agronomy

MDPI

Article

From Bench to Conveyor: A Four-Generation Deployment Study of an Automated Grading System for Pine Mushroom

Jeon-Seong Kang [1], Jonghun Jeong [1], Beom-Joon Park [1], Junwon Yoon [1], Ha-Yoon Song [1], Hyun-Joon Chung [1,*] and Dong-Geol Choi [2]

[1] AI Robotics R&D Division, Korea Institute of Robotics & Technology Convergence (KIRO), 4 Jagok-ro 7-gil, Gangnam-gu, Seoul 06372, Republic of Korea; kjs2605@kiro.re.kr (J.-S.K.); jonghn01@kiro.re.kr (J.J.); parkbj26@kiro.re.kr (B.-J.P.); jyoon@kiro.re.kr (J.Y.); hayoon@kiro.re.kr (H.-Y.S.)
[2] Department of Information and Communication Engineering, Hanbat National University, 125 Dongseo-daero, Yuseong-gu, Daejeon 34158, Republic of Korea; dgchoi@hanbat.ac.kr
* Correspondence: hjchung@kiro.re.kr

Abstract

This is a longitudinal deployment study and not a new model. We report four development cycles (2025–2026) of a grading system for pine mushroom (*Tricholoma matsutake*), the highest-value non-timber forest product traded in the Republic of Korea, as it moved from a bench dataset to a sorting conveyor, and we separate image-based agreement with recorded inspector grades, engineering measurements, functional field demonstrations, and a belt-level experiment. All figures were re-derived by re-running the delivered models on their test partitions. Generation 1 combined a dual-view CNN ensemble with a morphometric rule (96.70% agreement, 8.5 FPS). Generation 2 reached 30 FPS by removing a backbone, because three networks on one GPU at batch 1 contended rather than overlapped. Generation 3 moved to three cameras; on new data, the earlier architecture fell to 70.0%, and, in a first field trial, frozen–thawed specimens defeated segmentation while the classifiers kept emitting grades. Generation 4 agreed with the inspector on 60 of 61 specimens (98.36%; 95% CI 91.3–99.7%) of a cleaned internal split that was also used for checkpoint selection, at 55 FPS inference throughput; a second trial demonstrated PLC-driven sorting, and a belt-level experiment that fed the 61 test specimens through a laboratory conveyor sorter sent 60 (98.36%) to the lane of their inspector grade—the same result as on the stored images. The re-analysis showed that the Generation 3 pipeline agreed on only 32.8% of the later Generation 4 collection; that a speed optimisation decorrelated the stipe thickness estimate ($r = -0.32$); and that, in Generation 4, both classifiers yielded identical predictions, one had collapsed onto a single camera, and the rule alone agreed on 52.5%. The accuracies are not a progress curve, the indicators are unvalidated image-derived proxies, and the system is a single-site prototype; we state the evaluations that a commercial claim would require.

Keywords: pine mushroom; matsutake; automated grading; non-timber forest products; deployment study; domain gap; instance segmentation; multi-view fusion; attention fusion; TensorRT; real-time inference; object tracking; human-in-the-loop interface

Check for updates

Academic Editor: Xiangjun Zou

Received: 29 September 2026
Revised: 6 November 2026
Accepted: 12 November 2026
Published: 17 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.

1. Introduction

Pine mushroom (*Tricholoma matsutake* S. Ito & S. Imai, Singer) is an ectomycorrhizal fungus fruiting in association with the roots of Japanese red pine (*Pinus densiflora*). Despite decades of effort, it has resisted commercial cultivation, so supply comes entirely from wild collection in a short autumn season [1]. In the Republic of Korea, the harvest is

organised through village-level collection and sold through cooperative centres, where it is one of the main cash incomes of mountain communities [2]. The price per kilogramme differs by an order of magnitude between the top and bottom grades. Grade is assigned based on the external morphology—whether the veil is intact and the cap still closed, how thick the stipe is, and whether the specimen is damaged—and it is assigned by eye, by trained inspectors, at the collection centre. Manual throughput is limited, the product loses freshness while waiting, and judgement is subjective: two inspectors may disagree on a borderline specimen, and the collector receives no objective record justifying the grade that set the payment.

Automation is an obvious application of modern computer vision. Deep learning for produce quality assessment represents a mature body of literature [3,4]; the real-time grading of fruit on a line has been demonstrated with standard backbones [5]; and mushrooms specifically have been graded by size and appearance for several cultivated species [6–9]. What the literature rarely reports is the distance between a model that grades a held-out image set well and a system that grades physical objects moving on a belt—a distance that surveys of deployed machine learning instruments identify as the dominant source of failure in practice and one that is rarely visible in a single evaluation [10,11]. This paper focuses on this distance. Over four development cycles between January 2025 and May 2026, we built, deployed, broke, and rebuilt a grading system for pine mushroom on the sorting conveyor of a robot-based collection and distribution centre for non-timber forest products. We report the whole trajectory, including two field trials—one that failed and one that succeeded—because the failures were more instructive than the final accuracy figure.

Figure 1 shows the conveyor sorter on which the belt-level experiment of Section 3.4.2 was carried out: an imaging tunnel over the infeed belt, followed by flip-bar lanes that divert each specimen to the chute for its grade.

