. *Camera* is the trivial in-deployment robustness variable: if a model performs consistently across hardware that differs in sensor, optics, and platform mount, that consistency is itself a useful in-deployment robustness property.

a) Controlled perturbation.: We perturb the 345-image Sexbierum test split with two perturbation families at five severity levels each, applied to the RGB channel only. *Motion blur* uses the ImageNet-C directional-kernel mechanism of [7] with the angle fixed to a row-aligned direction approximating tractor motion, and the severity-to-kernel-size mapping calibrated against the measured Laplacian-variance values of [6] for three cameras at three platform speeds. The Laplacian variance is a standard image-sharpness measure: high values for sharp images, low for blurry. [6] report 27, 260, and 416 for the Logitech, RealSense D435, and ZED-2 respectively; we match our synthetic severity ladder against these by choosing motion-blur kernel sizes on a log-spaced grid such that s_0 is unperturbed, s_5 matches the Logitech (worst measured) sharpness, and the intermediate severities s_1–s_4 interpolate between the RealSense D435 / ZED-2 anchors and the Logitech anchor on the Laplacian-variance axis.

b) Wet-and-sunny.: The wet-and-sunny perturbation is a bespoke photometric synthesis: gamma brightness elevation plus Gaussian highlight blobs placed within the broccoli-mask region, simulating wet-surface specular reflection under direct sunlight, with blob count and intensity varying across severity by a deterministic schedule. Representative severity examples for both perturbation families are shown in Appendix A, Fig. 3. Five models are evaluated on every (image, severity) combination (Mask R-CNN, ORCNN, Christofi's YOLOv26s-seg, the COCO-pretrained YOLO, and Mateijsen's YOLOv8n); detection rate is over all 345 images at each severity, and IoU and diameter MAE are over detected frames only.

c) Cross-camera characterisation.: The same five models are additionally evaluated on the LAR Broccoli dataset [6] across its four cameras (RealSense_A, RealSense_B, Logitech, ZED-2). LAR captures a different cultivar in a different field, so absolute detection and IoU values are not directly comparable to the in-distribution Sexbierum evaluation; what is well-defined is whether each model is *consistent* across the four cameras. We report this as a within-LAR consistency check (representative frames per camera in Appendix A).

D. RQ4: training-label sensitivity of the downstream MLP

H4. *Training-label sizing noise propagates approximately linearly to MLP test MAE; the slope is small enough that cleaning labels below current state-of-the-art sizing-MAE yields only marginal downstream gains.*

The downstream model is Mateijsen's growth-prediction multi-layer perceptron [9]: a $[128, 256, 128]$ architecture with ReLU activations and no dropout, Adam at $\eta = 10^{-3}$, batch size 256, 20 epochs, L_1 loss, MinMaxScaler on 15 weather features. [9] reports the MLP's growth-rate context against two reference datasets: *Verdonk*, the in-distribution training/test cultivar at 1.06 cm/day, and *Tanashi*, an external slower-growing comparison cultivar at 0.68 cm/day. We use both growth-rate references when converting RQ4 downstream MAE reductions into hours of biological head growth.

a) Noise injection.: For each target added sizing-MAE ε we draw zero-mean noise whose expected absolute value equals ε and add it to Mateijsen's training-set diameters; the injected MAE is interpreted additively on top of his inherent labels-versus-ground-truth error. The ε sweep covers 20 levels from 0 to 16 mm, dense (0.5 mm steps) below 3 mm where the response changes most and 1 mm steps above.

b) Retraining and cross-validation.: For each ε we run $B = 200$ independent noise realisations and retrain the MLP from scratch. The protocol is wrapped in stratified $K = 10$ cross-validation at the plant level, matching Mateijsen's

own evaluation procedure. A noise-free ensemble baseline of $K = 10$ unperturbed MLPs is averaged to provide a stable reference for the initialisation-noise floor. The test split is held at Mateijsen's original (unperturbed) values throughout, so test MAE versus ε measures the response of the MLP's fit to training-label noise with downstream test-label noise held constant. Test MAE is reported overall and per plant-size stratum (small < 5 cm, medium 5–10 cm, large > 10 cm) with 95% cluster bootstrap CIs at the plant level (Section III-E). The RQ4 plant-size bins differ from the RQ2 head-diameter bins (< 100 / 100–150 / > 150 mm) because the two stratify different populations: RQ2 mature-plant head diameters on Blok's test set, RQ4 whole-cycle growth states on Mateijsen's Verdonk training set, where plants spend most of their life below 5 cm.

E. Statistical aggregation: cluster bootstrap

All confidence intervals use the cluster bootstrap [1], [12] with plant identity as the unit of independence: for each metric we resample plant IDs with replacement, take all rows from sampled plants, recompute, and use the 2.5th and 97.5th percentiles as the interval limits (the standard bounds for a 95% percentile bootstrap CI [12]). Image-level resampling would treat the multiple images per plant on the Sexbierum test set as independent and shrink CIs by $\approx \sqrt{k}$ for k images per plant; the cluster bootstrap avoids this. Paired comparisons compute the per-image difference first (on rows where both models predicted) and bootstrap with the same clustering. Iterations: $B = 10{,}000$ for RQ2, $B = 5{,}000$ for RQ3, $B = 200$ realisations for RQ4 wrapped in 10-fold CV. No multiple-comparison correction is applied: the headline ranking flips (occlusion bin 0.0–0.1, the size flip on small heads) have CIs whose magnitudes are large enough to survive any standard correction, but starred cells whose CI barely excludes zero (notably the $+1.5$ and $+0.9$ mm differences at the highest visibility bins) should be read as suggestive rather than confirmed at family-wise α.

IV. Results

A. RQ1: metric heterogeneity and the MoSCoW protocol

The seven core broccoli papers report metric sets that cannot be directly compared across papers (see Appendix A). No single metric is reported by more than three of the seven papers, and no two papers report the same combination of detection, segmentation, and sizing metrics. The closest pair ([17] and [8]) agree on accuracy / precision / recall / F1 / IoU but disagree on pixel-level versus instance-level granularity. No surveyed paper reports confidence intervals on its headline numbers, although [9] reports a fold-mean $\pm$ fold-std on MLP MAE that approximates one under a Gaussian assumption.

