foods

MDPI

Article

Cultivar Shift in YOLO-Based Potato Quality Grading

Marko Simonič *, Jernej Hernavs, Simon Klančnik and Tadej Peršak

Faculty of Mechanical Engineering, University of Maribor, Smetanova 17, SI-2000 Maribor, Slovenia; jernej.hernavs@um.si (J.H.); simon.klancnik@um.si (S.K.); tadej.persak@um.si (T.P.)
* Correspondence: marko.simonic@um.si

Abstract

An optical potato grading line may process several cultivars. We examined how much accuracy a detector trained on one cultivar lost on another, whether a jointly trained model matched cultivar-specific models and whether compact models matched larger ones. Twenty-five You Only Look Once (YOLO) variants (five families, five sizes each) were trained on Belmonda, Prada or both (75 runs, one seed) to detect tubers and grade them as edible, feed or rotten. Models were evaluated on held-out images from the same acquisition sessions (95 per cultivar). Relative to target-domain specialists, the mean shortfall in mean average precision (mAP50–95) was 0.336 from Belmonda to Prada and 0.162 in the reverse direction; it was larger from Belmonda to Prada for all 25 variants. Grade confusions, mostly involving the feed class, accounted for 96–97% of false negatives, whereas missed tubers were rare. Joint models differed from the specialists by −0.014 to +0.017 mAP50–95 on each cultivar; small joint models met a 0.01 margin against larger same-family specialists in 17 of 30 comparisons. For these two cultivar domains, a jointly trained model performed similarly to same-architecture specialists; as cultivar was confounded with lot and season, validation on independent lots is required.

Keywords: potato quality grading; object detection; YOLO; cultivar shift; domain shift; joint training; model capacity; computer vision

1. Introduction

Potatoes (*Solanum tuberosum* L.) are among the most important staple crops worldwide [1]. Post-harvest handling, transport and storage can cause quantitative and qualitative losses [2,3]. In many small and medium-sized operations, tubers are still graded by manual visual inspection, which is labor-intensive and subjective [4,5]. Grading determines whether a tuber is sold for consumption, used as animal feed or discarded, so systematic grading errors translate directly into economic losses or reduced quality of the marketed product.

Industrial optical sorters combine spectral, near-infrared or X-ray sensing with mechanical ejection and achieve high throughput [6,7], but their cost limits their use on smaller farms [5,7]. Machine vision for automatic produce grading has a long history [8,9], and deep learning has made red–green–blue (RGB) image inspection with commodity hardware increasingly practical [10,11]. In manufacturing, deep models are also combined with simulation-based digital twins for quality prediction and control [12]. Single-stage detectors from the You Only Look Once (YOLO) family [13,14] can localize and classify many objects in one pass and are widely used for industrial visual inspection, including within the parameter and computational limits of embedded deployment [15]. Current YOLO releases provide each architecture in several sizes, from nano to extra-large, that trade accuracy against computation [14,16]. However, such models are usually compared on a

Check for updates

Academic Editor: Luca Laghi

Received: 16 October 2026
Revised: 13 November 2026
Accepted: 17 November 2026
Published: 18 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

single domain, so it remains unclear whether additional capacity also improves robustness to a change of cultivar. Compact detectors running on embedded accelerators have, for example, enabled real-time fruit detection in orchards [17], which makes them attractive for low-cost grading equipment. For potatoes, YOLO-based detectors have been applied to external defect detection and edibility assessment [18,19], and related deep-learning methods to defect segmentation and disease classification [20–22]. Image color and texture descriptors have also been used to recognize potato cultivars together with storage damage [23], indicating that the visual appearance on which a grading model relies differs between cultivars. Within the OPTIFARM project, a low-cost imaging platform and a three-class grading scheme (edible, feed and rotten) were developed and evaluated with instance segmentation [24] and with a benchmark of 25 YOLO configurations [25].

The performance of models trained on images from one acquisition setting can degrade substantially when the data distribution changes [26,27]. In plant disease recognition, a classifier that was highly accurate on held-out images of its own dataset reached only 31.4% accuracy on images collected under different conditions [28]. Large multi-site datasets have been assembled because differences in genotypes and observation conditions limit detectors calibrated on narrow datasets [29]. Cross-domain degradation has also been documented for agricultural detection tasks [30,31], and transferring knowledge between related product variants has improved industrial quality inspection [32].

Approaches to this problem differ in the target-domain information they assume is available. Unsupervised domain adaptation uses unlabeled target images [33,34], domain generalization aims to generalize to unseen domains without target data [35], and supervised multi-domain training pools labeled data from all domains of interest, so that a single detector can serve several domains [36]. In a large-scale comparison of image classifiers, most interventions intended to improve robustness provided little protection against natural distribution shifts, whereas training on larger and more diverse data was the main exception [37]. For a grading line that processes several cultivars, this raises two practical questions: whether larger detectors transfer better between cultivars, and whether one model trained on labeled images of all cultivars can replace cultivar-specific models. Separate models for each cultivar would also require separate training, validation and maintenance, whereas one joint model would simplify deployment on a grading line.

In our previous study, models trained at one farm and tested at another maintained high detection rates for edible tubers. Their F1 score, however, decreased from at least 0.906 on the internal test set to 0.792–0.918 at the external location, mainly because of edible–feed confusion [25]. That design evaluated a single transfer direction, did not separate a change of cultivar from other differences between the locations and did not examine training on pooled domains.

This study extends [25] by adding the Prada cultivar domain, measuring transfer in both directions against target-domain specialists, evaluating supervised joint training and examining whether small models match the accuracy of larger models of the same family; it introduces no new detector architecture. Three training regimes (Belmonda only, Prada only and joint) were crossed with 25 YOLO variants, and every model was evaluated on the same frozen test sets.

For a grading line that processes more than one cultivar, three research questions (RQs) follow; these were specified before testing:

- RQ1: How much accuracy is lost when a model trained on one cultivar is applied to the other, and which quality classes drive it?
- RQ2: Does a single model trained on both cultivars perform as well as the cultivar specialists?

- RQ3: Do compact models reach the accuracy of larger models of the same family, and does this depend on joint training?

In addition, false detections on empty roller-bed scenes were recorded as a practical check.

2. Materials and Methods

2.1. Image Acquisition and Quality-Class Annotation

All images were acquired with the imaging chamber described in [25]. Tubers rested on parallel aluminum rollers (diameter 60 mm) inside an enclosure of aluminum profiles and panels that shielded the scene from ambient light. They were imaged from above by an industrial RGB camera (Basler acA2500-14uc; Basler AG, Ahrensburg, Germany) with a 6 mm lens (Basler C125-0618-5M, F1.8) mounted 954 mm above the rollers. A 60 × 60 cm light-emitting diode (LED) panel (MAXXO ZB1224 URG19, 36 W, 4000 K, 4320 lm; EMOS spol. s r.o., Přerov, Czech Republic) above the camera provided homogeneous illumination. Images were recorded with fixed acquisition parameters at 2520 × 1835 pixels and stored as Portable Network Graphics (PNG) files.

For each image, an operator placed a group of tubers of mixed quality on the rollers, captured and immediately annotated the scene in the acquisition application. The application stored the class and bounding box of every visible tuber in YOLO format (class identifiers 0, edible; 1, feed; 2, rotten) [25]. Between consecutive images, the tubers were turned, rearranged or replaced, and no two image files were identical. The same tubers could nevertheless appear in several images in different positions or orientations; tuber identities were not recorded. The classes were defined as in [25]: edible, a marketable tuber without major visible defects; feed, a non-marketable tuber with visible external defects, discoloration, surface damage or reduced visual quality that did not amount to rot; and rotten, a tuber with clearly visible advanced decay. Borderline tubers were assigned to a class according to their dominant visible condition. The class names denote operational visual grades and do not certify suitability for human or animal consumption or the absence of internal, chemical or microbiological hazards.

Belmonda images were recorded between 4 December 2025 and 13 March 2026 and correspond to the Kranj dataset [25]; Prada images were recorded on 1, 6 and 28 July 2026. Both cultivars were obtained from the same farm in Kranj, Slovenia. Imaging of each cultivar began approximately one month after harvest. Lot identifiers and storage conditions were not recorded. All images were annotated by the same reference labeler using the same visual-grade definitions. Fifty-nine images of the empty roller bed were recorded on 17 April 2026. Here, cultivar domain denotes a cultivar together with its recorded acquisition conditions. Because the two cultivars were harvested and imaged at different times of the year and came from unrecorded lots, cultivar effects cannot be separated from season-, lot- and storage-related differences; the cross-cultivar results of this study therefore describe transfer between two cultivar domains rather than an effect of the cultivar alone.

An automated audit of all 2005 source images confirmed that each image had a label file with valid class identifiers, normalized coordinates and box extents, and that all Secure Hash Algorithm 256-bit (SHA-256) hashes were distinct (no file-identical duplicates). The source images contained 26,236 annotations (16,724 edible, 5428 feed and 4084 rotten), of which 25,391 remained in the 1949 images used after the exclusion described in Section 2.2. Because tubers were imaged repeatedly, annotations do not represent independent tubers. Representative images are shown in Figure 1.

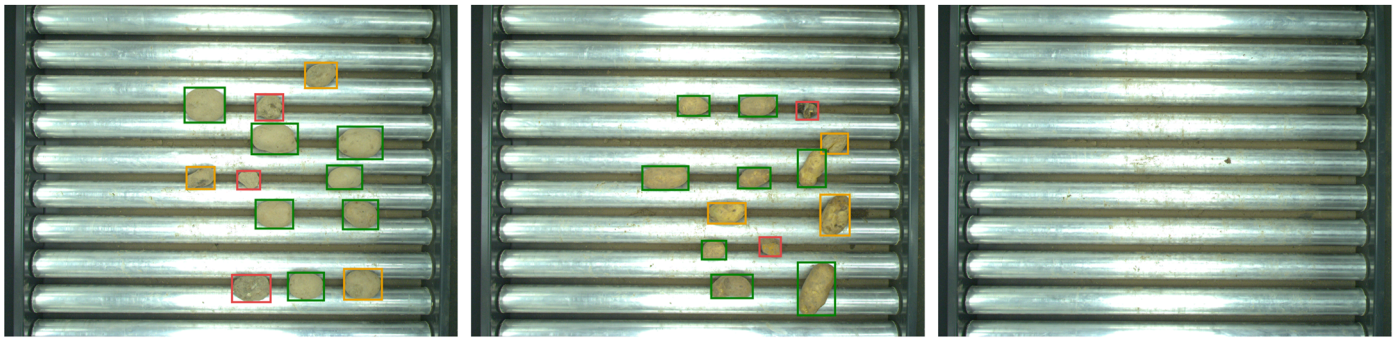


Figure 1. Representative annotated images: (**a**) Belmonda; (**b**) Prada; (**c**) empty roller bed. Bounding boxes denote the annotated visual grade: edible (green), feed (yellow) and rotten (red).