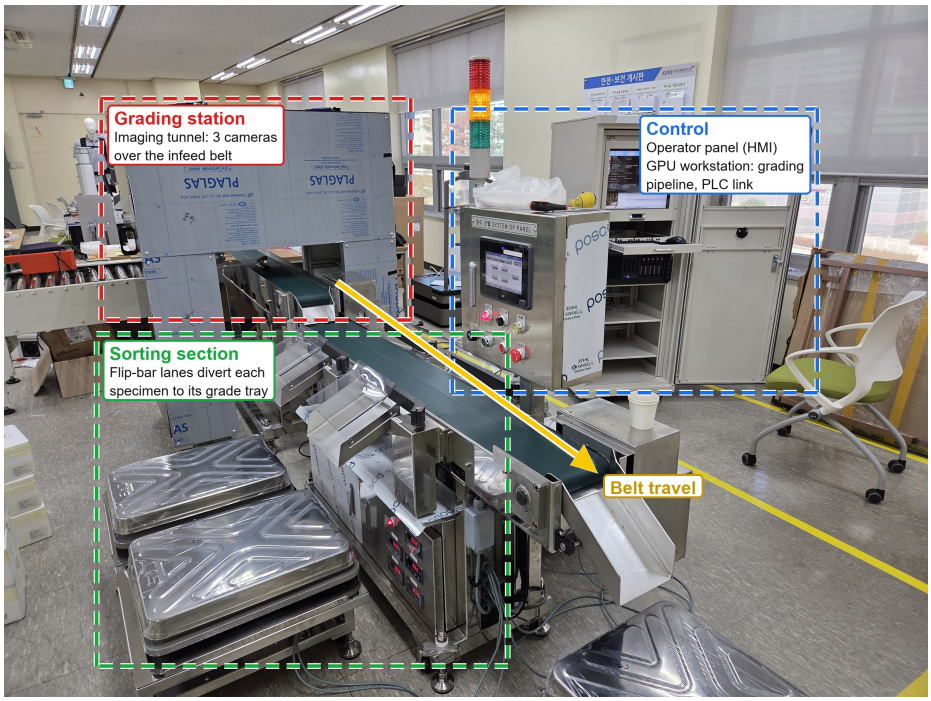


Figure 1. The laboratory conveyor sorter used for the belt-level experiment (Section 3.4.2). The red dashed box marks the grading station: the imaging tunnel housing the three cameras over the infeed belt. The blue dashed box marks the control units: the operator panel (HMI touchscreen, start/stop and emergency stop) and the GPU workstation that runs the pipeline of Figure 2 and writes each grade to the PLC. The green dashed box marks the sorting section: flip-bar lanes that, on the PLC's

command, divert each specimen into the tray for its grade. The yellow arrow gives the direction of belt travel; specimens not diverted leave by the end-of-belt chute. The Korean words on the operator panel plate mean 'pine mushroom sorting', and the board in the background is the institute's safety and health notice board.

Three properties of the task shaped the design throughout. First, the official grade specification issued by the Korea Forest Service is written in physical terms—the stipe thickness and degree of cap opening—so a system whose output cannot be traced to these quantities is difficult to certify and more difficult to dispute when a collector challenges a payment. Second, the boundaries between grades are commercial parameters rather than physical constants: they are renegotiated between the centre and the producers' association, they vary by region, and inspectors deliberately change them with the season's yield, passing a marginally small Grade 2 specimen as Grade 1 in a poor year. Third, the conveyor imposes a hard real-time budget. Every generation of the system therefore combined a learned appearance branch with an explicit, operator-editable morphometric rule, and every generation had to earn its accuracy inside a latency budget.

Figure 2 shows the deployed fourth-generation pipeline.

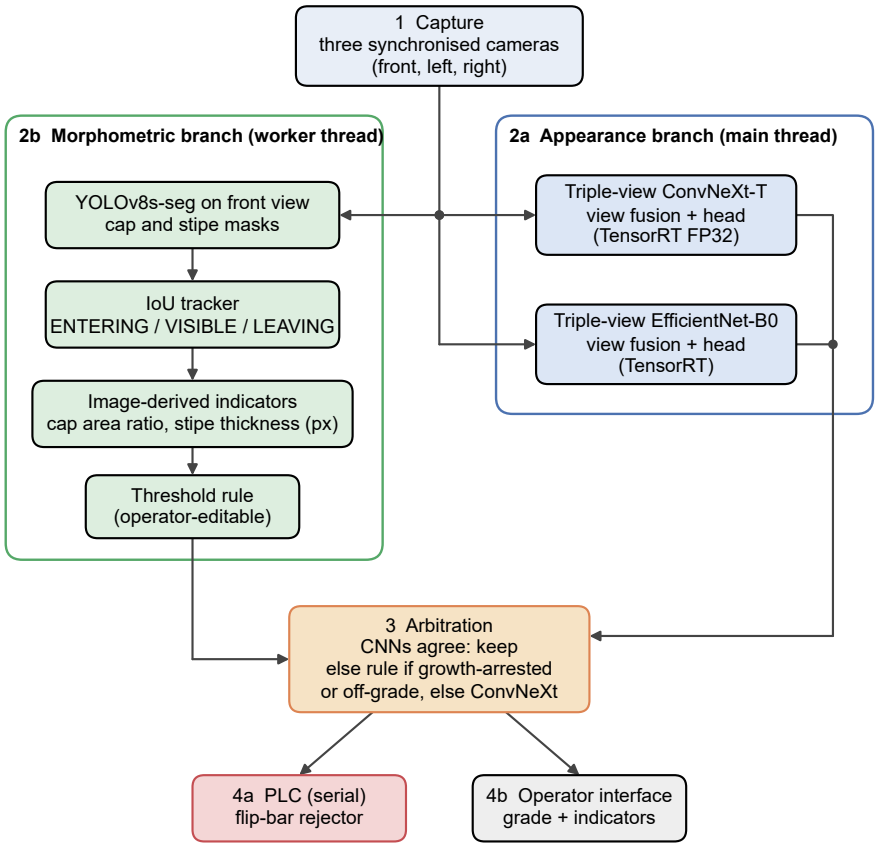


Figure 2. The deployed fourth-generation pipeline, read top to bottom in four numbered stages. (**1**) Three synchronised cameras. (**2a**) Appearance branch: two triple-view classifiers, each fusing the three views and predicting a grade. (**2b**) Morphometric branch, on a separate thread: segmentation of the front view, tracking of each specimen across frames, image-derived indicators, and an operator-editable threshold rule. (**3**) Arbitration of the three predictions (Section 2.8). (**4a**) The grade is sent to the flip-bar rejector over a serial PLC link and (**4b**) shown, with the indicators, to the operator.

Our contribution is not a novel architecture. It is a documented account, with measurements at each step, of how four successive deployment contexts affected a grading system that was accurate on the bench—together with the specific engineering responses

that worked. We are explicit throughout that the four headline accuracies (96.70%, 93.41%, 95.45%, 98.36%) were measured on four different test sets and do not form a progress curve; the reasons that they are not comparable are themselves among the findings.

Four types of evidence appear in this paper, and they support different claims (Table 1). Agreement with recorded inspector grades on stored image sets is the only evidence about grading; it is internal to each generation's collection except for two re-evaluations on later collections. Engineering measurements—latency, numerical precision, segmentation availability, and agreement between implementations—are about the pipeline and not about grades. Field trials demonstrated that the chain from camera to rejector operates; they measured nothing about grading or sorting accuracy. A belt-level experiment on a laboratory sorter counted where the test specimens physically went (Section 3.4.2). Each result below is labelled with the type of evidence that it rests on.