The heterogeneity has an operational explanation. The seven papers serve at least two distinct deployment use cases, and each rewards a different metric prioritisation. We make these use cases explicit and propose a MoSCoW grid that maps each (metric, condition) cell to a priority class per case (Table VI). The use-case framework is itself the methodological contribution of RQ1: it explains the metric heterogeneity, defines a minimum reporting unit per case, and the explicit case declaration becomes the precondition for any cross-paper synthesis.

Case A, end-of-life-cycle harvest decision (served by [2]). The model is deployed at the end of the growth cycle to decide which heads to harvest on a given pass; the target population is mature heads at or near the marketable range, and the deployment context is a controllable daytime window. *Case B, whole-cycle growth monitoring* (served by [9], partially by [16]). The model is deployed across the full life cycle to track each head's size trajectory with longitudinal-consistency requirements on the growth increment, with repeated weekly passes across time of day and seasonal weather.

The full MoSCoW grid is in Appendix A; the headline priorities differ between the two cases in operationally meaningful ways. Case A's MUST cells are recall (every head needs to be considered for harvest, and a false positive is cheap because a cutting arm that approaches an empty location simply sees no head and skips) and the diameter MAE on the medium and large head bins; small-head MAE, longitudinal sizing MAE, and harvest-day deviation are COULD because the harvest-decision pipeline neither targets small heads nor tracks individuals over time. Case B inverts the precision/recall priority (a smaller set of accurately-sized heads produces a cleaner growth curve than a larger noisy set) and elevates diameter MAE across *all* size bins to MUST, since heads spend much of the growth cycle below 100 mm. Stratification by occlusion and by time of day and robustness to the two perturbations are SHOULD for both cases. The MoSCoW grid carries three overlays in prose: drone deployments upgrade motion-blur robustness to MUST because platform speeds magnify blur relative to tractor-mounted capture; on cultivars with a stable occlusion range across the population, the occlusion-stratified MAE is MUST for the cultivar's bin and downgrades to COULD elsewhere; and the priority on F1 depends on the harvest mechanism, since a per-plant detection miss is a missed economic event under selective harvesting but partly absorbed at the field-aggregate level under region-wise harvesting.

B. RQ2: stratified evaluation reveals two ranking flips

A *ranking flip* here means that the relative ordering of two models on the same metric reverses between strata of a stratification variable: for example, model A is more accurate than model B on one stratum but less accurate than B on another.

a) Occlusion: a ranking flip between visible-mask and amodal-mask sizing.: Stratifying by the visible-pixel ratio of [2] (Table I) reproduces that paper's amodal-mask finding and adds cluster-bootstrap confidence intervals [2] did not report. At the most heavily occluded bin (visible ratio 0.0–0.1), ORCNN's MAE is 28.0 mm $[12.8, 48.4]$ versus Mask R-CNN's 99.2 mm $[67.0, 123.1]$, a paired reduction of -66.8 mm $[-81.0, -56.0]$. The advantage shrinks monotonically as visibility increases, becomes statistically indistinguishable for bins

0.4–0.7, and at 0.8–1.0 the ranking reverses: Mask R-CNN beats ORCNN by approximately 1 mm with a paired CI that excludes zero. The marginal MAE summary (8.59 mm for ORCNN versus 12.38 mm for Mask R-CNN on the test set) would have obscured the fact that ORCNN's win is concentrated in a narrow regime of heavy occlusion, and that under near-perfect visibility ORCNN's amodal inference actively adds error by extrapolating occluded mass that does not exist.

b) Size: a second ranking flip in the opposite direction.: Stratifying by ground-truth diameter (full table in Appendix A) reveals a qualitatively different ranking flip. On large heads (> 150 mm) ORCNN beats Mask R-CNN by -4.9 mm $[-8.7, -1.3]$ and on medium heads (100–150 mm) by -4.4 mm $[-6.2, -2.8]$. On small heads (< 100 mm) the ranking reverses: Mask R-CNN reaches an MAE of 5.3 mm $[2.7, 7.8]$ while ORCNN sits at 13.6 mm $[7.0, 19.0]$, a $+8.2$ mm penalty $[+2.6, +11.8]$. The small-head bin contains only 20 images and 3 plants on ORCNN's held-out evaluation subset, so the magnitude of the penalty is uncertain, but the direction of the flip is robust at this confidence level.

The size flip has a concrete operational reading: a grower monitoring an immature field (heads predominantly < 100 mm) is better served by Mask R-CNN, while one scheduling harvest of a mature field (100–150 mm and $>$ 150 mm bins) prefers ORCNN, by the same amodal-versus-visible mechanism analysed in Section V.

c) Time of day: substrate-dependent on YOLOv8n, no signal on Mask R-CNN or ORCNN.: On the full Sexbierum substrate, Mateijsen's YOLOv8n shows a midday-best signature with disjoint 95% CIs: 23.8 mm $[20.6, 26.9]$ at midday versus 30.6 mm $[26.7, 34.9]$ in the morning and 31.4 mm $[28.1, 34.9]$ in the afternoon (Table X in the appendix). On the matched 345-image substrate the CIs overlap heavily and the midday advantage disappears; Mask R-CNN and ORCNN show no separation on their substrate either. The separated YOLOv8n signature is therefore partly a property of the full dataset's time-of-day distribution rather than an architectural fact. The operational reading is conditional: midday scheduling may help on a full-Sexbierum deployment with a cross-domain detector, but it is not a universal property of the model.

C. RQ3: controlled perturbation reveals robustness tiers; cross-camera does not

a) Motion blur differentiates models clearly.: Table II shows the per-severity behaviour under directional motion blur calibrated to Hardy's Laplacian variances. *ORCNN is the standout*: $100 \rightarrow 95.4\%$ detection from clean to s_5, amodal IoU $0.84 \rightarrow 0.81$, diameter MAE constant at 8.6 mm. *Mask R-CNN* shows a graceful detection drop ($100 \rightarrow 71.3\%$) but its mask quality on detected frames is preserved (visible IoU $0.89 \rightarrow 0.88$). *Mateijsen's YOLOv8n* detection collapses catastrophically ($96.2 \rightarrow 11.9\%$). The lightest model is the least blur-robust by a wide margin.