2.2. Frozen Dataset Splits and Negative Images

The Belmonda collection contained 1001 non-empty images and the Prada collection 945. To match the image counts of the two domains, 56 Belmonda images were removed before splitting. Instead of removing images at random, a precautionary annotation-similarity criterion was used so that the removed images were those whose annotated scenes most closely resembled another Belmonda image. The images were not removed because they were identical, as all source images were distinct (Section 2.1). The criterion used neither model outputs nor previous split assignments.

For each of the 500,500 pairs of non-empty Belmonda images, boxes of the same class were matched one-to-one in order of decreasing intersection over union (IoU), accepting matches with IoU ≥ 0.8. The similarity of images a and b was defined as follows:

$$S(a, b) = 2\,\Sigma_k\,\mathrm{IoU}_k/(n_a + n_b), \quad (1)$$

where the sum runs over the matched box pairs and n_a and n_b are the numbers of boxes in the two images. Pairs were processed in order of decreasing similarity. When both images of a pair were still retained, the image with the lexicographically larger file name was removed, until 56 images had been removed (similarity to the most similar retained image 0.871–0.957). No removed image had annotations identical to those of a retained image (maximum similarity 0.957). The procedure therefore matched the image counts while preferentially removing scenes with similar same-class annotation geometry. The excluded and retained Belmonda images contained on average 15.1 and 15.3 annotations per image, respectively, with edible/feed/rotten proportions of 61.1/22.0/16.9% and 62.0/21.1/16.8%; the exclusion removed 845 of the 15,310 Belmonda annotations (5.5%) and, as 32 of the 56 removed images were recorded on 13 March 2026, slightly changed the composition of acquisition days. These summaries describe annotation density and class composition; the selection may still have altered scene diversity or difficulty, and its effect on model performance was not tested.

The 945 non-empty images of each cultivar and the 59 empty images were then randomly split at the image level into training, validation and test subsets (approximately 80/10/10; Mulberry32 generator with a Fisher–Yates shuffle, seed 42). This gave 756/94/95 non-empty images per cultivar and 47/6/6 empty images (Table 1).

The domains differed in annotation density and class composition: Belmonda training images contained on average 15.3 annotations and Prada training images 11.6, with edible/feed/rotten proportions of 61.7/21.3/17.0% and 66.2/19.9/13.9%, respectively, and similar proportions in the test sets.

Table 1. Frozen dataset composition. Annotation counts are given as edible/feed/rotten (total).

Subset	Belmonda Images	Belmonda Annotations	Prada Images	Prada Annotations	Empty Images
Training	756	7113/2457/1963 (11,533)	756	5785/1744/1216 (8745)	47
Validation	94	961/280/236 (1477)	94	716/225/144 (1085)	6
Test	95	900/320/235 (1455)	95	733/216/147 (1096)	6 [1]
Removed before splitting	56	–	0	–	0

[1] Separate negative test set.

2.3. Experimental Design and Training Regimes

Three training regimes were crossed with 25 YOLO variants, giving 75 training runs, each with seed 42 (Figure 2):

- Belmonda specialist (B): 756 Belmonda images and 47 empty images (803 entries);
- Prada specialist (P): 756 Prada images and the same 47 empty images (803 entries);
- Joint (J): 756 Belmonda and 756 Prada images and the 47 empty images listed twice (1606 entries, 1559 unique images).

Listing the empty images twice in the joint regime kept the proportion of empty entries equal in all regimes (5.85%); the data loader was verified to retain the duplicates. Validation sets were domain-matched: 94 Belmonda and 6 empty images for B, 94 Prada and the same 6 empty images for P, and all 188 non-empty validation images with the 6 empty images for J. Every checkpoint was evaluated on the same three test sets (95 Belmonda, 95 Prada and 6 empty images), giving 225 test evaluations. The joint regime differed from the specialists in data diversity, data quantity and, with the same epoch limit, number of optimizer updates (about twice as many on average); joint–specialist differences reflect the combination of these factors.

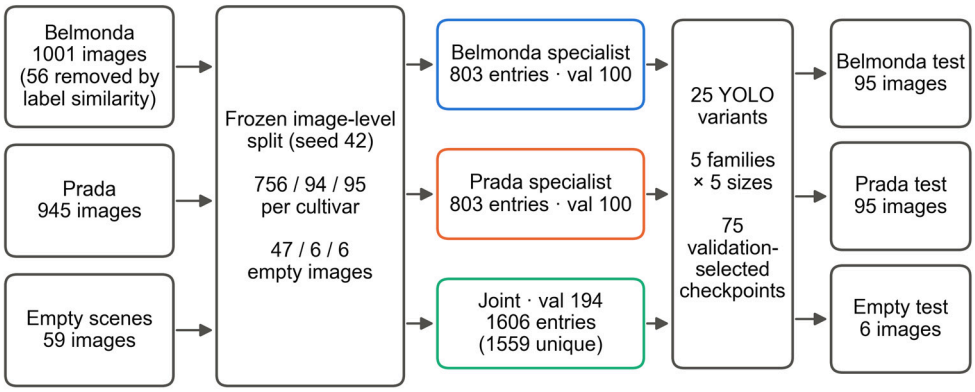


Figure 2. Experimental workflow: frozen splits, three training regimes, 25 YOLO variants (75 checkpoints selected on domain-matched validation sets) and evaluation of every checkpoint on all three test sets. Counts in the frozen-split box are training/validation/test image counts; val denotes the number of validation images.

2.4. YOLO Variants and Model Capacity

Five detector families were included: YOLOv8 [16], YOLOv9 [38], YOLOv10 [39], YOLO11 [16,40] and YOLO26 [41,42], each in five sizes. The official n/s/m/l/x variants were used for YOLOv8, YOLOv10, YOLO11 and YOLO26; for YOLOv9, the t, s, m, c and e variants served as the five size levels (Table 2). Because size letters do not denote equal capacity across families, capacity was characterized directly. All models were initialized from the official weights pretrained on Common Objects in Context (COCO) [43] distributed with the Ultralytics library [16]. YOLOv8, YOLOv9 and YOLO11 use non-maximum suppression

(NMS); YOLOv10 and YOLO26 were evaluated through their end-to-end, NMS-free inference path. Parameter counts and floating-point operation counts in billions (GFLOPs) were computed for the fused inference graph (without the auxiliary one-to-many heads of the end-to-end models) in 32-bit floating-point precision (FP32) for an input of 1 × 3 × 1280 × 1280 using thop, with MACs denoting multiply–accumulate operations (GFLOPs = 2 × MACs). These describe computational size and are not proxies for latency.

Table 2. YOLO variants and model capacity. Parameter counts are in millions (M).

Variant	Size Level	Parameters (M)	GFLOPs
YOLOv8n	N	3.01	32.3
YOLOv8s	S	11.13	113.8
YOLOv8m	M	25.84	314.8
YOLOv8l	L	43.61	659.3
YOLOv8x	X	68.13	1029.6
YOLOv9t	N	1.97	30.4
YOLOv9s	S	7.17	106.9
YOLOv9m	M	20.02	306.1
YOLOv9c	L	25.32	409.3
YOLOv9e	X	57.38	756.6
YOLOv10n	N	2.27	26.1
YOLOv10s	S	7.22	85.7
YOLOv10m	M	15.31	235.4
YOLOv10l	L	24.31	480.1
YOLOv10x	X	29.40	639.9
YOLO11n	N	2.58	25.3
YOLO11s	S	9.41	85.2
YOLO11m	M	20.03	270.6
YOLO11l	L	25.28	346.3
YOLO11x	X	56.83	777.7
YOLO26n	N	2.38	20.8
YOLO26s	S	9.47	82.1
YOLO26m	M	20.35	271.4
YOLO26l	L	24.75	344.4
YOLO26x	X	55.64	773.5

2.5. Training and Domain-Matched Validation

All variants used the same user-configurable training settings where applicable (Table 3), without model-specific tuning; architecture-specific losses and end-to-end heads were retained. Rectangular batches at a nominal image size of 1280 pixels gave a training tensor of 8 × 3 × 960 × 1280. Because the geometric augmentation stage of the installed library version produced square outputs even with rectangular batching, a project-level adapter set its output to the data-loader shape without changing any augmentation value. The 960 × 1280 training tensor dimensions were verified for every run. Augmentation was limited to hue and saturation jitter, horizontal flipping and translation, without mosaic or other image-composition augmentations.

The checkpoint with the highest validation mAP50–95 was retained. For the joint regime, this metric was computed on the pooled validation images rather than as an equal-weight mean of the two domain values. Test images were not used for checkpoint selection, early stopping, hyperparameter choices or split decisions. Before testing, every retained checkpoint was re-validated, and its mAP50–95 agreed with the value logged during training for all 75 runs (maximum absolute difference 0.0008). Runs interrupted by a hardware fault or an implementation check were repeated from scratch with unchanged data, settings and selection rule. All runs stopped early, after 49–141 epochs (Appendix A).

Table 3. Training settings shared by all 75 runs.

Setting	Value
Initial weights	COCO-pretrained weights of each variant
Input/batching	imgsz 1280, rectangular batches (training tensor 960 × 1280)
Batch size	8 (nominal batch size 64 by gradient accumulation)
Optimizer	AdamW; initial learning rate 0.001; final factor 0.01 (linear decay); momentum (β_1) 0.937; weight decay 0.0005
Warm-up	3 epochs (momentum 0.8, bias learning rate 0.1)
Epochs/early stopping	maximum 150; patience 30 epochs on validation mean average precision (mAP50–95)
Loss gains	box 7.5; classification 0.5; distribution focal loss 1.5
Color augmentation	hue 0.02; saturation 0.30; value 0
Geometric augmentation	horizontal flip $p = 0.80$; translation 0.10; all others disabled
Mosaic/mixup/copy-paste	disabled
Precision/seed	automatic mixed precision; seed 42; deterministic mode
HPC	Maister high-performance computing (HPC) system (University of Maribor, Maribor, Slovenia)
GPU configuration	Four NVIDIA Tesla V100-PCIE-32GB graphics processing units (GPUs; NVIDIA Corporation, Santa Clara, CA, USA) per node; one GPU allocated to each run.
CPU and memory	Two 6-core Intel central processing units (CPUs; Intel Corporation, Santa Clara, CA, USA), providing 24 logical CPUs, and approximately 256 GB of random-access memory (RAM) per node; six logical CPUs and 48 GB RAM allocated to each run.
Software	Python 3.12.3; Ultralytics 8.4.75; PyTorch 2.7.1; torchvision 0.22.1 (CUDA 12.6)
Operating system	Linux kernel 4.18.0-553.42.1.el8_10 (Rocky Linux 8.10 on the login node)

2.6. Test Evaluation and Detection Metrics

Each checkpoint was evaluated on the Belmonda, Prada and empty test sets with image size 1280, rectangular batches and batch size 16. Inference used 16-bit floating-point precision (FP16) on a Tesla V100 GPU, no test-time augmentation, at most 300 detections per image and, for the NMS-based families, class-aware NMS with an IoU threshold of 0.7. The primary metric was mAP50–95. For each class and IoU threshold, the average precision (AP) summarizes the interpolated precision–recall curve obtained by ranking the predictions by confidence; a prediction is a true positive only if it is matched one-to-one to an annotation of the same class with an IoU at or above the threshold, whereas duplicate and unmatched predictions are false positives. mAP50 averages the AP at an IoU threshold of 0.50 over the three classes, whereas mAP50–95 additionally averages over ten IoU thresholds from 0.50 to 0.95 in steps of 0.05. Because higher thresholds require tighter localization, mAP50–95 reflects localization accuracy as well as grade assignment, and each class receives equal weight regardless of its number of annotations [43,44]. AP was computed with the Ultralytics validator [16] at a confidence threshold of 0.001 to retain low-confidence predictions for the precision–recall analysis; mAP50 and per-class AP50–95 were secondary metrics. These metrics describe individual detections and do not directly measure the reliability of final tuber-level grading decisions (Section 4.2).