Table 1. Types of evidence in this study and what each can and cannot support.

Evidence	Where	Supports	Does Not Support
Image-based agreement, own split	G1–G4 test splits (91, 22, 61 specimens), all re-run for this revision	Agreement with one recorded inspector label on curated stored images	Accuracy on the line, on later harvests, or against a consensus grade
Image-based agreement, later collection	G2-type model on raw G3 data (40); G3 pipeline on G4 data (61)	How much an internal split figure overstates generalisation	Accuracy of the G4 system itself on later data
Engineering measurement	Latency, FP16/FP32, mask availability, indicator agreement between implementations	Properties of the pipeline	Grading accuracy; physical correctness of indicators
Field demonstration	Trials on 18 March and 13 May 2026	That capture, grading, PLC signalling, and rejection operate end to end	Belt-level accuracy, correct lane rate, missed or false triggers, actuation timing
Belt-level experiment	Laboratory conveyor sorter of Figure 1, the 61 G4 test specimens (Section 3.4.2)	That the physical line reproduces the image-level grades: correct-lane rate and missed specimens	Performance on unseen specimens, at the collection centre's throughput, on later harvests, or against a consensus grade; the timing margin and false trigger rate

1.1. Background and Related Work

1.1.1. Grade Specification

The grades follow the specification for non-timber forest products announced by the Korea Forest Service, as operationalised with the collection centre. The specification is stated in two physical quantities—the stipe length and the fraction of the cap that has opened—plus a list of defects (Figure 3, Table 2). Two premium grades have a closed or almost closed cap and are separated by stipe length; a growth-arrested category and an opened-cap category are both traded as third grade but are morphologically distinct and are kept separate by the models; and an off-grade category covers deformed, damaged, insect-eaten, water-soaked, or fully opened specimens. Two features of this standard shaped what follows. It is length-primary, which Section 3.1.1 shows is not what the measured data support, and its cap criterion is a fraction of opening, which is what the cap area ratio of Section 2.6 approximates.

1.1.2. Why a Hybrid Design Throughout

Every generation kept a learned branch and an explicit measurement branch, for three reasons that never changed. The first was auditability: a grade sets a price, so, when it is

disputed, the operator must be able to show the quantity that produced it—"cap area ratio 0.39, stipe thickness 26 px, therefore Grade 3" can be checked, while a softmax score cannot. The second was adjustability: exposing the boundaries as editable thresholds is what allows an automated system to follow the seasonal flexing that inspectors already practice, rather than challenging it. The third was coverage of residual cues: soil contamination, bruising, insect damage, and the texture change of an opening veil are learned far more easily than formalised. Section 3.1.3 quantifies the last point directly: on the first-generation test set, the physical rule alone reached 82.42%, with thresholds fitted on the same set, where the appearance branch reached 95.60%.

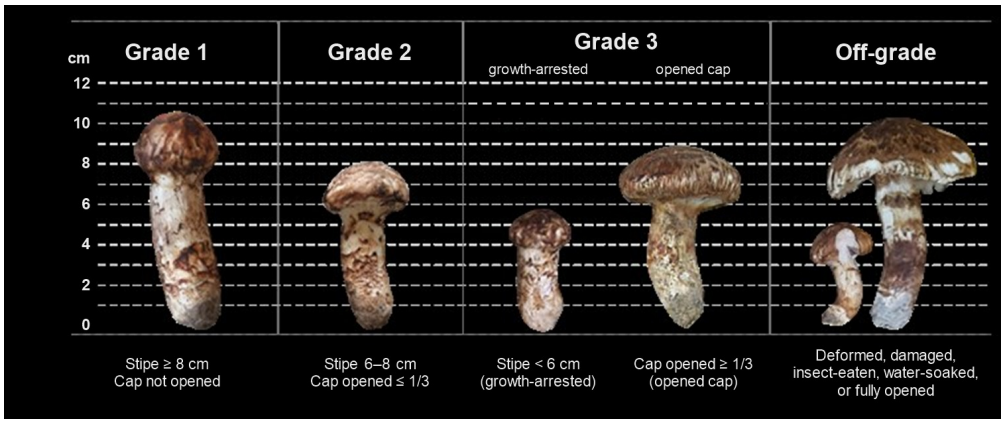


Figure 3. The grade standard as illustrated for inspectors, with the reference photographs against a 1 cm grid (labelled every 2 cm). Left to right: Grade 1 (stipe $\geq$ 8 cm, cap not opened); Grade 2 (stipe 6–8 cm, cap opened by at most one third); Grade 3, growth-arrested (stipe $<$ 6 cm) and opened cap (cap opened by one third or more); off-grade (deformed, damaged, insect-eaten, water-soaked, or fully opened). Labels are translated from the Korean original.

Table 2. Grade categories, their criteria in the official standard, and the code index used in each generation. The labels translate the Korean designations 1-, 2-, and 3-deunggeup (first, second, and third grade) and deungoe-pum (off-grade); the two third-grade categories are saengjang-jeongji-pum (growth-arrested) and gaesan-pum (opened cap).

Grade	Criterion in the Standard	G1–G2	G3–G4
Grade 1	Stipe $\geq$ 8 cm; cap not opened	0	1
Grade 2	Stipe 6–8 cm; cap opened $\leq$1/3	1	2
Grade 3 (growth-arrested)	Stipe $<$ 6 cm, growth arrested	2	4
Grade 3 (opened cap)	Cap opened $\geq$1/3	3	3
Off-grade	Deformed, damaged, insect-eaten, water-soaked, or fully opened	4	5

1.1.3. Related Work

Deep learning for agricultural quality inspection is well established [3,4], and instance segmentation is the standard tool when a task requires part-level measurement rather than whole-object classification [12,13]. Mushroom inspection follows the same pattern. A review of computer vision in the mushroom industry finds most work concentrated on species identification, harvesting, and size grading in cultivated species [6]; representative examples grade dried shiitake by appearance with a VGG variant [8], size-grade antler mushrooms by combining YOLOv5 detection with PSPNet segmentation [7], and size-grade *Agaricus bisporus* on a conveyor from video at over a thousand specimens per minute [9]. Cultivated mushrooms are uniform in appearance and graded chiefly by cap

diameter; a wild-collected species with a written multi-criterion standard, graded partly by cap opening and partly by defects, has, to our knowledge, not been addressed. Multi-view fusion resolves the pose ambiguity of single-camera inspection in three-dimensional objects [14], and attention-based fusion allows a model to weight views per sample rather than uniformly [15]. Simple IoU association between successive detections is sufficient for tracking rigid objects moving predictably in a fixed camera [16], which is the case on a belt. Work on quantisation and graph compilation reports that the inference cost can usually be cut substantially at a negligible accuracy cost [17,18]; Section 3.2.2 qualifies this expectation for LayerNorm-based architectures.