Christofi and the *COCO baseline* display the same selectivity-threshold artefact in opposite magnitudes: Christofi detection drops mildly ($94.5 \rightarrow 88.1\%$) and its amodal IoU on the kept frames actually rises ($0.65 \rightarrow 0.71$), while the COCO baseline detection collapses by more than half ($77.4 \rightarrow 31.9\%$) and visible IoU on the kept frames also rises ($0.82 \rightarrow 0.84$). In both models the IoU rise has the same cause: the confidence-threshold drops the harder, more occluded frames first and leaves an easier residual subset, so the IoU averaged over *detected* frames goes up even though the model's actual ability to handle blur has degraded. The single-number metric that catches this by construction is mAP at IoU thresholds [0.5:0.95], which couples detection and mask quality and therefore tracks the underlying degradation. Only [6] among the seven broccoli papers reports this metric, and we recommend it as the headline robustness number when only one summary statistic is reported.

b) Wet-and-sunny differentiates models more gently and the ordering only partially agrees with motion blur.: Table III shows the wet-and-sunny sweep. Mask R-CNN and ORCNN are unaffected (100% detection at every severity, IoU stable to within the CI), Christofi is stable on both axes, Mateijsen drops mildly ($96.2 \rightarrow 90.1\%$), and the COCO baseline shows a $\sim$10 percentage point detection drop ($77.4 \rightarrow 67.0\%$) and the same selectivity pattern.

The two perturbations agree on the qualitative tiers (the Blok detectron2 family and Christofi cluster at the top of both, the COCO baseline and Mateijsen sit at the bottom of both) but they do not preserve a strict rank order. Mateijsen is the worst-by-far model under motion blur (detection collapses by 84 pp) and outperforms the COCO baseline under wet-and-sunny (6 pp drop versus 10 pp), and Mask R-CNN drops considerably under blur (29 pp) but is unaffected by wet-and-sunny. The disagreement is plausibly because motion blur disrupts the mid-frequency texture the lighter YOLO family relies on, whereas wet-and-sunny mainly shifts brightness statistics that the un-fine-tuned COCO baseline handles worst. The practical reading is that a single perturbation variable is not enough to characterise robustness on this set of models, so both should be reported.

c) Cross-camera on Hardy LAR yields no robustness ordering.: Table IV shows per-camera marginal detection rate and IoU. None of the five models was trained on Hardy's cultivar, so the cross-camera evaluation is purely a check of *whether each model is consistent across the four cameras* (amodal-model IoU on LAR is cross-target; see Table IV caption). The five models all fluctuate substantially between cameras, no model is meaningfully more camera-robust than the others, and the partial differences that do appear do not align with the controlled-perturbation ordering above. We therefore do not draw any model-robustness conclusion from the cross-camera data, and recommend controlled perturbation as the robustness instrument for future broccoli vision evaluations.

TABLE I

DIAMETER MAE (MM) STRATIFIED BY LEAF OCCLUSION (EVERY OTHER VISIBLE-RATIO BIN SHOWN FOR COMPACTNESS; FULL TABLE IN APPENDIX A). 95% CLUSTER BOOTSTRAP CIS CLUSTERED ON PLANT IDENTITY; ORCNN − MRCNN IS THE PAIRED DIFFERENCE, ENTRIES MARKED * HAVE A CI EXCLUDING ZERO.

Visible ratio	n	**MRCNN MAE**	**ORCNN MAE**	**ORCNN − MRCNN**
0.0–0.1	4	99.2 [67.0, 123.1]	28.0 [12.8, 48.4]	−66.8 [−81.0, −56.0]*
0.2–0.3	36	22.9 [17.7, 29.1]	12.1 [8.6, 16.0]	−10.6 [−17.9, −4.5]*
0.4–0.5	31	9.5 [6.5, 12.3]	6.6 [4.9, 8.6]	−3.1 [−6.4, +0.6]
0.6–0.7	30	7.8 [5.5, 10.6]	7.1 [5.4, 9.3]	−0.7 [−3.6, +2.3]
0.8–0.9	25	3.7 [2.6, 5.1]	5.3 [4.0, 6.6]	+1.5 [+0.4, +2.8]*

TABLE II

MOTION-BLUR SWEEP ON BLOK'S 345-IMAGE TEST SPLIT. SEVERITY IS CALIBRATED TO HARDY'S MEASURED LAPLACIAN VARIANCES. DETECTION RATE OVER ALL 345 IMAGES; MASK IOU OVER DETECTED FRAMES ONLY WITH 95% CLUSTER BOOTSTRAP CI CLUSTERED ON PLANT IDENTITY.

Model	s_0	s_1	s_2	s_3	s_4	s_5
Detection rate (%)						
Mask R-CNN	100	99.7	98.0	95.1	83.5	71.3
ORCNN	100	100	100	98.8	97.4	95.4
Christofi	94.5	94.2	91.9	88.7	89.3	88.1
Baseline	77.4	73.9	65.2	56.2	40.9	31.9
Mateijsen	96.2	56.5	23.2	17.1	15.1	11.9
Diameter MAE (mm), MRCNN + ORCNN only						
Mask R-CNN	12.4	12.4	12.8	13.8	12.9	10.3
ORCNN	8.6	8.8	9.0	9.0	8.8	8.6

TABLE III

WET-AND-SUNNY SWEEP ON BLOK'S 345-IMAGE TEST SPLIT. COLUMNS AND PROCEDURE AS IN TABLE II. THE QUALITATIVE TOP/BOTTOM TIERS AGREE WITH THE BLUR SWEEP BUT THE STRICT RANK ORDER DOES NOT (MATEIJSEN AND THE COCO BASELINE SWAP).