Operating-point metrics were computed at a fixed confidence threshold of 0.25. Predictions and annotations were matched one-to-one at IoU ≥ 0.50, without regard to class, by processing candidate pairs in order of decreasing IoU, then decreasing confidence, then prediction and annotation index. A matched pair with the correct class was a true positive

(TP). A matched pair with the wrong class counted as a false positive (FP) of the predicted class and a false negative (FN) of the true class. Unmatched predictions were FPs, and unmatched annotations were FNs. Precision, recall and the F1 score (the harmonic mean of precision and recall) were computed from these counts per class and as micro-averages over the three classes, and the same matching produced the confusion matrices.

2.7. Transfer, Joint-Training and Capacity Contrasts

Let $A(r, d, m)$ denote the mAP50–95 of model m trained in regime $r \in \{\mathrm{B}, \mathrm{P}, \mathrm{J}\}$ and tested on domain $d \in \{\mathrm{B}, \mathrm{P}\}$. All contrasts are absolute differences on the 0–1 scale between models evaluated on the same test images and are summarized in Table 4. The transfer shortfall (RQ1) is the deficit of the transferred model relative to the target-domain specialist of the same architecture:

$$\begin{aligned} L(\mathrm{B} \to \mathrm{P}, m) &= A(\mathrm{P}, \mathrm{P}, m) - A(\mathrm{B}, \mathrm{P}, m), \\ L(\mathrm{P} \to \mathrm{B}, m) &= A(\mathrm{B}, \mathrm{B}, m) - A(\mathrm{P}, \mathrm{B}, m), \end{aligned} \tag{2}$$

the directional asymmetry (RQ1) is:

$$D(m) = L(\mathrm{B} \to \mathrm{P}, m) - L(\mathrm{P} \to \mathrm{B}, m), \tag{3}$$

and the joint-training gain (RQ2) is:

$$G(d, m) = A(\mathrm{J}, d, m) - A(d, d, m), d \in \{\mathrm{B}, \mathrm{P}\}. \tag{4}$$

For RQ3, a small joint model s was compared with a larger in-domain specialist ℓ of the same family:

$$C(d, s, \ell) = A(\mathrm{J}, d, s) - A(d, d, \ell). \tag{5}$$

All within-family pairs N→{M, L, X} and S→{M, L, X} were pre-specified (30 pairs, 60 domain comparisons); in every pair the smaller model had fewer parameters and GFLOPs. A pair met the criterion on a domain if $C \geq -0.01$, i.e., if the small model was no more than 0.01 below the larger specialist. It met the criterion on both domains only if it met it on each. The criterion was evaluated before rounding. It is a pre-specified descriptive margin for a single training realization, not a test of equivalence or non-inferiority. It corresponds to one hundredth of the mAP50–95 scale and is small compared with the transfer shortfalls observed in this study (all at least 0.127). No operational acceptance threshold was established in this study, so the margin was chosen as a small descriptive tolerance; it was also smaller than the subsequently observed spread among the variants within each in-domain condition (0.022–0.038). Post hoc linear regressions across the variants within each in-domain condition associated a difference of 0.01 in mAP50–95 with differences of approximately 0.003–0.014 in micro-F1 at the operating point; these associations do not establish an acceptable grading error rate. As an additional baseline defined after testing, the same pairs were evaluated with the small model trained on the target cultivar alone, $A(d, d, s) - A(d, d, \ell)$. To identify the quality classes that drive these differences (RQ1), L and G were recomputed with per-class AP50–95 and complemented by confusion matrices. Spearman rank correlations between capacity and the contrasts describe the fixed set of tested variants.

2.8. Negative Test, Descriptive Analyses and Scope of Inference

As an additional practical check on the six empty test images, the number of detections per image and the number of images with at least one detection were recorded at a confidence threshold of 0.25. AP and recall are undefined on this set. To describe performance in relation to temporal proximity to the training images, each test image was assigned to

a stratum according to the time gap between its capture and that of the nearest training image of the same cultivar (≤5 s, 6–10 s and >10 s); the same target-domain strata were used for all models, including transferred models not trained on that cultivar, and micro-F1 was recomputed within each stratum from the summed TP, FP and FN counts at the operating point. To characterize visual differences between the cultivar domains, we used the central 60% of each bounding box's width and height across all retained subsets. Pixels were treated as standard RGB (sRGB) and converted to the International Commission on Illumination L*a*b* (CIELAB) color space with a D65 white point. Mean L*, a* and b*, within-crop lightness standard deviation (SD(L*)) and full bounding-box area were computed for each annotation and averaged within each cultivar and grade. One annotation of about 1 × 5 pixels yielded an empty crop and was excluded from these descriptors. Both analyses are descriptive and were defined after testing. Each configuration was trained once, so run-to-run training variability was not estimated, and differences of approximately 0.01 between individual models are interpreted descriptively. No confidence intervals are reported because the test images are not independent sampling units.

Table 4. Contrasts used to answer the research questions. All contrasts are differences in mAP50–95 between models evaluated on the same test images.

Contrast	Definition	Meaning	RQ
$L(\mathrm{B} \to \mathrm{P}, m)$, $L(\mathrm{P} \to \mathrm{B}, m)$	Equation (2)	Shortfall of a model transferred to the other cultivar relative to the target-domain specialist of the same architecture; positive values mean that the transferred model is worse	RQ1
$D(m)$	Equation (3)	Difference between the two transfer directions; positive values mean a larger shortfall from Belmonda to Prada	RQ1
$G(d, m)$	Equation (4)	Joint model minus the specialist of the same architecture on test domain d; positive values favor the joint model	RQ2
$C(d, s, \ell)$	Equation (5)	Small joint model minus a larger in-domain specialist of the same family; a pair meets the criterion if $C \geq -0.01$ on both test domains	RQ3
$A(d, d, s) - A(d, d, \ell)$	Section 2.7	As C, but with the small model trained on the target cultivar alone (baseline defined after testing)	RQ3

3. Results

3.1. Performance Across Training Regimes and Test Domains

All 75 training runs and 225 test evaluations were completed (Figure 3; Table 5, Tables S1 and S2). In the four in-domain conditions (specialist and joint models tested on domains represented in training), mAP50–95 ranged from 0.853 to 0.910. The spread across the 25 variants within a condition (0.022–0.038) was several times smaller than the mean transfer shortfalls (0.162 and 0.336). Larger models tended to score slightly higher (Spearman ρ between mAP50–95 and GFLOPs 0.21–0.60), and the smallest size level (N) had the lowest mean in-domain mAP50–95 across all four conditions.

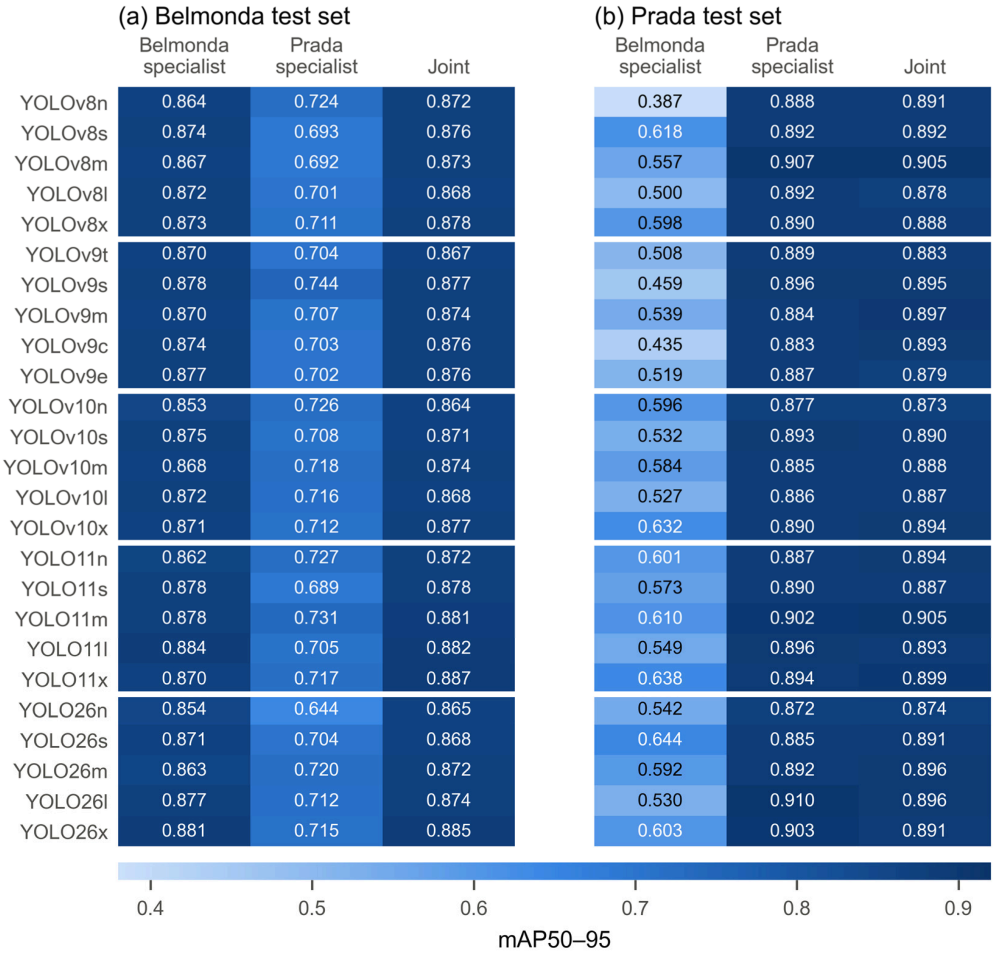


Figure 3. Heat maps of mAP50–95 of the 25 YOLO variants (rows) under the three training regimes (columns) on (**a**) the Belmonda test set and (**b**) the Prada test set. Each cell shows the value of one model; both panels share the same color scale, with darker cells denoting higher values. The lighter middle column in (**a**) shows Prada specialists transferred to Belmonda; the lighter left column in (**b**) shows Belmonda specialists transferred to Prada.

Table 5. Test performance across the 25 YOLO variants: mean (minimum–maximum). Precision, recall and F1 are micro-averaged at a confidence threshold of 0.25 and IoU 0.50.