Domain shift between training and deployment conditions is a recognised failure mode of learned vision systems in general [19,20]—classifiers lose accuracy even on a fresh test set drawn by the same protocol as the original [21]—and of agricultural vision in particular, where the lighting, season, cultivar, handling, and imaging device all vary between the development collection and the field [22,23]. Small training sets amplify the effect [22]. Reports of this kind usually describe a gradual loss of accuracy; the visible failure of one branch of a pipeline while the other fails silently is less often documented, and it is the case that Section 3.3.5 reports. Surveys of deployed machine learning instruments place monitoring and the detection of such silent failures among the least addressed stages of deployment [10,11].

Two earlier reports from our group cover single-generation snapshots of this work: a Dino-SAM2 and ResNet-50 ensemble that reached 90% when a mask could be produced [24] and a two-camera hybrid of YOLOv8n with dual-view ConvNeXt and EfficientNet under majority voting that reached 94.51% on 466 pairs—6.6 points above the best single view and 19.78 above the morphometric rule alone [25]. Neither reports what happened when the system met the conveyor.

What is uncommon in this literature is a longitudinal account. Published systems are typically evaluated once, on data collected under the conditions in which they were developed. Our four cycles allowed us to observe the same architecture family evaluated under four successive definitions of the task, and the differences between these evaluations are the substance of this paper.

1.2. Key Contributions

- A four-generation deployment record of an automated grader for a high-value non-timber forest product, including a field trial that failed, in which every reported figure has been re-derived from the delivered models and test partitions, so that per-specimen, per-class, and component-level analyses are available for all generations.
- Evidence, from two re-evaluations on later collections, of how far internal split agreement overstates generalisation in this setting: 90%-plus fell to 70.0% (G2-type architecture on G3 data) and 95.45% fell to 32.8% (G3 pipeline on G4 data).
- Quantified failure modes and their observability: under the frozen–thawed shift, the segmentation failure was visible (no usable mask for 36% of specimens) while the larger classifier failure was not; a speed optimisation silently decorrelated the thickness estimate; three networks on one GPU at batch 1 contended rather than overlapped; and FP16 behaviour was architecture-dependent.
- An account of what the ensembles, the view fusion module, and the morphometric rule actually contributed and of the validation that a commercial claim would require—an independent later-season collection, an uncurated belt run, a multi-inspector study, caliper validation, and operating-level sorting metrics—none of which we claim here.

2. Materials and Methods

2.1. The Four Generations

Table 3 summarises the trajectory. Each generation was a separate development cycle with its own objective, and each changed the system in response to a limitation exposed by the previous one.

Figure 4 reproduces the pipeline diagrams of the first three generations; the fourth is given in Figure 2.

Table 3. The four development generations. Accuracies are not comparable across rows: each was measured on the generation's own test set (Section 2.4). Accuracy is given as the exact count with its 95% Wilson score interval.

Gen.	Period	Objective	Principal Components	Correct/*n*	Acc. (%) [95% CI]	FPS
G1	March–July 2025	Define morphometric indicators; rule-based grading	YOLOv12-s seg; dual-view ConvNeXt-T + ResNet-50; hard majority vote; 2 cameras	88/91	96.70 [90.8, 98.9]	8.5
G2	May–October 2025	Reach 30 FPS; operator threshold interface	TensorRT compilation; YOLOv8n + dual-view ConvNeXt-T only; Gradio UI; 2 cameras	85/91	93.41 [86.4, 96.9]	30.64
G3	December 2025–March 2026	Three-camera capture; first field trial	Triple-network ConvNeXt-S + MobileNetV3; YOLOv8n-seg; 3 cameras	21/22	95.45 [78.2, 99.2]	30.5
G4	March–May 2026	Frozen–thawed domain gap; end-to-end operation with physical sorting	Camera–attention fusion; ConvNeXt + EfficientNet + morph; YOLOv8s-seg; IoU tracking; PLC rejector; 3 cameras	60/61	98.36 [91.3, 99.7]	55

2.2. Re-Analysis of the Delivered Models

Each generation was delivered as source code, trained checkpoints, and the test partition on which it was evaluated. For this revision, we re-ran the delivered checkpoints of all four generations on their delivered test partitions with the delivered pre-processing, segmentation, indicator, and decision code, using PyTorch 1.13.1 on an NVIDIA GeForce RTX 3060 Ti in place of the TensorRT engines built from the same checkpoints (for G3, the engines were verified to match PyTorch on every test prediction; Section 3.2.2). Every headline count was reproduced exactly: G1 88/91 (ConvNeXt 87, ResNet-50 86, rule 75), G2 85/91, G3 21/22, and G4 60/61. The re-run yields per-specimen predictions for every component, from which this revision adds per-class and rank-aware metrics for all generations, paired McNemar tests, component and view ablations, segmentation availability, and indicator agreement between implementations. The same procedure applied the delivered G3 pipeline to the later G4 test partition (Section 3.3.6). Two corrections to the submitted description followed from the re-run: the delivered G3 pipeline used YOLOv8n-seg, not YOLO26m, and the G4 rule and arbitration assign the thin-stipe branch to the growth-arrested category, not to off-grade (Equations (6) and (7)). The scripts and per-specimen outputs are available from the corresponding author.