Model	s_0	s_1	s_2	s_3	s_4	s_5
Detection rate (%)						
Mask R-CNN	100	100	100	100	100	100
ORCNN	100	100	100	100	100	100
Christofi	94.5	95.7	97.1	96.2	95.9	97.4
Baseline	77.4	76.5	74.2	71.6	68.1	67.0
Mateijsen	96.2	95.7	94.8	94.5	92.8	90.1

D. RQ4: the downstream MLP is largely insensitive to sizing-label noise

Fig. 2 shows the response of Mateijsen's MLP test MAE to injected training-label sizing noise. In the operational range $\varepsilon \geq 2$ mm the response is linear with slope 0.0089 cm of test MAE per mm of label MAE (95% CI $[0.0070, 0.0104]$, $R^2 = 0.996$). Below $\varepsilon \approx 2$ mm the response plateaus at the initialisation-noise floor ≈ 0.61 cm.

a) Stratified slopes: small heads dominate the sensitivity.: Per-size-bin slopes (cm of test MAE per mm of label MAE): small heads (< 5 cm), 0.0122 $[0.0103, 0.0143]$; medium (5–10 cm), 0.0085 $[0.0070, 0.0102]$; large (> 10 cm), 0.0063 $[0.0041, 0.0097]$. The label-noise effect is roughly twice as large for small heads as for large ones; combined with their lower baseline MAE ($\sim$0.29 cm vs $\sim$0.91 cm for large), the same absolute MAE reduction is a proportionally larger gain for the small-head sub-problem.

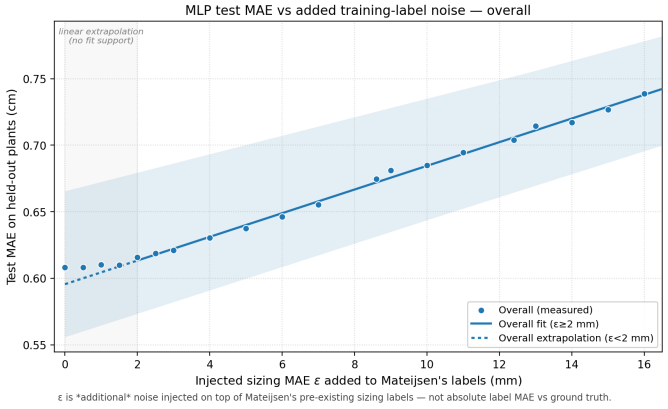

Fig. 2. MLP test MAE on Mateijsen's held-out plants versus injected sizing-MAE ε added to his training labels. Solid line: per-fold linear fit over $\varepsilon \geq 2$ mm. Dotted line: extrapolation. The x-axis is *additional* noise on top of Mateijsen's existing labels, not absolute label-MAE versus ground truth. Per-size-bin slopes are reported in the prose below.

b) An upper bound on what perfect sizing labels could buy.: ORCNN's sizing MAE on Blok's Sexbierum test set is 8.6 mm against caliper-measured ground truth [2], measured on a Sexbierum-trained ORCNN evaluated on a Sexbierum test split. The counterfactual we ask is what happens if ORCNN replaces Mateijsen's bounding-box-from-detection sizing pipeline on the Verdonk dataset that Mateijsen's MLP is trained on. This requires an explicit assumption: that an ORCNN *trained on Verdonk* would reach a similar $\sim$ 8.6 mm MAE there. Verdonk and Sexbierum are different cultivars and different fields, so this is a working assumption, not a measured fact, but [2]'s 8.6 mm value is the best calibrated reference point in the literature for what an occlusion-aware mask-based sizing pipeline can deliver.

Applying our slope, improving the upstream sizing pipeline from this 8.6 mm baseline to a hypothetical 0 mm MAE would reduce downstream MLP MAE by 0.077 cm overall (95% CI $[0.060, 0.089]$), and by 0.054 to 0.105 cm across plant-size strata (Table V). This 0.077 cm figure is an upper bound. The four plateau points below $\varepsilon \approx 2$ mm in Fig. 2 suggest the response flattens at the initialisation-noise floor, so the linear extrapolation overshoots, and an absolute 0 mm

TABLE IV

HARDY LAR CROSS-CAMERA MARGINAL DETECTION RATE AND VISIBLE-MASK IOU WITH 95% CLUSTER BOOTSTRAP CI CLUSTERED ON PER-CAMERA TRACK ID. MATEIJSEN'S YOLOV8N IS BOUNDING-BOX ONLY AND PRODUCES NO MASK, SO ITS IOU COLUMN IS REPORTED AS "—" (NOT APPLICABLE); ITS DETECTION-RATE VALUES ARE FULLY COMPARABLE TO THE OTHER MODELS. ORCNN/CHRISTOFI IOU VALUES ARE REPORTED BUT ARE CROSS-TARGET ON THIS DATASET (HARDY HAS ONLY VISIBLE GT) AND ARE MECHANICALLY BOUNDED BY THE VISIBLE-PIXEL RATIO.

Camera	**Mask R-CNN**	**ORCNN**	**Christofi**	**Baseline**	**Mateijsen**
Detection rate					
Logitech	1.4%	0.8%	19.3%	10.8%	11.1%
RealSense_A	12.2%	12.2%	42.6%	14.8%	41.7%
RealSense_B	18.3%	19.2%	45.0%	20.5%	46.7%
ZED-2	83.1%	81.8%	69.7%	29.9%	63.4%
Mean IoU on detected frames, vs visible GT [95% CI]					
Logitech	0.21 [0.00, 0.47]	0.43 [0.12, 0.75]	0.47 [0.41, 0.52]	0.55 [0.45, 0.65]	—
RealSense_A	0.44 [0.15, 0.64]	0.36 [0.06, 0.57]	0.48 [0.34, 0.60]	0.57 [0.30, 0.80]	—
RealSense_B	0.64 [0.55, 0.71]	0.57 [0.49, 0.64]	0.54 [0.49, 0.58]	0.66 [0.59, 0.71]	—
ZED-2	0.17 [0.14, 0.20]	0.18 [0.15, 0.21]	0.13 [0.11, 0.16]	0.20 [0.16, 0.25]	—

TABLE V

PREDICTED DOWNSTREAM MLP MAE REDUCTION (CM) IF THE SIZING PIPELINE WERE IMPROVED FROM ORCNN'S MEASURED 8.6 MM MAE TO A HYPOTHETICAL 0 MM, AND THE EQUIVALENT TIME-OF-GROWTH UNDER EACH OF MATEIJSEN'S TWO GROWTH-RATE REFERENCES.