Training Regime	Test Set	mAP50–95	mAP50	Precision	Recall	F1
Belmonda	Belmonda	0.871 (0.853–0.884)	0.958 (0.937–0.969)	0.900 (0.863–0.931)	0.937 (0.919–0.953)	0.918 (0.890–0.942)
Belmonda	Prada	0.555 (0.387–0.644)	0.638 (0.451–0.732)	0.531 (0.317–0.700)	0.608 (0.374–0.719)	0.566 (0.343–0.708)
Prada	Prada	0.891 (0.872–0.910)	0.962 (0.949–0.975)	0.921 (0.883–0.938)	0.947 (0.928–0.964)	0.934 (0.905–0.951)
Prada	Belmonda	0.709 (0.644–0.744)	0.811 (0.763–0.844)	0.768 (0.688–0.811)	0.821 (0.788–0.842)	0.793 (0.735–0.826)
Joint	Belmonda	0.874 (0.864–0.887)	0.959 (0.949–0.967)	0.906 (0.869–0.934)	0.937 (0.925–0.950)	0.922 (0.897–0.941)
Joint	Prada	0.890 (0.873–0.905)	0.961 (0.946–0.974)	0.919 (0.887–0.941)	0.946 (0.932–0.957)	0.933 (0.910–0.949)

3.2. Cross-Cultivar Transfer and Directional Asymmetry (RQ1)

Every model fell short of the target-domain specialist when applied to the other cultivar, and the shortfall was larger from Belmonda to Prada for all 25 variants (Figure 4a; Table A1). $L(\mathrm{B} \rightarrow \mathrm{P})$ ranged from 0.241 (YOLO26s) to 0.501 (YOLOv8n), with a mean of 0.336, and $L(\mathrm{P} \rightarrow \mathrm{B})$ from 0.127 (YOLOv10n) to 0.211 (YOLO26n), with a mean of 0.162. Belmonda specialists thus reached on average 62% of the mAP50–95 of Prada specialists on Prada images, and Prada specialists 81% of that of Belmonda specialists on Belmonda images. The asymmetry D was positive for all variants (0.074–0.361; mean 0.174). Shortfalls varied between architectures, particularly from Belmonda to Prada (standard deviation 0.064 versus 0.019), but showed no consistent monotonic association with capacity (Spearman ρ with parameters or GFLOPs between −0.06 and 0.20). Within YOLOv9, for example, the Belmonda → Prada shortfall was lowest for YOLOv9m (0.345) and highest for YOLOv9c (0.448). YOLOv8l also fell further short than YOLOv8s (0.392 versus 0.274).

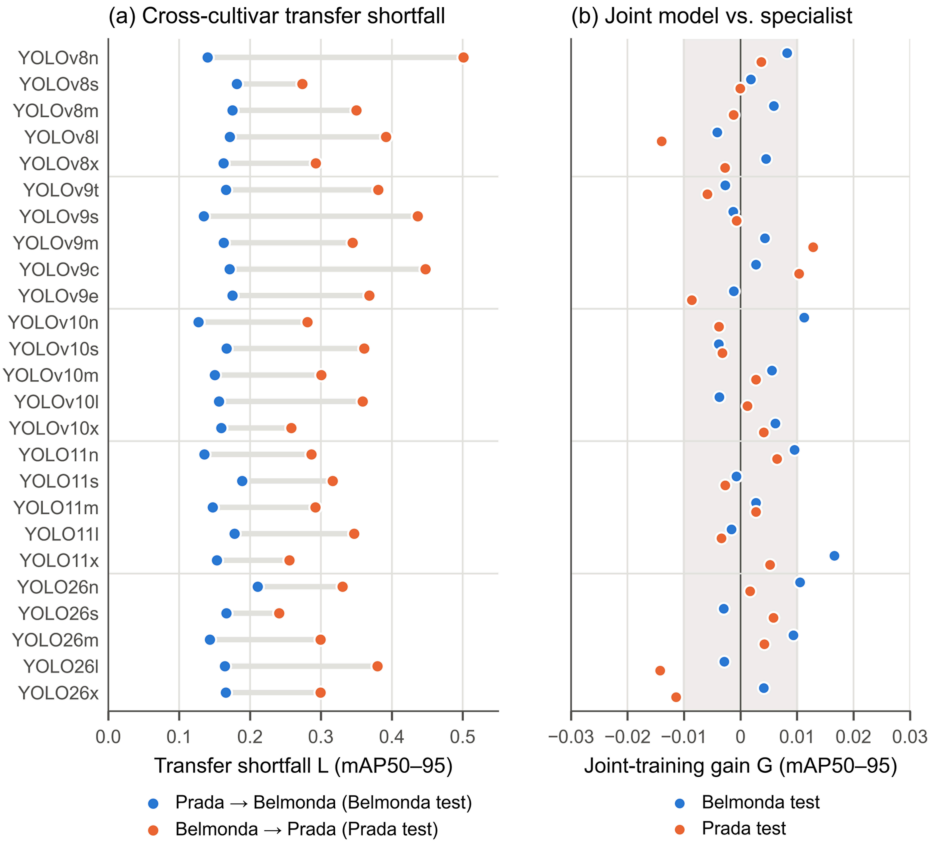


Figure 4. (**a**) Transfer shortfall L of each YOLO variant relative to the target-domain specialist of the same architecture, in both directions (orange, Belmonda → Prada evaluated on the Prada test set; blue, Prada → Belmonda evaluated on the Belmonda test set). (**b**) Joint-training gain G (joint model minus specialist) on each test set (blue, Belmonda; orange, Prada); the shaded band marks ±0.01 mAP50–95 as a reference margin, not a confidence interval. Each colored point represents one of the 25 variants in one transfer direction or test domain; gray lines in (**a**) connect the two directions for the same variant. Values are from a single training seed. Positive values in (**a**) denote a transfer shortfall, and positive values in (**b**) favor the joint model.

3.3. Class-Specific Performance and Error Patterns (RQ1)

The feed class was the most sensitive to the cultivar change (Table 6). Averaged over the 25 variants, the per-class shortfall from Belmonda to Prada was 0.140 for edible, 0.552 for feed and 0.316 for rotten annotations, and from Prada to Belmonda 0.082, 0.287 and 0.117, respectively.

Table 6. Mean per-class AP50–95 across the 25 variants (training regime → test set) and mean per-class transfer shortfalls. Test-set support: Belmonda 900/320/235 and Prada 733/216/147 edible/feed/rotten annotations.

Regime → Test Set	Edible	Feed	Rotten
Belmonda → Belmonda	0.908	0.831	0.874
Prada → Belmonda	0.826	0.544	0.757
Joint → Belmonda	0.911	0.834	0.878
Prada → Prada	0.927	0.869	0.876
Belmonda → Prada	0.788	0.317	0.560
Joint → Prada	0.929	0.868	0.874
Mean L(B → P)	0.140	0.552	0.316
Mean L(P → B)	0.082	0.287	0.117

At the fixed operating point, errors under the cultivar change were dominated by grade confusion rather than missed tubers (Figure 5; Section 3.6). Under same-cultivar testing, the main confusion was between edible and feed (12.8% of feed annotations were predicted as edible by Belmonda specialists). Under transfer, the confusions became directional. Belmonda specialists applied to Prada predicted 34.6% of edible annotations as feed and 25.8% of feed annotations as rotten, i.e., they assigned Prada tubers to a lower grade. Conversely, Prada specialists applied to Belmonda predicted 32.0% of feed annotations as edible and 15.2% of rotten annotations as feed. Figure 6 illustrates the dominant error on a Prada image. In this image, the Belmonda specialist detected all 14 tubers but assigned 3 edible tubers to feed and 1 feed tuber to rotten, and it produced overlapping edible and feed boxes for 2 further edible tubers, whereas the joint model graded all 14 tubers correctly. The example illustrates, in one selected scene, the grade confusions that dominate the errors of transferred models across the test sets (Section 3.6).

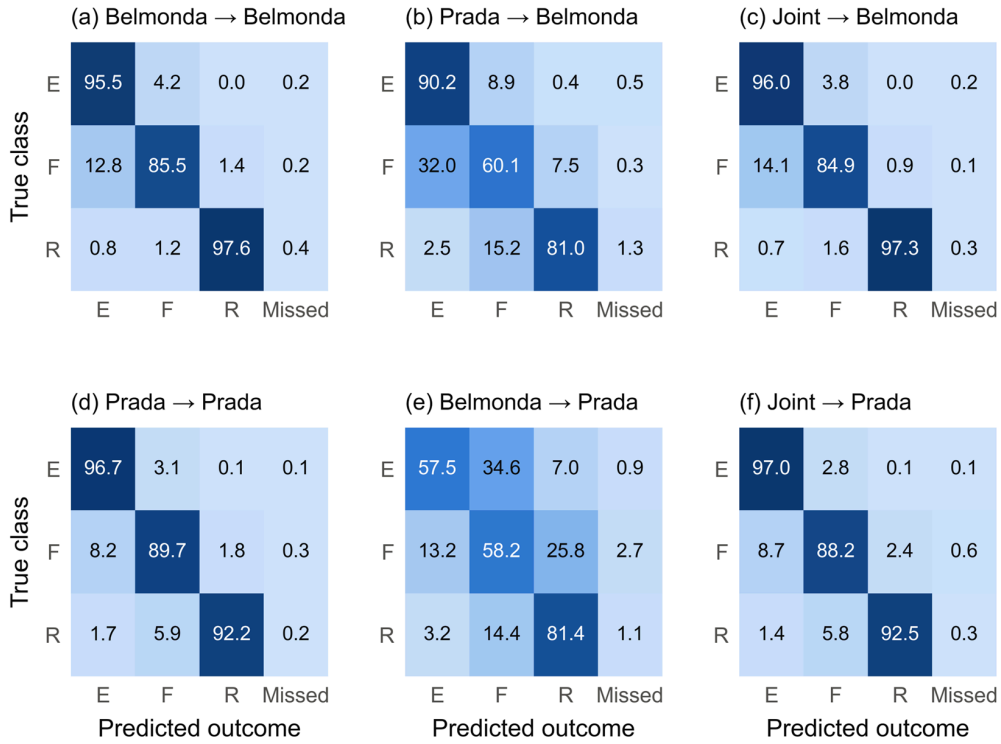


Figure 5. True-class outcome profiles pooled across the 25 model variants at a confidence threshold of 0.25 and IoU 0.50 (training regime → test set). Panels (**a**–**c**) show Belmonda → Belmonda, Prada → Belmonda and Joint → Belmonda, respectively; panels (**d**–**f**) show Prada → Prada, Belmonda → Prada

and Joint → Prada, respectively. Rows denote true classes, and columns denote predicted classes or Missed. Values are percentages of annotation-level outcomes within each true class on a common 0–100% scale; Missed denotes annotations without a matched prediction. Unmatched predictions are reported in Section 3.6; percentages may not sum to 100 because of rounding. E, edible; F, feed; R, rotten.