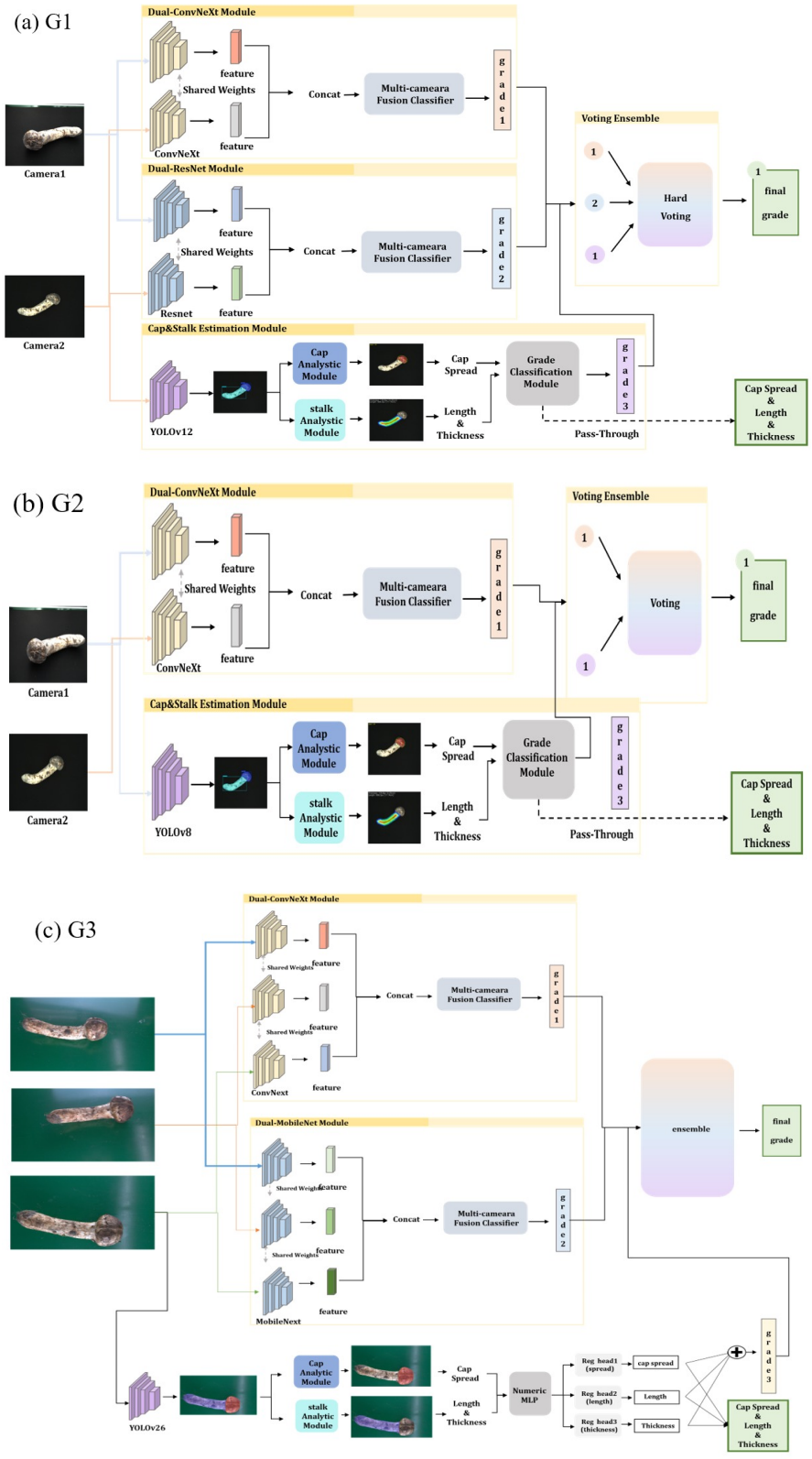


Figure 4. Pipeline diagrams of the first three generations. (**a**) G1: two cameras feed a dual-view ConvNeXt module, a dual-view ResNet module, and a cap/stalk estimation module (YOLOv12 segmentation, cap and stalk analytic modules, threshold rule); a hard-voting ensemble emits the grade, and the measurements pass through. (**b**) G2: the ResNet module is removed; the dual-view ConvNeXt and the YOLOv8 segmentation branch run as compiled engines and are reconciled by the two-branch rule. (**c**) G3 as designed at the interim stage: three cameras feed triple-view ConvNeXt and MobileNetV3 modules; YOLO26 segments the front view, and a numeric MLP with three regression

heads refines the cap spread, length, and thickness estimates before the ensemble. The delivered G3 pipeline replaced YOLO26 with YOLOv8n-seg and applied the rule to the unrefined indicators (Section 2.2). In all panels, thin coloured lines trace the image of each camera to the modules that use it; the beige, blue and purple blocks are the ConvNeXt, ResNet or MobileNet, and YOLO backbones; the vertical boxes labelled grade 1, grade 2 and grade 3 are the grades emitted by the three branches; the numbered circles show these grades entering the vote; green boxes are the system outputs; and the dashed arrow is the pass-through of the measured cap spread, length and thickness.

2.3. Imaging Rigs

Generations 1–2 (two cameras). Specimens were imaged on the sorting conveyor by two industrial colour cameras producing synchronised 1280 × 1024 px RGB pairs (Figure 5). Camera 2 viewed the specimen against a matte black background under diffuse illumination and was the metric view; Camera 1 provided the complementary viewpoint used only by the appearance branch. The optical scale was calibrated once at the working distance, giving $s = 60\ \mathrm{px\,cm^{-1}}$, so that indicators could be reported in centimetres.



Figure 5. A synchronised pair from the first-generation two-camera rig. (**Left**) Camera 1, with the conveyor belt edge visible in-frame. (**Right**) Camera 2, against a matte black background; in the first two generations, all physical measurements were taken from this view.

Generations 3–4 (three cameras). Three Basler industrial cameras (front, left, right) were mounted above the conveyor, producing synchronised 1920 × 1080 px triples (Figure 6).

The exposure, gain, and frame rate are set per camera through the operator interface, and cameras are addressed by IP or serial number.

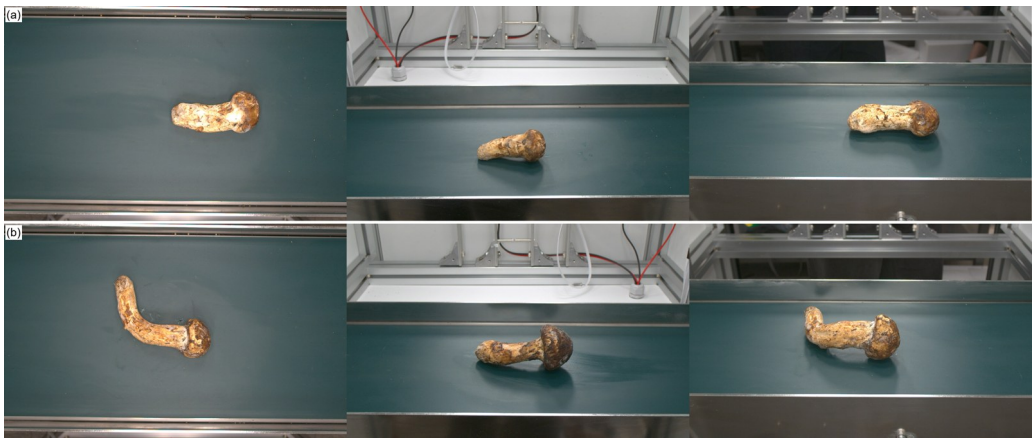


Figure 6. Three-camera triples from the G3–G4 rig (left, front, right). (**a**) A consistent triple. (**b**) A cross-camera pair mismatch: the three frames were recorded as one specimen but show different specimens; such triples were removed in G4 (Section 2.4).