Stratum	**ΔMAE (cm)**	**Verdonk (h)**	**Tanashi (h)**
Overall	0.077	1.7	2.7
Small (<5 cm)	0.105	2.4	3.7
Medium (5–10 cm)	0.073	1.7	2.6
Large (>10 cm)	0.054	1.2	1.9

MAE is operationally unrealistic anyway. More importantly, the 8.6 mm baseline is measured on a test set of fully grown broccoli where ORCNN sizes most accurately per the RQ2 size stratification, and on that big-head-dominated regime the relevant stratified slope is the large-head 0.0063 cm/mm rather than the population-averaged 0.0089 cm/mm, so the actual downstream reduction is closer to 0.054 cm (large row of Table V). One further caveat: the injected ε is i.i.d. per-diameter, whereas real sizing error is plausibly within-plant correlated and heteroscedastic with size and occlusion, both of which propagate differently through a per-plant time-series MLP than matched-marginal i.i.d. noise. Anchored in Mateijsen's growth-rate framing of 1.06 cm/day on Verdonk and 0.68 cm/day on the Tanashi external dataset [9, §4.5], these upper-bound reductions convert to between one and four hours of biological head growth, against a baseline MLP error of 13 to 21 hours and a harvest decision cadence of approximately two days. The maximum credible operational gain from sizing-pipeline cleanup is therefore on the order of a few percent of one harvest cycle.

V. DISCUSSION

a) One mechanism explains both RQ2 ranking flips.: The occlusion ranking flip and the size ranking flip in RQ2 look like two unrelated regime changes but trace back to a single underlying cause that ties them together: the amodal-versus-visible distinction from [4] maps directly onto the conditions under which ORCNN beats Mask R-CNN and the conditions under which it does not. At low visibility there is genuine occluded mass to extrapolate and ORCNN's amodal head is doing the work it was trained for. At high visibility, and at small head sizes that are typically more compact and less occluded than mature heads in the same field, there is no occluded mass to predict and the amodal extension actively manufactures area that does not exist. The practical implication is that a cross-paper comparison between an amodal-trained model and a visible-trained one is determined as much by the visibility-and-size distribution of the evaluation population as by the models themselves, a confound that the broccoli vision literature has not previously been explicit about.

b) The RQ4 slope reaches the random-noise component, not the systematic one.: Our 0.0089 cm/mm slope was measured by adding zero-mean noise on top of Mateijsen's existing labels, so the propagation it characterises is the propagation of *random* label error, not of systematic bias. Per [9, §5.2] the current bounding-box-from-depth-at-centre sizing pipeline systematically under-reads true diameter; the operational case for switching to segmentation-based sizing therefore rests on removing that systematic component, which our slope cannot quantify. The one-to-four-hour upper bound applies specifically to the random-noise channel and is consistent with Mateijsen's recommendation in [9, §8.1] only when read in that direction. Separating the two components requires caliper-measured ground-truth diameters on a Verdonk subset, listed under Future Work.

c) What transfers beyond broccoli, and what does not.: The transferability the abstract and conclusion claim is deliberately scoped to the methodological apparatus rather than the specific MoSCoW cell contents. Use-case-parameterised reporting, cluster bootstrap on the natural unit of independence, and severity-calibrated controlled perturbation are all directly portable to other smart-farming computer-vision pipelines and to agricultural sizing and yield-estimation problems more generally: the same skeleton (a two-or-more use case decomposition, a per-case metric grid, cluster-bootstrap CIs on the relevant identity unit, and at least two perturbation axes

calibrated to deployment-anchored sharpness or photometric measurements) would apply directly to apple harvest detection or to lettuce maturity estimation by re-deriving the per-crop use cases and the per-crop grid. The grid in Appendix A is itself broccoli-specific in its metric inventory and in the two deployment cases it parameterises, and would need to be re-derived per crop and per pipeline.

d) Reflection on the experimental design.: Three things the present design could not reach are worth acknowledging. The RQ4 sweep retrains the MLP on noisy sizing labels but does not retrain the upstream segmentation model itself, so the propagation it measures is downstream-only; a more complete pipeline experiment would also perturb the segmentation model. The wet-and-sunny perturbation is internally consistent but not externally calibrated, so its severity values are comparable across the models in this paper but not directly to other work without an equivalent of Hardy's Laplacian-variance anchor. And the cross-camera comparison on Hardy LAR turned out to be uninformative not because camera invariance is unimportant but because all five models were out-of-distribution on Hardy's cultivar; a follow-up using a model fine-tuned on LAR Broccoli would isolate camera behaviour from cultivar-domain mismatch.

VI. Conclusion

Single-number, single-substrate evaluation hides regime-dependent behaviour in any deep-learning pipeline whose deployment regime varies meaningfully; the broccoli vision literature is a particularly clear instance. Stratification by leaf occlusion and by head size each flips the aggregate model ranking, so no single architecture is operationally optimal across the broccoli growth curve and the choice between visible-mask and amodal-mask sizing depends on the deployment regime. Two controlled perturbation variables agree on the qualitative robustness tiers but not on the strict rank order, so a single perturbation axis is insufficient to characterise robustness; a cross-camera evaluation on a cultivar-mismatched dataset produces no usable ordering at all. Upstream sizing-label noise propagates downstream only weakly: cleaning sizing labels below the most accurate currently published sizing pipeline buys at most one to four hours of biological head growth, well under one harvest-decision cycle, so the operational lever for downstream growth-prediction performance sits elsewhere in the pipeline. Together, stratified cluster-bootstrap CIs, controlled perturbation as the primary robustness instrument, and MoSCoW use-case-aware reporting provide a transferable evaluation template for the broader smart-farming computer-vision literature.

VII. Future Work

Four lines of future work follow directly from this paper's findings. A spectrally measured calibration of the wet-and-sunny per-severity blob intensities on field-wet broccoli would turn that perturbation's severity ladder from an internally-consistent ranking into a calibrated benchmark comparable to the Laplacian-variance anchoring of [6] we use for motion blur. Closer to a real benchmark, roughly 50 heads imaged under each of four conditions (wet, dry, sunny, overcast) at a fixed mounting height would let the synthetic wet-and-sunny perturbation be validated against real paired imagery and released as a public dataset. Looking further ahead, diffusion-based scene synthesis would extend the perturbation inventory to visually richer conditions without hand-engineering each new variable, provided a validation procedure verifies predicted-diameter stability against an unperturbed reference to preserve ground truth. Finally, hand-measured GT diameters for a small Verdonk subset would convert RQ4's upper-bound slope argument into a directly measured labels-versus-ground-truth gap, closing the systematic-component question our slope analysis cannot reach [9, §5.2].