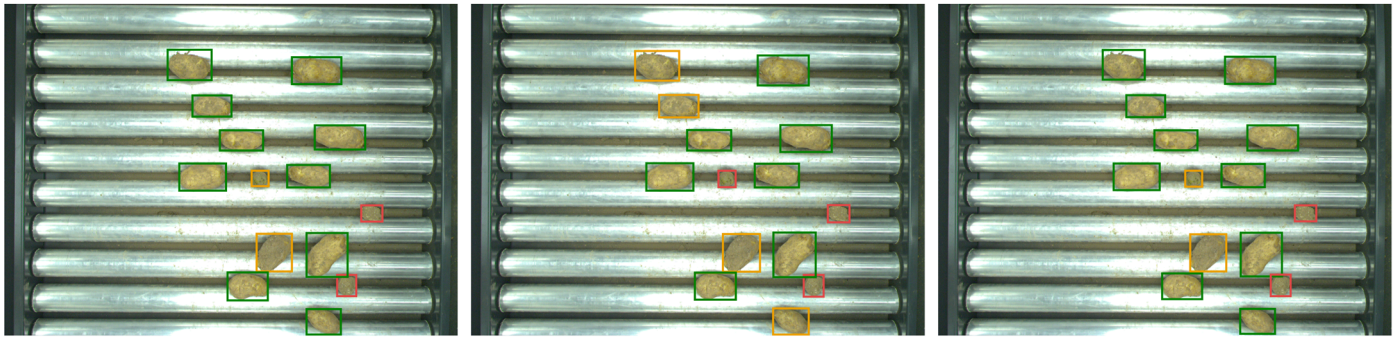


Figure 6. Illustrative Prada test image: (**a**) annotations; YOLO11m predictions after (**b**) Belmonda and (**c**) joint training, at a confidence threshold of 0.25; edible (green), feed (yellow), rotten (red). The image was selected from the archived GPU (FP16) evaluation results by a fixed rule (10–15 annotations, specialist F1 closest to its test-set median, joint F1 $\geq$ 0.9). The displayed predictions were regenerated on a CPU in FP32; specialist TP/FP/FN were 9/7/5 (F1 = 0.600) for this inference versus 10/6/4 (F1 = 0.667) in the archived evaluation, and the joint counts were 14/0/0 (F1 = 1.000) in both. The panel was generated with Ultralytics 8.3.191 and is a selected illustration, not the output underlying the quantitative results.

3.4. Joint Training Versus Cultivar Specialists (RQ2)

Joint models achieved mAP50–95 close to the specialists of the same architecture on both test sets (Table 5; Figure 4b; Table A1). *G* ranged from −0.004 to +0.017 on Belmonda (mean +0.003) and from −0.014 to +0.013 on Prada (mean −0.0004). For 16 and 15 of the 25 variants, respectively, the absolute difference was below 0.005. Only three joint models, all on Prada, were more than 0.01 below their specialist (YOLO26l, YOLOv8l and YOLO26x). Joint-training gains showed no consistent association with capacity (maximum $|\rho| = 0.18$). Per class, joint-training gains were also small on average (−0.002 to +0.003), although individual models varied more for the less frequent classes (feed −0.041 to +0.031; rotten −0.033 to +0.031). The class-outcome profiles of joint models were similar to those of the corresponding specialists (Figure 5).

3.5. Accuracy–Capacity Trade-Offs and Small-Model Comparisons (RQ3)

Small joint models were no more than 0.01 below the larger specialist on both test sets in 17 of the 30 pre-specified pairs (Tables 7 and A2, Figure 7). Small single-cultivar specialists met the same criterion in 17 pairs, and the two outcomes agreed in 28 of 30 pairs, so compact-model competitiveness was not specific to joint training. The outcome depended on the family; the criterion was met in most YOLOv8, YOLOv9 and YOLO11 pairs but in only one of six YOLO26 pairs. Some N and S models failed only against a particularly strong M specialist on Prada (e.g., YOLOv8m, 0.907). For example, the joint YOLO11n (2.6 M parameters, 25 GFLOPs) differed by +0.002 and +0.000 from the YOLO11x specialist (56.8 M, 778 GFLOPs). These are single-seed descriptive outcomes, not equivalence tests.

Table 7. Number of pre-specified within-family pairs (N or S versus M, L or X) in which the small model was no more than 0.01 mAP50–95 below the larger in-domain specialist on both test sets. Complete pair values are given in Table A2.

Family	Small Joint Model	Small Single-Cultivar Specialist
YOLOv8	4/6	4/6
YOLOv9	5/6	6/6
YOLOv10	3/6	3/6
YOLO11	4/6	3/6
YOLO26	1/6	1/6
All families	17/30	17/30

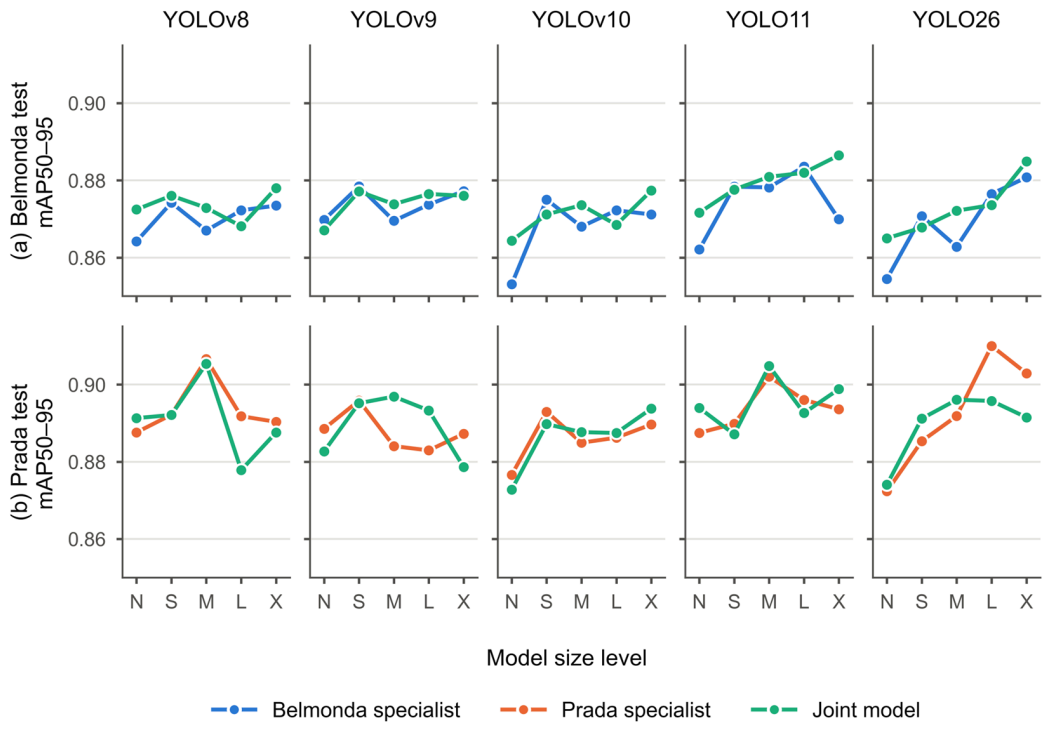


Figure 7. mAP50–95 of the in-domain specialists and joint models by YOLO family and size level on (**a**) the Belmonda test set and (**b**) the Prada test set. Size levels N–X correspond to the variants in Table 2 (for YOLOv9: t, s, m, c and e). Blue and orange lines denote the Belmonda and Prada specialists, respectively; green lines denote joint models. Lines connect the five size levels of each family. All panels use truncated vertical axes, which emphasize differences of about 0.01. Single-seed values; transferred models are shown in Figures 3 and 4.

3.6. Detection Errors and Empty Scenes

At the operating point, precision and recall decreased together under transfer (Table 5). Belmonda specialists reached a precision of 0.900 and a recall of 0.937 on Belmonda but 0.531 and 0.608 on Prada. Prada specialists reached 0.921 and 0.947 on Prada but 0.768 and 0.821 on Belmonda. Mean precision and recall for joint models differed from those of the specialists by at most 0.007. Table 8 decomposes the errors by source: grade confusions (Section 3.3) accounted for 96–97% of all FNs and 58–70% of all FPs in every regime–test combination. Under transfer, their mean number rose to 415 on Prada (Belmonda specialists) and 252 on Belmonda (Prada specialists), compared with 56 and 89 for the target-domain specialists. Mean missed rates were low (Figure 5), although individual-model missed rates per class reached 8.8%.

Table 8. Decomposition of the errors at the operating point (confidence threshold 0.25, IoU 0.50): mean (minimum–maximum) counts per evaluation across the 25 YOLO variants. Grade confusions are matched prediction–annotation pairs with different classes (one FP and one FN each); missed are unmatched annotations (FN); unmatched are predictions without a matching annotation (FP). The Belmonda and Prada test sets contain 1455 and 1096 annotations, respectively. Counts were rounded independently.

Training Regime	Test Set	TP	Grade Confusions (Events)	Missed (FN)	Unmatched (FP)
Belmonda	Belmonda	1363 (1337–1386)	89 (66–114)	4 (1–21)	63 (33–98)
Belmonda	Prada	667 (410–788)	415 (255–683)	14 (2–71)	185 (72–364)
Prada	Prada	1038 (1017–1057)	56 (38–77)	2 (0–5)	33 (22–58)
Prada	Belmonda	1195 (1147–1225)	252 (221–304)	8 (3–33)	111 (61–217)
Joint	Belmonda	1364 (1346–1382)	88 (71–105)	3 (1–12)	53 (22–99)
Joint	Prada	1037 (1022–1049)	56 (44–73)	3 (0–13)	36 (20–62)

No detections above a confidence of 0.25 occurred on any of the six empty test images for any of the 75 models (Table S3).

3.7. Performance Stratified by Temporal Proximity to Training Images

Specialist and joint-model performance did not decrease monotonically with increasing time gap to the nearest training image of the same cultivar (Table 9). For the same-cultivar specialists, mean micro-F1 in the nearest (≤5 s) and most distant (>10 s) strata was 0.915 and 0.921 on Belmonda and 0.932 and 0.948 on Prada. The transfer shortfalls in F1 remained substantial in both strata (0.353 and 0.364 from Belmonda to Prada; 0.115 and 0.126 in the reverse direction). However, timestamps do not identify tubers, and even the most distant Prada test images were captured within 32 s of a training image. This diagnostic therefore does not establish independence between the subsets or exclude optimistic estimates caused by repeated views.

Table 9. Mean micro-F1 across the 25 variants by the time gap between each test image and its nearest training image of the same cultivar. The same target-domain strata were used for all models; transferred models had not been trained on that target domain.

Training Regime	Test Set	≤5 s	6–10 s	>10 s
Belmonda	Belmonda	0.915	0.918	0.921
Belmonda	Prada	0.578	0.549	0.584
Prada	Prada	0.932	0.926	0.948
Prada	Belmonda	0.800	0.788	0.796
Joint	Belmonda	0.917	0.924	0.924
Joint	Prada	0.935	0.925	0.943

Number of test images per stratum (≤5 s/6–10 s/>10 s): Belmonda 32/41/22; Prada 23/44/28.