The third viewpoint was added because a single metric view cannot disambiguate a cap that is opening asymmetrically and because a specimen's orientation on the belt is uncontrolled. No global pixel-to-centimetre calibration was carried into these generations; the fourth-generation rule operates on the pixel-domain thickness and on a dimensionless area ratio (Section 2.7).

2.4. Datasets

Table 4 gives the four datasets. They differ in camera count, resolution, size, and—critically—curation policy, which is why the accuracies of Table 3 cannot be compared directly.

Table 4. Datasets by generation. "Pairs"/"triples" count physical specimens; each contributes 2 or 3 images.

Gen.	Cameras	Resolution	Train	Test	Curation
G1–G2	2	1280 × 1024	373	91	None reported
G3 (raw)	3	1920 × 1080	150	40	Cross-camera pair repair
G3 (filtered)	3	1920 × 1080	78	22	+Outlier removal by stipe length
G4	3	1920 × 1080	235	61	+Pair mismatch, label error, and duplicate removal; specimen-level stratified split

Collection and partitioning. In every generation, one specimen contributes one image pair or triple, and the train–test split was drawn at the specimen level by grade-stratified random sampling within the generation's pooled collection; no split was performed by collection session or batch. Specimens from the same session therefore appear on both sides of every partition. In G1–G2, the 91 test specimens come from three capture sessions (29 June, 22 and 23 October 2024) that each contribute to four or five of the five grades, so the test split shares sessions—and any lighting, handling, and storage conditions that these sessions had in common—with the training split. G3 and G4 file names carry no session information. In G4, a specimen-level split was enforced only after duplicates were merged: an earlier image-level split had placed re-photographed specimens on both sides and scored 100%. Duplicate detection was manual, so residual duplicates between the G4 training and test splits cannot be excluded; the training split was not retained, so it cannot be checked, but, within the test split, the most similar pairs of images were inspected and were different specimens.

Exclusions. Table 5 reports what was removed before evaluation. In G3, 18 of 40 raw test specimens and 72 of 150 raw training specimens were removed via the outlier criteria listed below; the per-grade composition of the raw partitions was not retained, so only the composition after filtering can be reported. In G4, 132 of 428 specimens were removed before the split; the per-grade counts are known, but the record does not itemise how many were removed for each of the three reasons. In both generations, the reported accuracy therefore applies to a cleaned internal test set, not to the specimens arriving at the centre.

Storage conditions. G4 added frozen-and-thawed specimens to the collection, but the storage conditions of each specimen were not recorded, so the numbers of fresh and frozen–thawed specimens in the G4 training and test splits, and the accuracy by condition, cannot be reported. Section 3.3.6 gives the closest available evidence: the G3 pipeline, trained only on fresh material, evaluated on the G4 test split.

Table 5. Specimens excluded before evaluation by grade, where known. G4 counts are image counts from the cleaning record divided by three cameras.

Gen.	Stage	Gr. 1	Gr. 2	Gr. 3 Opened	Gr. 3 Arrested	Off	Total
G3	Raw train/test			Not retained			150/40
G3	After filtering, train	9	18	11	17	23	78
G3	After filtering, test	4	5	3	3	7	22
G3	Removed (train/test)			Not retained by grade			72/18
G4	Before cleaning	77	65	83	118	85	428
G4	After cleaning	53	45	62	67	69	296
G4	Removed (mismatch, label, duplicate)	24	20	21	51	16	132

Ground truth. In every generation, the reference grade for a specimen is the grade assigned by the collection centre's inspectors at the time of collection, recorded as a single label: an inspector graded each specimen at intake against the standard as operated at the centre (Table 2), and the specimen was then imaged. There was no blind re-grading, no second inspector, and no adjudication step. No specimen was independently re-graded by a second inspector, so the labels carry the inter-inspector disagreement that Section 1 gives as part of the motivation, and the agreement rates reported below are agreement with one recorded inspector judgement rather than with a consensus. Where a label was contradicted by visual review (G4) or by the standard's own length criteria (G3), the specimen was removed rather than re-adjudicated. In G4, these exclusions were applied to the whole collection before the final specimen-level split was drawn so as to construct a technically consistent training and evaluation collection; they are not a correction of the inspectors' labels and should not be read as one. In G3, the outlier filter was applied to an existing 150/40 partition, reducing both sides of it, so the filtered G3 test set is a subset of the raw one. The consequences of this are taken up in Section 4.9, and Section 4.8 states the inter-rater study that would bound them.

The G1–G2 test partition contains 91 specimens distributed 10/17/31/20/13 across the five grades. Segmentation masks for this generation were produced semi-automatically: hand-drawn bounding boxes were supplied as visual prompts to YOLOE [26], and the predicted masks were denoised and simplified into two-class polygons (`cap`, `stipe`).

The G3 dataset comprises 570 images—190 specimens × 3 cameras—and required manual re-labelling to repair file name mismatches between Camera 3 and Cameras 1–2 for the opened-cap class. It was roughly 60% smaller in training specimens than G1–G2, which proved consequential (Section 3.3.1). It was subsequently filtered against explicit outlier criteria—Grade 1 specimens with stipes under 8 cm, Grade 2 under 6 cm, growth-arrested specimens over 6 cm, opened-cap specimens with cap damage or stipe deformity, and off-grade specimens under 6 cm or plausibly third grade—leaving 78 training and 22 test specimens.

The G4 collection began at 1284 images (428 specimens) and was cleaned based on three criteria (Figure 7). Specimens missing an image from any of the three cameras, or whose three frames showed different specimens (Figure 6b), were removed entirely to preserve triple consistency. Specimens whose grade label disagreed with visual inspection were removed. Finally, the same physical specimen had, in many cases, been photographed more than once at different positions on the belt; a random image-level split had placed such duplicates on both sides of the partition, and a preliminary model trained on this split scored 100% on its test set. Duplicates were merged and the split was re-performed at the specimen level. The result is 296 specimens (888 images), divided 235/61 by a stratified 8:2 split (Table 6). The growth-arrested class lost the most (354 to 201 images), which is where

the duplicates were concentrated. Frozen-and-thawed specimens were collected and added at this stage for the reason given in Section 3.3.5.