VIII. Responsible Research

a) Open data and reproducibility.: All experimental code, perturbation generation pipelines, statistical aggregation routines, and figure-generation scripts are released as a single repository alongside this paper[1], together with the exact parameter values used (perturbation severity-to-kernel mappings, MoSCoW grid definitions, ε grid for RQ4, bootstrap iteration counts, fixed random seeds). The perturbed Blok images we generate are derivative works of Blok's publicly released test split and are distributed under the same terms; releasing them increases reproducibility because the perturbed images form a known reference set that future authors can score their models against without re-implementing the perturbation pipeline.

b) Use of publicly released models and datasets.: The Mask R-CNN, ORCNN, and Mateijsen YOLOv8n model weights are sourced from publicly available checkpoints released by their respective authors, as is the COCO-pretrained YOLOv26s-seg baseline. Christofi's YOLOv26s-seg model is not publicly released; the results reported here use it under direct collaboration with the author and the weights are not re-distributed by this paper. The private Verdonk dataset that underlies [9] was not used in raw form; we work only with the publicly released derived pickles that accompany Mateijsen's thesis. The other primary datasets (the Sexbierum split from [2], LAR Broccoli from [6]) are publicly released by their respective institutions, so no private data was at risk of being leaked or re-distributed in the course of this work.

c) Labour displacement from agricultural automation.: The operational success of this research direction implies the displacement of seasonal harvest labour in the Dutch fresh-market broccoli sector. The aggregate labour-availability framing used by [9] and the Dutch climate agreement does not fully address the individual workers whose livelihoods depend on this work. We flag this honestly rather than treating the labour shortage as a complete justification for automation, and we recommend that future agricultural automation work include transition planning (reskilling support, alternative employment in adjacent sectors) as a substantive part of its responsible-research section.

[1] Available at https://github.com/pieterhaspelino/CSE-3000-broccoli-eval-repo

d) LLM use disclosure.: Large language models were used as a writing aid in preparation of this paper, and for code review and refactoring suggestions on the experimental scripts. All numerical claims in the tables and prose were independently verified against the source CSV files that came out of the experiments. A session-level log of LLM use is maintained in the project repository.

REFERENCES

[1] Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5:1089–1105, 2004.

[2] P. M. Blok, E. J. van Henten, F. K. van Evert, and G. Kootstra. Image-based size estimation of broccoli heads under varying degrees of occlusion. *Biosystems Engineering*, 208:213–233, 2021.

[3] Pieter M. Blok, Ruud Barth, and Wim van den Berg. Machine vision for a selective broccoli harvesting robot. In *IFAC-PapersOnLine*, volume 49, pages 66–71, October 2016.

[4] Patrick Follmann, Rebecca König, Philipp Härtinger, Michael Klostermann, and Tobias Böttger. Learning to see the invisible: End-to-end trainable amodal instance segmentation. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[5] Greenports Nederland. Nationale tuinbouwagenda 2019–2030: Circular horticulture in practice – the priorities of the horticulture cluster united in greenports nederland. Technical report, Greenports Nederland, June 2019.

[6] O. Hardy, K. Seemakurthy, and E. I. Sklar. A comparative dataset of annotated broccoli heads recorded with depth cameras from a moving vehicle. *Agronomy*, 14(5):964, 2024.

[7] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

[8] S. Kang, D. Li, B. Li, J. Zhu, S. Long, and J. Wang. Maturity identification and category determination method of broccoli based on semantic segmentation models. *Computers and Electronics in Agriculture*, 217:108633, 2024.

[9] N. Mateijsen. Field-based predictive growth modeling of broccoli (brassica oleracea var. italica) within precision agriculture. Master's thesis, Delft University of Technology, 2025.

[10] Akshay Patel, Senthil Vinayagamoorthy, Hyun Park, Sahil Misra, Jeongki Heo, and Hua Wang. Hydra: Accurate multi-modal leaf wetness sensing with mm-wave and camera fusion. *arXiv preprint arXiv:2508.02409*, 2025.

[11] Rijksoverheid Nederland. Klimaatakkoord. Technical report, Rijksoverheid Nederland, June 2019.

[12] Gael Varoquaux and Olivier Colliot. Evaluating machine learning models and their diagnostic value. In Olivier Colliot, editor, *Machine Learning for Brain Disorders*, volume 197 of *Neuromethods*, pages 601–630. Humana (Springer), 2023.

[13] L. R. Walton and J. H. Casada. Evaluation of broccoli varieties for mechanical harvesting. *Applied Engineering in Agriculture*, 4(1):5–7, 1988.

[14] Haozhou Wang, Tang Li, Erika Nishida, Yoichiro Kato, Yuya Fukano, and Wei Guo. Drone-based harvest data prediction can reduce on-farm food loss and improve farmer income. *Plant Phenomics*, 5:Article 0086, 2023.

[15] World Agritech. Roboveg unveils autonomous broccoli harvester. Industry news, October 2021.

[16] Chenzi Zhang, Xiaoxue Sun, Shuxin Xuan, Jun Zhang, Dongfang Zhang, Xiangyang Yuan, Xiaofei Fan, and Xuesong Suo. Monitoring of broccoli flower head development in fields using drone imagery and deep learning methods. *Agronomy*, 14(11):2496, 2024.

[17] C. Zhou, J. Hu, Z. Xu, J. Yue, H. Ye, and G. Yang. A monitoring system for the segmentation and grading of broccoli head based on deep learning and neural networks. *Frontiers in Plant Science*, 11:402, 2020.

APPENDIX

TABLE VI

MoSCoW prioritisation of (metric, condition) cells for the two broccoli vision deployment use cases. The Varoquaux column marks cells whose presence is directly supported by the recommendations of [12] for ML evaluation. Three deployment-context overlays modify these defaults: drone deployments upgrade both robustness rows to MUST; stable-occlusion cultivars make the occlusion-stratified MAE MUST only for the relevant bin and downgrade the remaining bins to COULD; and the priority on F1 depends on the harvest mechanism. See Section IV-A for the prose summary.