3.8. Image Characteristics of the Two Cultivar Domains

Prada crops had a higher mean b* (a more yellow appearance) and greater within-crop lightness variability (SD(L*)) than Belmonda crops in every grade (Table 10). In both cultivars, lightness decreased and its variability increased from edible to feed to rotten annotations, but the Prada values were shifted: edible Prada crops were closer to Belmonda feed crops in mean L* (65.2 versus 64.7) and SD(L*) (5.06 versus 4.75) than to Belmonda

edible crops (67.5 and 3.17), and Belmonda feed crops were closer to Prada edible crops than to Prada feed crops (60.8 and 6.40). Belmonda images contained more annotations per image than Prada images (15.3 versus 11.6, SD 4.4 versus 1.2), whereas the box areas of the same grade were similar. These descriptors characterize recorded image appearance and do not measure calibrated skin color or surface roughness; the crops do not represent independent tubers.

Table 10. Annotation-weighted mean image descriptors by cultivar domain and annotated grade. CIELAB color coordinates and SD(L*) were computed from the central 60% of each bounding box's width and height. The Crops column gives the number of analyzed annotations; SD(L*) denotes within-crop lightness standard deviation; area denotes full bounding-box area in square pixels (px^2).

Cultivar	Grade	Crops	L*	a*	b*	SD(L*)	Area (px^2)
Belmonda	Edible	8973	67.5	−6.6	24.1	3.17	31,401
Belmonda	Feed	3057	64.7	−6.5	22.9	4.75	29,541
Belmonda	Rotten	2434	56.6	−6.1	20.2	5.81	21,591
Prada	Edible	7234	65.2	−4.4	32.3	5.06	32,882
Prada	Feed	2185	60.8	−4.5	29.0	6.40	33,945
Prada	Rotten	1507	54.9	−4.7	22.9	7.38	22,271

4. Discussion

4.1. Cross-Cultivar Transfer, Asymmetry and Class Errors (RQ1)

Detectors trained on one cultivar domain fell well short of target-domain specialists on the other, and the shortfall depended strongly on the direction. Because the asymmetry appeared in all 25 variants, it reflects a difference between the two recorded domains rather than a property of a particular detector, consistent with the dataset-bias and agricultural domain-shift literature [26,30,31]. In contrast, cross-validation among five potato cultivars gave satisfactory generalization for the detection of tuber eyes [45]. This suggests that sensitivity to the cultivar depends on the task. Quality grading, which relies on subtle differences in surface appearance, was strongly affected here. The Belmonda → Prada shortfall was also larger than the within-model cross-location drops reported with the same platform in [25], but the contrasts and evaluation protocols differ, which prevents attributing the difference to cultivar rather than location.

The class-level results indicate that the dominant failure was not the detection of tubers but the position of the grade boundaries. Grade confusions accounted for 96–97% of all FNs (Table 8), at least at IoU 0.50 and the chosen operating point. Belmonda-trained models downgraded Prada tubers (edible → feed and feed → rotten), whereas Prada-trained models upgraded Belmonda feed tubers to edible. Because grading is ordinal, such shifts are consistent with domain differences in the visual cues the models associate with each grade, such as skin color and texture, brightness, soil adherence or the appearance of surface defects. A similar edible–feed confusion was attributed to subtle differences in skin tone across locations in [25]. The image descriptors are consistent with this interpretation (Section 3.8; Table 10). In both cultivars, lightness decreased and its variability increased from edible to feed to rotten annotations, but Prada crops were darker and had greater lightness variability within each grade. Edible Prada crops were closer to Belmonda feed crops than to Belmonda edible crops in these lightness descriptors, and Belmonda feed crops were closer to Prada edible crops than to Prada feed crops. This partial resemblance is consistent with the downgrading of Prada tubers by Belmonda-trained models and the upgrading of Belmonda feed tubers by Prada-trained models. The descriptors do not, however, establish which cues the detectors used. The chromatic coordinates also differed, the domains differed in annotation density and class composition, and cultivar remains confounded with harvest season, acquisition period and unrecorded lot and storage conditions.

The feed class was the least robust across all analyses; it had the lowest same-cultivar AP, the largest transfer shortfalls and was the main source of confusion in both directions, in agreement with [25]. Feed is an intermediate grade whose boundaries with edible and rotten depend on the extent of visible defects, which makes it sensitive to a shift of the decision boundaries. Feed is defined by exclusion, with regard to tubers with visible defects that do not amount to advanced decay, and this grade therefore covers a broad and heterogeneous range of appearances that may overlap with both neighboring grades; in both cultivars, the mean values of feed crops lay between those of edible and rotten crops in lightness and lightness variability (Table 10). A narrower or defect-specific definition of this grade might reduce such confusions but would require re-annotation, and agreement between labelers on the boundaries of this grade was not measured. Averaged across the 25 transferred models, 3.2% of rotten annotations were predicted as edible for Belmonda → Prada and 2.5% in the reverse direction. Individual-model rates, however, reached 10.9% (YOLO26m) and 6.8% (YOLOv8l), respectively. Pooled averages therefore do not bound the error rate of a deployed model. Downgrading edible tubers to feed or rotten, which represents an economic loss, was frequent from Belmonda to Prada. These errors concern operational visual grades, not validated consumption suitability.

Increasing model size did not reduce the shortfall consistently across the tested variants. YOLO26l, the configuration that transferred best across locations in [25], showed an above-median Belmonda → Prada shortfall here (0.380), whereas YOLO26s had the lowest (0.241). Robustness observed on one external test set did not necessarily carry over to this domain shift.

4.2. Joint Training, Model Capacity and Implications for Grading (RQ2 and RQ3)

Joint models achieved mAP50–95 close to the corresponding specialists, with differences from −0.014 to +0.017. This supports considering one joint model for the two recorded domains, but it does not establish statistical equivalence or the preservation of every class-specific error rate; individual feed AP decreased by up to 0.041. Joint training also changed data quantity, domain diversity and training budget, and their contributions were not isolated. Nevertheless, equal image counts did not ensure transferability. Both specialists were trained on 756 non-empty images, yet transferred models fell 0.127–0.501 mAP50–95 below the target-domain specialists, whereas joint models, trained on twice as many non-empty images, showed only small mean differences from the specialists on each cultivar (mean G of +0.003 on Belmonda and −0.0004 on Prada). These comparisons support including labeled examples of the target domain but do not isolate the contribution of diversity from that of additional data and optimizer updates; quantity- and update-matched controls are needed. Mlinarič et al. [32] reported benefits from transferring hyperparameters and feature importance between motor-subtype classifiers. This provides an industrial analogy rather than direct evidence for pooled visual-domain training. A supervised joint model differs from domain adaptation and generalization methods that assume no target labels [33,35].

The competitiveness of compact models was not specific to joint training; small joint models and small single-cultivar specialists each met the 0.01 margin against larger specialists in 17 of 30 comparisons. Within each family, mAP50–95 varied by only 0.009–0.038 across the five size levels and increased monotonically with size in only 2 of the 20 curves defined by family, training regime and test domain (Figure 7). YOLO26, the family with the largest range, was the only one whose small models met the margin in just one of six pairs (Table 7). The practical distinction is that one joint model serves both recorded domains, whereas the specialist baseline requires a separate model for each. Because the in-domain spread across the 25 variants was only 0.022–0.038, individual pair outcomes may change with a different seed and should not be interpreted as stable family rankings.

Reductions in parameters and GFLOPs are relevant for embedded deployment [15] but do not by themselves establish latency or energy savings. Inference speed, peak memory use and energy consumption were not measured in this study. In this benchmark, for example, the joint YOLO11n reached the mAP50–95 of the YOLO11x specialist with about 3% of its GFLOPs, but the relative inference speed on an embedded device depends on the accelerator, numerical precision and software stack. An engineering comparison should therefore first identify the models that meet the throughput and memory requirements of the grading line on the target device and then select, among them, the model with the lowest rate of consequential grading errors; a smaller model that meets these requirements may leave computing headroom for tracking and multi-view aggregation.

With regard to model selection, joint YOLOv9s and YOLOv10s met the descriptive margin against all three larger specialists in their respective families on both test domains, and most other N and S variants of YOLOv8, YOLOv9 and YOLO11 met it in two of three pairs, whereas joint YOLOv10n and YOLO26n met it in none (Table A2). A compact YOLOv8, YOLOv9, YOLOv10 or YOLO11 model trained jointly on all relevant cultivars therefore appears to be a reasonable starting point for deployment evaluation within the recorded domains. This does not establish a stable family ranking, because seed variability was not estimated and differences between families were small; the final choice requires device-specific speed measurements and validation of consequential grade-confusion rates.

Whether a joint model trained on three or more cultivars would perform equally well on each of them cannot be inferred from two cultivar domains. Within these two domains, joint training recovered near-specialist performance, whereas increasing capacity did not consistently reduce the transfer shortfall. With more cultivars, however, grade appearances that conflict between cultivars, such as the partial resemblance between edible Prada and feed Belmonda crops in the lightness descriptors, may compete within one model. Additional cultivars should therefore be introduced and evaluated incrementally, with results reported separately for each cultivar and on independent lots.

For grading practice, a model trained on one cultivar should not be assumed to grade another cultivar correctly, even if it detects every tuber. A systematic shift between adjacent grades changes the proportions of tubers directed to consumption, feed or waste streams. Within the recorded sessions, including labeled images of each cultivar in a joint training set restored performance close to that of cultivar-specific models. In a grading line, the final decision concerns a tuber rather than a single detection. When a tuber is tracked over several frames, the per-frame predictions can be aggregated, for example by majority vote or by giving priority to the most severe grade observed; such aggregation may reduce sporadic errors when complementary views are available, but correlated errors and systematic grade shifts such as those observed under transfer may persist across views. Tracking, aggregation rules and tuber-level reliability were not evaluated here.

No detections above a confidence of 0.25 occurred on the six recorded empty test images. This limited check under the recorded conditions does not estimate an operational false-alarm rate, isolate the effect of training with empty images or establish reliability under practical operating conditions, such as soil, debris, plant residues or changing illumination.

4.3. Limitations and Future Work

Several limitations restrict the generality of these results. First, repeated views of unidentified tubers may cross training, validation and test, creating dependence that can yield optimistic estimates [46]; the results therefore do not establish performance on independent tubers or lots. The temporal-proximity diagnostic (Section 3.7) cannot quantify such inflation or establish independence. Second, each cultivar domain combines a cultivar with its acquisition period, so cultivar effects cannot be separated from lot and season effects.

Third, each configuration was trained once, so differences of about 0.01 between individual models are not reliable. A single seed was used to prioritize the coverage of 25 variants and three training regimes (approximately 98.5 GPU-hours for the completed runs, excluding repeated runs); the consistent direction of the transfer shortfall across all variants strengthens the descriptive pattern but does not replace repeated training. Fourth, the joint regime differed from the specialists in data quantity, diversity and number of updates, and no quantity-matched control was included. Fifth, the negative test contained six images from a single day, which is insufficient to establish reliability in practical grading environments. Sixth, the Belmonda images had been used in previous research [25], although all models in the present study were trained from COCO-pretrained weights. Finally, the evaluation concerned individual detections on still images under controlled illumination; tuber-level decisions, tracking across frames, rejection mechanics and on-device inference were not evaluated, and inference latency, memory use and energy consumption on embedded hardware were not measured.