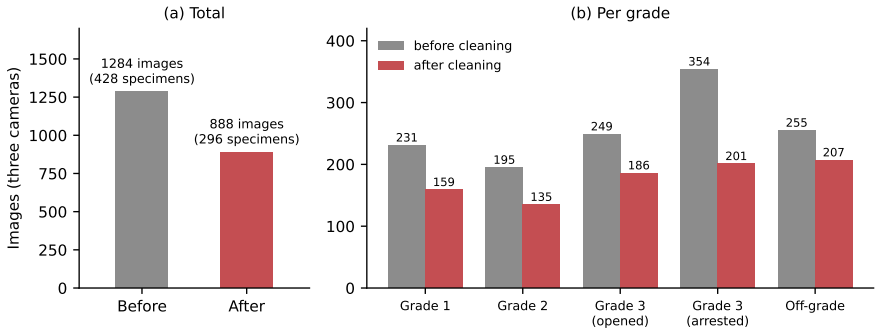


Figure 7. G4 dataset cleaning. (**a**) Total images before and after removal of cross-camera mismatches, mislabelled specimens, and duplicate specimens. (**b**) The same per grade; the growth-arrested class lost 43% of its images.

Table 6. G4 dataset by grade after cleaning. Specimen counts are the report's image counts divided by three cameras; the test counts were read from the delivered test partition.

	Grade 1	Grade 2	Gr. 3 (Opened)	Gr. 3 (Arrested)	Off-Grade	Total
Specimens	53	45	62	67	69	296
Train	42	36	49	53	55	235
Test	11	9	13	14	14	61

2.5. Appearance Branch

2.5.1. Generations 1–2: Dual-View Concatenation

Both views were resized to 384 × 384 px in G1 and 416 × 416 px in G2 and normalised with ImageNet statistics. A single ConvNeXt-Tiny backbone [27] pre-trained on ImageNet-22k [28], with its head removed, was applied to both views with shared weights, and the two 768-dimensional descriptors were concatenated and mapped to five grades by one linear layer:

$$\mathbf{z}_i = \Phi(\mathbf{x}_i) \in \mathbb{R}^{768}, \qquad \hat{\mathbf{p}} = W[\mathbf{z}_1; \mathbf{z}_2] + \mathbf{b}, \qquad W \in \mathbb{R}^{5\times 1536}. \tag{1}$$

Fusing before a single linear layer, rather than averaging two softmax outputs, allows the classifier to learn cross-view interactions—a wide cap silhouette in one view is benign when the other shows an intact veil and diagnostic when it does not. A ResNet-50 [29] variant with an identical structure ($2048 + 2048 \to 5$) served as the second ensemble member in G1.

2.5.2. Generations 3–4: Triple-View Learned Attention Fusion

With three cameras, uniform fusion is wasteful: for any given specimen, some viewpoints are more informative than others, and this varies with how the specimen lies on the belt. G3 extended the concatenation scheme to three inputs; G4 replaced it with a learned attention fusion that weighted the views per sample. Let $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3 \in \mathbb{R}^d$ be the per-camera backbone descriptors. A small gating network produces one logit per camera from their concatenation, and a temperature-scaled softmax converts these into convex weights:

$$\boldsymbol{\ell} = W_2\, \sigma(W_1[\mathbf{f}_1; \mathbf{f}_2; \mathbf{f}_3]) \in \mathbb{R}^3, \qquad W_1 \in \mathbb{R}^{128\times 3d},\ W_2 \in \mathbb{R}^{3\times 128}, \tag{2}$$

$$\boldsymbol{\alpha} = \mathrm{softmax}\big(\boldsymbol{\ell} / \max(\tau, 0.1)\big), \qquad \mathbf{f}_{\text{fused}} = \sum_{i=1}^{3} \alpha_i \mathbf{f}_i, \tag{3}$$

where σ is ReLU with dropout 0.3, and τ is a learned scalar temperature initialised at 3.0. The fused descriptor passes through a $d \to 512 \to 5$ head with batch normalisation, ReLU, and dropout.

Two details were intended to aid stability. The gating weights are initialised with Xavier gain 0.1 and zero bias, and the temperature starts high, so that, early in training, $\boldsymbol{\alpha}$ is close to uniform; the intent was that the network should not collapse onto a single camera before the backbone has learned anything useful, after which the temperature is free to sharpen. This describes the ConvNeXt classifier. The EfficientNet classifier of G4 uses a simpler variant of the same module, without the temperature or the initialisation: a two-layer gating network of width d followed directly by a softmax and a $d \to 512 \to 5$ head with dropout 0.4. Because $\boldsymbol{\alpha}$ is returned alongside the logits, the fusion is, in principle, inspectable. Section 3.4.1 shows that, in the delivered G4 models, it carries little information: the ConvNeXt weights are almost constant across specimens, and the EfficientNet weights collapsed onto one camera.

Inputs are 384×384 px in G3–G4, with training augmentation consisting of horizontal flip ($p = 0.5$), vertical flip ($p = 0.3$), and rotation ($\pm 15°$). Backbones are shared across the three cameras (Figure 8). The deployed G4 appearance branch is an ensemble of a triple-view ConvNeXt and a triple-view EfficientNet [30], each with its own camera–attention fusion module and head, reconciled with the morphometric branch by the arbitration rule of Section 2.8. MobileNetV3 [31], which had served in the G3 ensemble, was evaluated in the G4 backbone screening (Section 3.4) but was not carried into the deployed configuration. All classifiers were trained with cross-entropy loss from ImageNet-22k initialisation with the full backbone unfrozen—the regime that Section 3.1.2 shows to be necessary.