Metric	**Condition**	**Case A**	**Case B**	**Varoquaux**
Recall	in-distribution	MUST	could	—
Precision	in-distribution	SHOULD	MUST	—
Diameter MAE	small bin (< 100 mm)	could	MUST	yes
Diameter MAE	medium + large bins	MUST	MUST	yes
Diameter MAE	stratified by leaf occlusion	SHOULD	SHOULD	yes
Diameter MAE	stratified by time of day	SHOULD	SHOULD	yes
Diameter RMSE	in-distribution	SHOULD	SHOULD	—
Harvest-day deviation	—	SHOULD	MUST	—
Longitudinal sizing MAE	per-head growth increment	n/a	MUST	—
Robustness to motion blur	calibrated severity	SHOULD	SHOULD	—
Robustness to wet-and-sunny	calibrated severity	SHOULD	SHOULD	—

Metric	**Task**	**Papers (of 7)**
Precision	Detection	3
Recall	Detection	3
F1	Detection	3
Accuracy	Detection	1
mAP @ IoU=0.5	Detection	2
mAP @ IoU=[0.5:0.95]	Detection	1
Visible-mask IoU	Segmentation	2
Amodal-mask IoU	Segmentation	1
Mid-IoU (custom)	Segmentation	1
Pixel accuracy	Segmentation	1
Mean pixel accuracy	Segmentation	2
Mean IoU	Segmentation	2
MAE	Sizing	2
MAD	Sizing	1
Median absolute error	Sizing	1
RMSE	Sizing/yield	3
R^2 / r^2	Sizing/yield	3
Cumulative error curve	Sizing	1
Pearson correlation	Sizing	1

TABLE VII

Source data for the heterogeneity finding in Section IV-A: no metric is reported by more than three of the seven papers, and no two papers report the same combination of detection, segmentation, and sizing metrics.

TABLE VIII

Full version of Table I (all ten visible-ratio bins).

Visible ratio	n	**MRCNN MAE**	**ORCNN MAE**	**ORCNN − MRCNN**
0.0–0.1	4	99.2 [67.0, 123.1]	28.0 [12.8, 48.4]	-66.8 $[-81.0, -56.0]^*$
0.1–0.2	19	46.3 [36.6, 57.1]	21.2 [14.9, 29.0]	-22.8 $[-30.5, -15.2]^*$
0.2–0.3	36	22.9 [17.7, 29.1]	12.1 [8.6, 16.0]	-10.6 $[-17.9, -4.5]^*$
0.3–0.4	53	16.4 [12.7, 20.9]	9.8 [6.6, 13.2]	-6.6 $[-12.0, -1.5]^*$
0.4–0.5	31	9.5 [6.5, 12.3]	6.6 [4.9, 8.6]	-3.1 $[-6.4, +0.6]$
0.5–0.6	41	8.8 [6.9, 10.8]	7.6 [5.1, 10.9]	-1.2 $[-4.8, +2.8]$
0.6–0.7	30	7.8 [5.5, 10.6]	7.1 [5.4, 9.3]	-0.7 $[-3.6, +2.3]$
0.7–0.8	21	5.3 [3.1, 7.9]	6.6 [4.8, 8.5]	$+1.1$ $[-2.2, +4.3]$
0.8–0.9	25	3.7 [2.6, 5.1]	5.3 [4.0, 6.6]	$+1.5$ $[+0.4, +2.8]^*$
0.9–1.0	85	4.9 [4.0, 5.8]	5.8 [4.9, 6.8]	$+0.9$ $[+0.2, +1.6]^*$

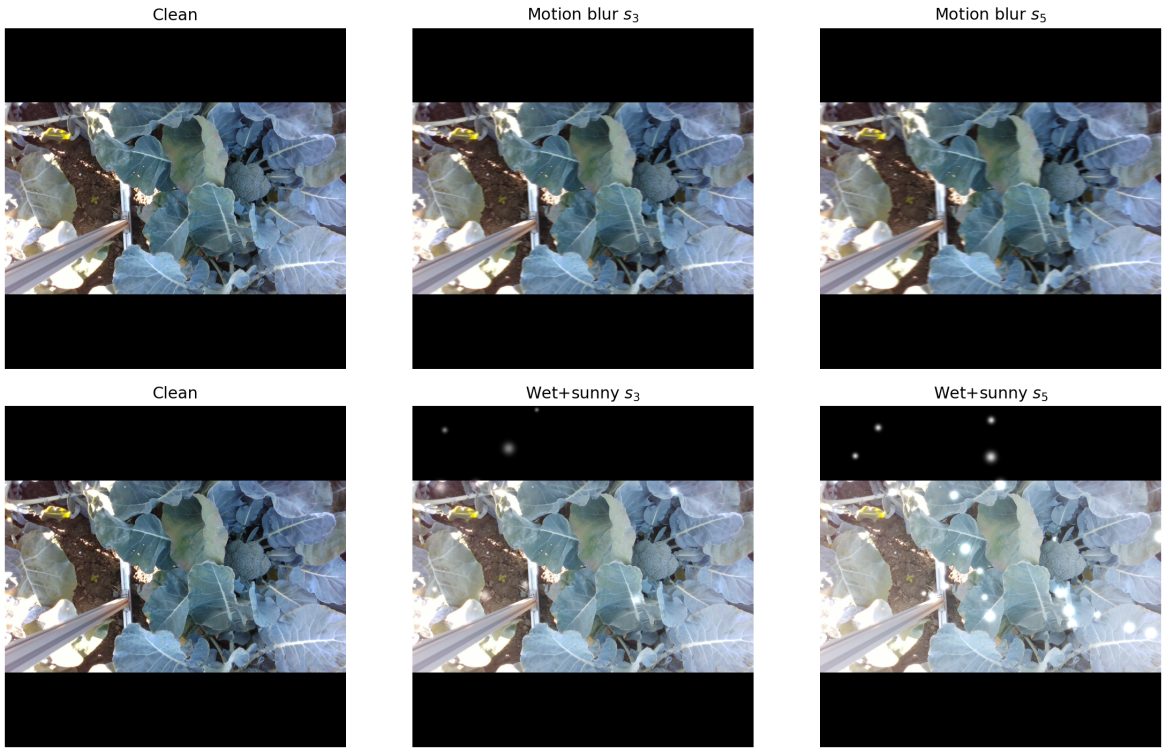

Fig. 3. Representative severity examples of the two perturbation families: motion blur (top) and wet-and-sunny (bottom), at severities s_0 (clean), s_3, s_5. All perturbations are applied to the RGB channel only and preserve segmentation ground-truth geometry by construction.