Future work should validate the findings on independently recorded lots, with separation by capture day and tuber identity, on additional cultivars and seasons and with several training seeds. A joint regime matched in image count and number of updates (e.g., 378 images of each cultivar) would help to assess the contributions of domain composition and training-data quantity. A separate system-level study will address embedded inference, latency, multi-frame tuber-level decisions and integration with the rejection mechanism of the grading line.

5. Conclusions

In this single-seed benchmark of 25 YOLO variants and three training regimes, models trained on one potato cultivar domain fell short of target-domain specialists by an average of 0.336 mAP50–95 from Belmonda to Prada and 0.162 in the reverse direction. The shortfall did not decrease consistently with model capacity. Transfer errors mainly involved confusion between adjacent visual grades rather than missed tubers, with the feed class most affected (RQ1). A model trained on one cultivar should therefore not be assumed to grade another cultivar correctly.

Including labeled images of both cultivars in training avoided this shortfall. Joint models differed from the specialists by only −0.014 to +0.017 mAP50–95 on each test set, and their mean precision and recall were within 0.007 of those of the specialists (RQ2). For the two recorded cultivar domains, a joint model thus performed similarly to the specialist of the same architecture on each cultivar. In practice, one model could serve a grading line that processes these cultivars, provided that labeled images of every cultivar are included in training. Compact models were often sufficient regardless of the training regime; small joint models and small single-cultivar specialists each met a one-sided 0.01 margin against larger specialists in 17 of 30 comparisons, with a strong dependence on the family (RQ3). Whether this translates into lower latency or energy consumption on low-cost hardware was not measured. No detections occurred on the six empty test images used as a practical check, which is insufficient to establish reliability in practical grading environments.

These results concern held-out images from shared acquisition sessions in which cultivar is confounded with lot and season; they do not establish statistical equivalence or generalization to independent lots or additional cultivars, which require further validation. The benefits of joint training and compact models therefore require validation on independent tubers and lots, additional cultivars and other acquisition conditions, with repeated training seeds and device-specific measurements of speed, memory use and energy consumption, before they guide the design of a grading line.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/foods15224138/s1, Table S1: Test metrics of all 225 evaluations; Table S2: Training records of all 75 runs; Table S3: Per-image results of the negative test.

Author Contributions: Conceptualization, M.S., J.H., S.K. and T.P.; methodology, M.S., J.H. and T.P.; software, M.S.; validation, M.S., J.H. and T.P.; formal analysis, M.S. and T.P.; investigation, M.S., J.H. and T.P.; data curation, M.S., J.H. and T.P.; writing—original draft preparation, M.S. and T.P.; writing—review and editing, M.S., J.H., S.K. and T.P.; visualization, M.S. and T.P.; supervision, M.S. and S.K.; project administration, M.S. and S.K.; funding acquisition, M.S. and S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Innovation Partnership (EIP) project OPTIFARM—Development and Implementation of a Low-cost Optical Sorter in Crop Production on Farms (IRP31). The authors acknowledge the financial support from the Slovenian Research Agency (Research Core Funding No. P2-0157).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The images, annotations, frozen split manifests, training configurations and analysis code supporting this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors acknowledge the use of the HPC Maister system at the University of Maribor for model training and evaluation.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AP	Average precision
CIELAB	International Commission on Illumination L*a*b* color space
COCO	Common Objects in Context
CPU	Central processing unit
EIP	European Innovation Partnership
FN	False negative
FP	False positive
FP16	16-bit floating point
FP32	32-bit floating point
GFLOPs	Giga floating-point operations
GPU	Graphics processing unit
HPC	High-performance computing
IoU	Intersection over union
LED	Light-emitting diode
MACs	Multiply–accumulate operations
mAP	Mean average precision
NMS	Non-maximum suppression
PNG	Portable Network Graphics
RAM	Random-access memory
RGB	Red, green, blue
RQ	Research question
SD	Standard deviation
SHA-256	Secure Hash Algorithm 256-bit
sRGB	Standard red, green, blue color space
TP	True positive
YOLO	You Only Look Once

Appendix A

Table A1 lists the per-model transfer shortfalls and joint-training gains shown in Figure 4. Table A2 presents the complete pre-specified capacity comparisons summarized in Table 7, and Table A3 summarizes training behavior by regime.

Table A1. Transfer shortfalls, directional asymmetry and joint-training gains (absolute mAP50–95 differences).

Model	$L(\mathbf{B} \rightarrow \mathbf{P})$	$L(\mathbf{P} \rightarrow \mathbf{B})$	D	G (B)	G (P)
YOLOv8n	0.501	0.140	0.361	0.008	0.004
YOLOv8s	0.274	0.181	0.092	0.002	0.000
YOLOv8m	0.350	0.175	0.175	0.006	−0.001
YOLOv8l	0.392	0.172	0.220	−0.004	−0.014
YOLOv8x	0.293	0.162	0.130	0.005	−0.003
YOLOv9t	0.381	0.166	0.215	−0.003	−0.006
YOLOv9s	0.436	0.135	0.302	−0.001	−0.001
YOLOv9m	0.345	0.163	0.182	0.004	0.013
YOLOv9c	0.448	0.171	0.276	0.003	0.010
YOLOv9e	0.368	0.175	0.193	−0.001	−0.009
YOLOv10n	0.281	0.127	0.154	0.011	−0.004
YOLOv10s	0.361	0.167	0.194	−0.004	−0.003
YOLOv10m	0.300	0.150	0.150	0.006	0.003
YOLOv10l	0.359	0.156	0.203	−0.004	0.001
YOLOv10x	0.258	0.159	0.099	0.006	0.004
YOLO11n	0.287	0.135	0.151	0.010	0.006
YOLO11s	0.317	0.189	0.128	−0.001	−0.003
YOLO11m	0.292	0.147	0.145	0.003	0.003
YOLO11l	0.347	0.178	0.169	−0.002	−0.003
YOLO11x	0.255	0.153	0.102	0.017	0.005
YOLO26n	0.330	0.211	0.120	0.011	0.002
YOLO26s	0.241	0.167	0.074	−0.003	0.006
YOLO26m	0.299	0.143	0.156	0.009	0.004
YOLO26l	0.380	0.165	0.215	−0.003	−0.014
YOLO26x	0.299	0.166	0.134	0.004	−0.011
Mean	0.336	0.162	0.174	0.003	0.000

Table A2. Pre-specified capacity comparisons. C is the mAP50–95 difference between the small model and the larger in-domain specialist of the same family, for the small joint model and, as an additional baseline, for the small single-cultivar specialist; ✓ marks $C \geq -0.01$ on both test sets (evaluated before rounding). Ratios refer to the small relative to the large model.

Small Model	Large Specialist	Param. Ratio	GFLOPs Ratio	Joint C (B)	Joint C (P)	Joint Both	Spec. C (B)	Spec. C (P)	Spec. Both
YOLOv8n	YOLOv8m	0.12	0.10	0.005	−0.015	–	−0.003	−0.019	–
YOLOv8n	YOLOv8l	0.07	0.05	0.000	−0.001	✓	−0.008	−0.004	✓
YOLOv8n	YOLOv8x	0.04	0.03	−0.001	0.001	✓	−0.009	−0.003	✓
YOLOv8s	YOLOv8m	0.43	0.36	0.009	−0.014	–	0.007	−0.014	–
YOLOv8s	YOLOv8l	0.26	0.17	0.004	0.000	✓	0.002	0.000	✓
YOLOv8s	YOLOv8x	0.16	0.11	0.003	0.002	✓	0.001	0.002	✓
YOLOv9t	YOLOv9m	0.10	0.10	−0.002	−0.001	✓	0.000	0.004	✓
YOLOv9t	YOLOv9c	0.08	0.07	−0.007	0.000	✓	−0.004	0.006	✓
YOLOv9t	YOLOv9e	0.03	0.04	−0.010	−0.005	–	−0.007	0.001	✓
YOLOv9s	YOLOv9m	0.36	0.35	0.008	0.011	✓	0.009	0.012	✓
YOLOv9s	YOLOv9c	0.28	0.26	0.003	0.012	✓	0.005	0.013	✓
YOLOv9s	YOLOv9e	0.12	0.14	0.000	0.008	✓	0.001	0.009	✓

Table A2. *Cont.*

Small Model	Large Specialist	Param. Ratio	GFLOPs Ratio	Joint *C* (B)	Joint *C* (P)	Joint Both	Spec. *C* (B)	Spec. *C* (P)	Spec. Both
YOLOv10n	YOLOv10m	0.15	0.11	−0.004	−0.012	–	−0.015	−0.008	–
YOLOv10n	YOLOv10l	0.09	0.05	−0.008	−0.013	–	−0.019	−0.010	–
YOLOv10n	YOLOv10x	0.08	0.04	−0.007	−0.017	–	−0.018	−0.013	–
YOLOv10s	YOLOv10m	0.47	0.36	0.003	0.005	✓	0.007	0.008	✓
YOLOv10s	YOLOv10l	0.30	0.18	−0.001	0.004	✓	0.003	0.007	✓
YOLOv10s	YOLOv10x	0.25	0.13	0.000	0.000	✓	0.004	0.003	✓
YOLO11n	YOLO11m	0.13	0.09	−0.007	−0.008	✓	−0.016	−0.015	–
YOLO11n	YOLO11l	0.10	0.07	−0.012	−0.002	–	−0.021	−0.009	–
YOLO11n	YOLO11x	0.05	0.03	0.002	0.000	✓	−0.008	−0.006	✓
YOLO11s	YOLO11m	0.47	0.31	−0.001	−0.015	–	0.000	−0.012	–
YOLO11s	YOLO11l	0.37	0.25	−0.006	−0.009	✓	−0.005	−0.006	✓
YOLO11s	YOLO11x	0.17	0.11	0.008	−0.006	✓	0.008	−0.004	✓
YOLO26n	YOLO26m	0.12	0.08	0.002	−0.018	–	−0.008	−0.019	–
YOLO26n	YOLO26l	0.10	0.06	−0.011	−0.036	–	−0.022	−0.038	–
YOLO26n	YOLO26x	0.04	0.03	−0.016	−0.029	–	−0.026	−0.031	–
YOLO26s	YOLO26m	0.47	0.30	0.005	−0.001	✓	0.008	−0.007	✓
YOLO26s	YOLO26l	0.38	0.24	−0.009	−0.019	–	−0.006	−0.025	–
YOLO26s	YOLO26x	0.17	0.11	−0.013	−0.012	–	−0.010	−0.018	–

YOLOv9t versus YOLOv9e (joint, Belmonda: −0.0102) does not meet the criterion although it rounds to −0.010.

Table A3. Training behavior by regime across the 25 variants: mean (minimum–maximum) number of epochs, mean best epoch, mean number of optimizer updates and training time on one Tesla V100 GPU.