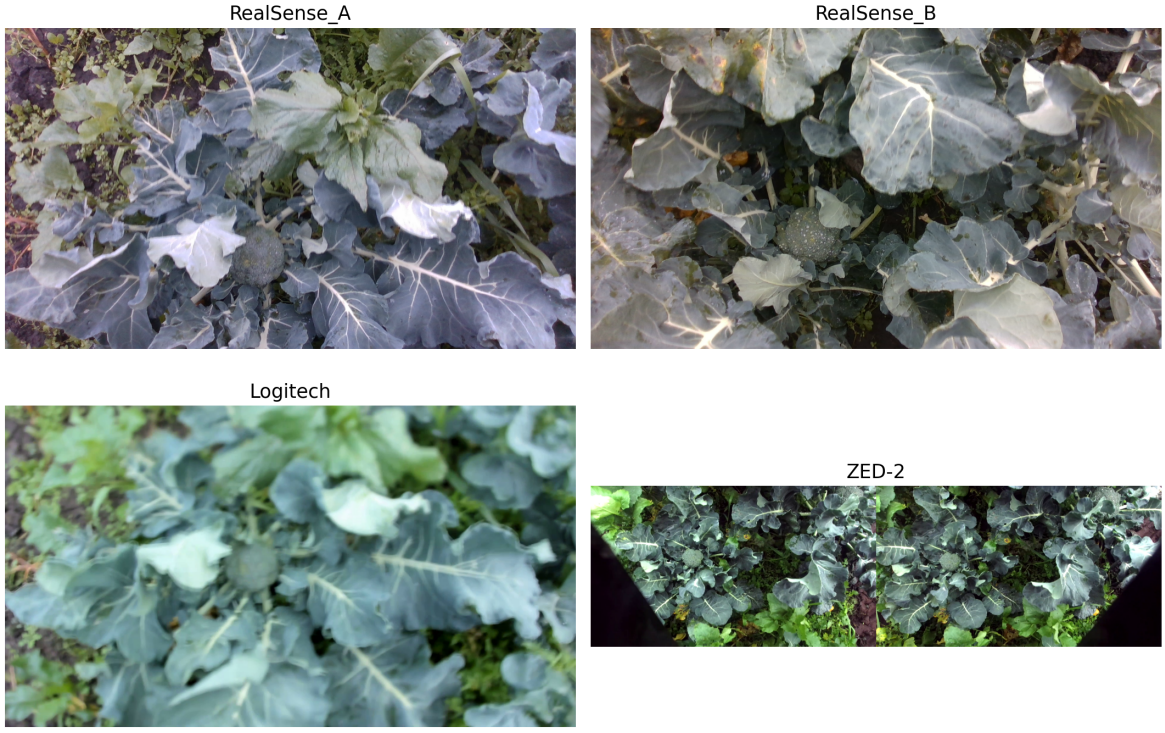

Fig. 4. Representative frames from each of Hardy LAR's four cameras: RealSense_A, RealSense_B, Logitech, ZED-2. The four captures differ visibly in field of view, distance to the heads, image resolution, and natural blur intensity. These differences are not controllable across cameras and are confounded with model behaviour in any per-camera comparison, which is why the cross-camera component of Section IV-C is framed as a within-LAR consistency check rather than as a cross-camera robustness test.

TABLE IX

DIAMETER MAE (MM) STRATIFIED BY GROUND-TRUTH HEAD DIAMETER. THE MASK R-CNN ADVANTAGE ON SMALL HEADS IS THE SECOND RANKING FLIP DISCUSSED IN SECTION IV-B.

Diameter bin	n	**MRCNN MAE**	**ORCNN MAE**	**ORCNN − MRCNN**
Small ($<$ 100 mm)	20	5.3 [2.7, 7.8]	13.6 [7.0, 19.0]	$+8.2$ $[+2.6, +11.8]^*$
Medium (100–150 mm)	222	11.9 [9.9, 14.1]	7.6 [6.5, 8.7]	-4.4 $[-6.2, -2.8]^*$
Large ($>$ 150 mm)	103	14.7 [10.9, 19.5]	9.8 [7.2, 13.2]	-4.9 $[-8.7, -1.3]^*$

TABLE X

DIAMETER MAE (MM) STRATIFIED BY TIME OF DAY WITH 95% CLUSTER BOOTSTRAP CIS CLUSTERED ON PLANT IDENTITY. MASK R-CNN AND ORCNN ARE EVALUATED ON BLOK'S 345-IMAGE HELD-OUT TEST SPLIT. MATEIJSEN'S YOLOV8N IS REPORTED TWICE: ON THE FULL 1,613-IMAGE SEXBIERUM SET (ITS NATIVE EVALUATION SUBSTRATE) AND RESTRICTED TO THE 345-IMAGE ORCNN SUBSTRATE TO MAKE THE COMPARISON SUBSTRATE-MATCHED. YOLOV8N SHOWS A MIDDAY-BEST SIGNATURE WITH DISJOINT CIS ON THE FULL SET BUT NO SEPARATION ON THE MATCHED 345 SUBSTRATE. MRCNN AND ORCNN SHOW NO SEPARATION ON THEIR SUBSTRATE EITHER.

Time of day	**MRCNN (345)**	**ORCNN (345)**	**YOLOv8n (full 1613)**	**YOLOv8n (matched 345)**
Morning ($<$ 11h)	11.6 [7.1, 17.5]	8.2 [5.8, 11.4]	30.6 [26.7, 34.9]	24.1 [18.3, 28.4]
Midday (11–13h)	14.2 [10.2, 19.4]	9.6 [7.1, 12.4]	23.8 [20.6, 26.9]	26.4 [20.7, 31.0]
Afternoon ($\geq$ 14h)	12.0 [9.8, 14.3]	8.4 [6.8, 10.1]	31.4 [28.1, 34.9]	27.5 [21.6, 33.7]