Regime	Epochs	Best Epoch	Optimizer Updates	Mean Time (h)	Total Time (h)
Belmonda	81.7 (49–124)	51.7	1081	0.91	22.9
Prada	96.3 (57–141)	66.3	1265	1.12	28.0
Joint	84.9 (66–103)	54.9	2237	1.91	47.7

All runs stopped early (patience 30). Total training time: 98.5 GPU-hours, excluding repeated runs. Totals were calculated before rounding.

References

1. Devaux, A.; Kromann, P.; Ortiz, O. Potatoes for Sustainable Global Food Security. *Potato Res.* **2014**, *57*, 185–199. [CrossRef]
2. Kader, A.A. Increasing Food Availability by Reducing Postharvest Losses of Fresh Produce. *Acta Hortic.* **2005**, *682*, 2169–2176. [CrossRef]
3. Kuyu, C.G.; Tola, Y.B.; Abdi, G.G. Study on Post-Harvest Quantitative and Qualitative Losses of Potato Tubers from Two Different Road Access Districts of Jimma Zone, South West Ethiopia. *Heliyon* **2019**, *5*, e02272. [CrossRef] [PubMed]
4. Bhargava, A.; Bansal, A. Fruits and Vegetables Quality Evaluation Using Computer Vision: A Review. *J. King Saud Univ. Comput. Inf. Sci.* **2021**, *33*, 243–257. [CrossRef]
5. Danielak, M.; Przybył, K.; Koszela, K. The Need for Machines for the Nondestructive Quality Assessment of Potatoes with the Use of Artificial Intelligence Methods and Imaging Techniques. *Sensors* **2023**, *23*, 1787. [CrossRef] [PubMed]
6. Lorente, D.; Aleixos, N.; Gómez-Sanchis, J.; Cubero, S.; García-Navarrete, O.L.; Blasco, J. Recent Advances and Applications of Hyperspectral Imaging for Fruit and Vegetable Quality Assessment. *Food Bioprocess Technol.* **2012**, *5*, 1121–1142. [CrossRef]
7. Cubero, S.; Aleixos, N.; Moltó, E.; Gómez-Sanchis, J.; Blasco, J. Advances in Machine Vision Applications for Automatic Inspection and Quality Evaluation of Fruits and Vegetables. *Food Bioprocess Technol.* **2011**, *4*, 487–504. [CrossRef]
8. Blasco, J.; Aleixos, N.; Moltó, E. Machine Vision System for Automatic Quality Grading of Fruit. *Biosyst. Eng.* **2003**, *85*, 415–423. [CrossRef]
9. Brosnan, T.; Sun, D.-W. Improving Quality Inspection of Food Products by Computer Vision—A Review. *J. Food Eng.* **2004**, *61*, 3–16. [CrossRef]
10. Zhang, B.; Gu, B.; Tian, G.; Zhou, J.; Huang, J.; Xiong, Y. Challenges and Solutions of Optical-Based Nondestructive Quality Inspection for Robotic Fruit and Vegetable Grading Systems: A Technical Review. *Trends Food Sci. Technol.* **2018**, *81*, 213–231. [CrossRef]
11. Kamilaris, A.; Prenafeta-Boldú, F.X. Deep Learning in Agriculture: A Survey. *Comput. Electron. Agric.* **2018**, *147*, 70–90. [CrossRef]

12. Zhang, Y.X. Simulation-Driven Explainable AI for Quality Prediction and Control. *Int. J. Simul. Model.* **2025**, *24*, 533–544. [CrossRef]
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
14. Terven, J.; Córdova-Esparza, D.-M.; Romero-González, J.-A. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1680–1716. [CrossRef]
15. Li, Y.X.; Chen, H.; Zhou, L. Industrial Multi-Object Detection for Automotive Battery Manufacturing Using a CBAM-Enhanced YOLOv5 Framework. *Adv. Prod. Eng. Manag.* **2026**, *21*, 163–177. [CrossRef]
16. Jocher, G.; Qiu, J.; Chaurasia, A. Ultralytics YOLO, Version 8.4.75. Available online: https://github.com/ultralytics/ultralytics (accessed on 15 October 2026).
17. Mazzia, V.; Khaliq, A.; Salvetti, F.; Chiaberge, M. Real-Time Apple Detection System Using Embedded Systems with Hardware Accelerators: An Edge AI Application. *IEEE Access* **2020**, *8*, 9102–9114. [CrossRef]
18. Li, X.; Wang, F.; Guo, Y.; Liu, Y.; Lv, H.; Zeng, F.; Lv, C. Improved YOLO V5s-Based Detection Method for External Defects in Potato. *Front. Plant Sci.* **2025**, *16*, 1527508. [CrossRef] [PubMed]
19. Li, J.; Gong, G.; Zhang, J.; Liu, X.; Jiang, S.; Wu, X.; Xie, H.; Gu, L. Multi-Task Learning Based Approach for Potato Edibility Detection and Defect Segmentation. *J. Food Compos. Anal.* **2025**, *148*, 108420. [CrossRef]
20. Zhang, K.; Wang, S.; Hu, Y.; Yang, H.; Guo, T.; Yi, X. Evaluation Method of Potato Storage External Defects Based on Improved U-Net. *Agronomy* **2023**, *13*, 2503. [CrossRef]
21. Arshaghi, A.; Ashourian, M.; Ghabeli, L. Potato Diseases Detection and Classification Using Deep Learning Methods. *Multimed. Tools Appl.* **2023**, *82*, 5725–5742. [CrossRef]
22. Wang, R.-F.; Su, W.-H. The Application of Deep Learning in the Whole Potato Production Chain: A Comprehensive Review. *Agriculture* **2024**, *14*, 1225. [CrossRef]
23. Przybył, K.; Boniecki, P.; Koszela, K.; Gierz, Ł.; Łukomski, M. Computer Vision and Artificial Neural Network Techniques for Classification of Damage in Potatoes during the Storage Process. *Czech J. Food Sci.* **2019**, *37*, 135–140. [CrossRef]
24. Verk, J.; Hernavs, J.; Klančnik, S. Using a Region-Based Convolutional Neural Network (R-CNN) for Potato Segmentation in a Sorting Process. *Foods* **2025**, *14*, 1131. [CrossRef] [PubMed]
25. Peršak, T.; Simonič, M.; Hernavs, J.; Ficko, M.; Klančnik, S. OPTIFARM: Benchmarking YOLO Architectures for Location-Robust Potato Quality Detection. *Foods* **2026**, *15*, 2121. [CrossRef] [PubMed]
26. Torralba, A.; Efros, A.A. Unbiased Look at Dataset Bias. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 1521–1528. [CrossRef]
27. Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; Vaughan, J.W. A Theory of Learning from Different Domains. *Mach. Learn.* **2010**, *79*, 151–175. [CrossRef]
28. Mohanty, S.P.; Hughes, D.P.; Salathé, M. Using Deep Learning for Image-Based Plant Disease Detection. *Front. Plant Sci.* **2016**, *7*, 1419. [CrossRef] [PubMed]
29. David, E.; Madec, S.; Sadeghi-Tehran, P.; Aasen, H.; Zheng, B.; Liu, S.; Kirchgessner, N.; Ishikawa, G.; Nagasawa, K.; Badhon, M.A.; et al. Global Wheat Head Detection (GWHD) Dataset: A Large and Diverse Dataset of High-Resolution RGB-Labelled Images to Develop and Benchmark Wheat Head Detection Methods. *Plant Phenomics* **2020**, *2020*, 3521852. [CrossRef] [PubMed]
30. Grbović, Ž.; Panić, M.; Stefanović, D.; Brdar, S.; Crnojević, V.; Lončar-Turukalo, T. Transfer Learning Models for Wheat Ear Detection on Multi-Source Dataset. *Sci. Rep.* **2025**, *15*, 44740. [CrossRef] [PubMed]
31. Rana, S.; Hensel, O.; Nasirahmadi, A. From Vineyard to Vision: Multi-Domain Analysis and Mitigation of Grape Cluster Detection Failures in Complex Viticultural Environments. *Results Eng.* **2026**, *29*, 108833. [CrossRef]
32. Mlinarič, J.; Pregelj, B.; Boškoski, P.; Petrovčič, J.; Dolanc, G. A Transfer Learning Approach to Machine Learning-Based End-of-Line Quality Inspection. *Adv. Prod. Eng. Manag.* **2025**, *20*, 277–290. [CrossRef]
33. Magistri, F.; Weyler, J.; Gogoll, D.; Lottes, P.; Behley, J.; Petrinic, N.; Stachniss, C. From One Field to Another—Unsupervised Domain Adaptation for Semantic Segmentation in Agricultural Robotics. *Comput. Electron. Agric.* **2023**, *212*, 108114. [CrossRef]
34. Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; Van Gool, L. Domain Adaptive Faster R-CNN for Object Detection in the Wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 3339–3348. [CrossRef]
35. Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; Loy, C.C. Domain Generalization: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4396–4415. [CrossRef] [PubMed]
36. Wang, X.; Cai, Z.; Gao, D.; Vasconcelos, N. Towards Universal Object Detection by Domain Attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7289–7298. [CrossRef]
37. Taori, R.; Dave, A.; Shankar, V.; Carlini, N.; Recht, B.; Schmidt, L. Measuring Robustness to Natural Distribution Shifts in Image Classification. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*; Curran Associates: Red Hook, NY, USA, 2020. [CrossRef]

38. Wang, C.-Y.; Yeh, I.-H.; Liao, H.-Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. In *Computer Vision—ECCV 2024*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2024; pp. 1–21. [CrossRef]
39. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*; Curran Associates: Red Hook, NY, USA, 2024; pp. 107984–108011. [CrossRef]
40. Ultralytics. YOLO11. Available online: https://docs.ultralytics.com/models/yolo11/ (accessed on 15 October 2026).
41. Ultralytics. YOLO26. Available online: https://docs.ultralytics.com/models/yolo26/ (accessed on 15 October 2026).
42. Sapkota, R.; Cheppally, R.H.; Sharda, A.; Karkee, M. YOLO26: Key Architectural Enhancements and Performance Benchmarking for Real-Time Object Detection. *arXiv* **2025**, arXiv:2509.25164. [CrossRef]
43. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Computer Vision—ECCV 2014*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2014; pp. 740–755. [CrossRef]
44. Tamilarasi, T.; Muthulakshmi, P.; Ashtiani, S.-H.M. Improved YOLO-Based Real-Time Brinjal Detection Algorithm for Vision Modules in Harvesting Robots. *Eng. Res. Express* **2025**, *7*, 035234. [CrossRef]
45. Divyanth, L.G.; Khanal, S.R.; Paudel, A.; Mattupalli, C.; Karkee, M. Efficient Detection of Eyes on Potato Tubers Using Deep-Learning for Robotic High-Throughput Sampling. *Front. Plant Sci.* **2024**, *15*, 1512632. [CrossRef] [PubMed]
46. Kaufman, S.; Rosset, S.; Perlich, C.; Stitelman, O. Leakage in Data Mining: Formulation, Detection, and Avoidance. *ACM Trans. Knowl. Discov. Data* **2012**, *6*, 15. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.