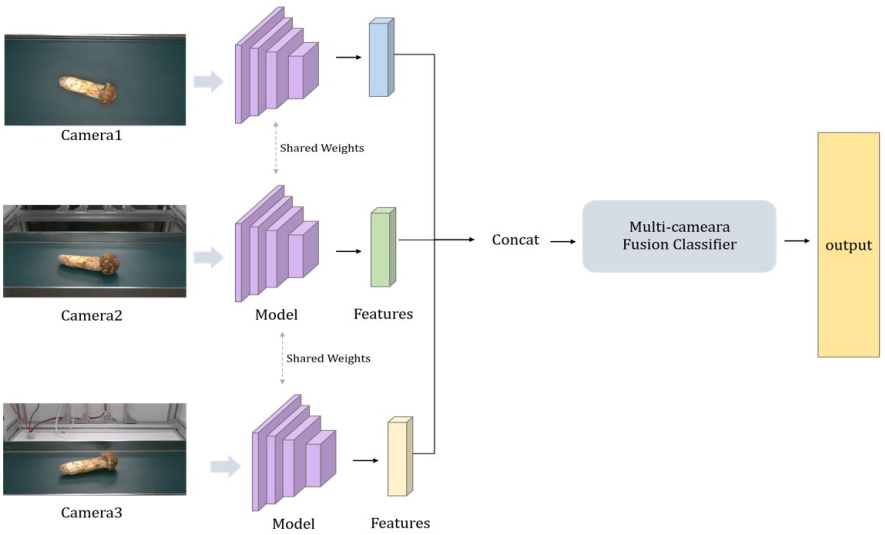


Figure 8. Triple-view classifier. The three camera images pass through one backbone with shared weights, compiled as a batch-1 TensorRT engine; the three feature vectors are concatenated, fused by the camera–attention module of Equations (2) and (3) (the multi-camera fusion classifier), and classified by a single head.

Training settings and model selection. The G3 and G4 checkpoints store their training configuration, reproduced in Table 7. During development, each generation's models were trained many times as the data, splits, and settings changed; we call these development runs, and examples are the preliminary model trained on an image-level split (Section 2.4), the backbone screenings, and the intermediate checkpoints of Section 3.3.4. Each delivered model is the product of one final training run in two stages—the head alone and then the whole network, with the backbone parameter group at one tenth of the head group's learn-

ing rate in the stored optimiser state—with an Adam family optimiser ($\beta = (0.9, 0.999)$), a decaying learning rate schedule, batch size 4, the augmentation given above, and label smoothing 0.1. The checkpoints also record how the reported model was chosen: the stored epoch is the one with the highest accuracy on the evaluation partition, and this accuracy equals the reported test accuracy (60/61 for both G4 classifiers, 19/22 for both G3 classifiers). The collections were split 8:2 into training and test partitions only; there was no separate validation partition, so the test partition was used to select the checkpoint. The G3 and G4 accuracies are therefore optimistically biased by model selection, by an amount that we cannot estimate from the available data. The G1–G2 checkpoints carry no metadata; their selection procedure was not recorded. No delivered configuration was retrained with different random seeds under otherwise identical data and settings (seed repeats), so training variability is unknown; the development runs differ in data or settings and cannot stand in for seed repeats. The rule thresholds were tuned on test data in G1 (Section 2.7).

Table 7. Training settings recovered from the stored G3 and G4 checkpoints. Learning rates are the configured values for the head-only stage and the fine-tuning stage; the selected epoch is the epoch of maximum accuracy on the test partition.

Gen.	Model	Epochs	Unfreeze at	LR Head Stage/Fine-Tune	Weight Decay	Selected Epoch	Acc. at Selection
G3	Triple ConvNeXt-S	200	15	$10^{-4}/10^{-4}$	0.01	53	19/22
G3	Triple MobileNetV3-L	200	15	$10^{-4}/10^{-4}$	0.01	45	19/22
G4	Triple ConvNeXt-T	50	15	$10^{-4}/10^{-5}$	0.01	36	60/61
G4	Triple EfficientNet-B0	100	40	$10^{-3}/10^{-4}$	10^{-4}	72	60/61

2.6. Morphometric Branch

2.6.1. Segmentation

The cap and stipe are separated by a YOLO instance segmentation model [13] applied to the front view (Figure 9). The choice was revisited in every generation (Section 3.2.4): G1 used YOLOv12-s; G2 used YOLOv8n at 416 × 416 with confidence 0.3; G3 evaluated YOLO26 [32] at the interim stage (mask mAP@0.5 of 0.988 at 105 FPS) but delivered YOLOv8n-seg, run at 384 × 384; and G4 used YOLOv8s-seg, trained at 640 × 640 and run at 384 × 384 in the delivered evaluation code, with confidence 0.25 and IoU threshold 0.5, compiled to TensorRT.

Annotation and splits. G1 masks were produced from hand-drawn boxes prompted into YOLOE (Section 2.4). The G3 and G4 segmentation models were trained on cap and stipe polygons from the corresponding collection; the delivered material does not document whether these polygons were drawn by hand or produced by the same prompt-assisted procedure or whether a human corrected them. The G4 contract report refers to validation on 42 images but does not contain the resulting table, so no segmentation accuracy is available for G4; Section 3.2.4 gives the most recent mask quality evidence.

Usable mask. We call a segmentation usable for a specimen when a cap mask and a stipe mask are both produced, neither touches the border band of the truncation gate, and all indicators that the rule needs can be computed from them. The rule abstains otherwise. Usability rates are reported for each generation in Section 4.3.

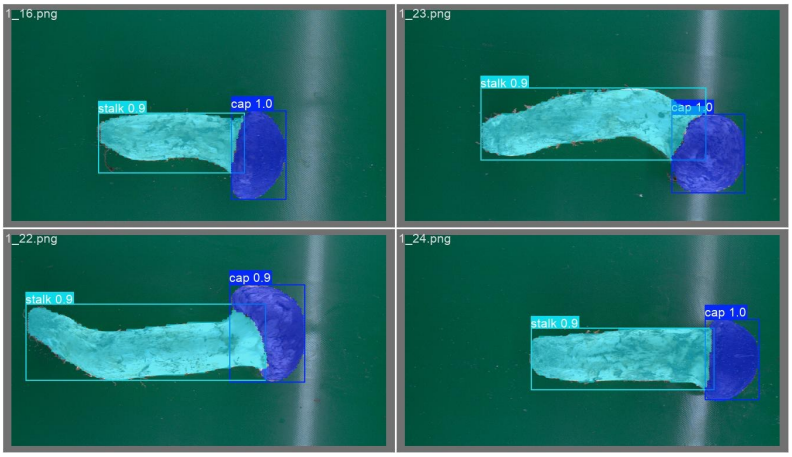


Figure 9. Cap (blue) and stipe (cyan) instance masks with confidence scores on front-view frames from the three-camera rig.

2.6.2. Indicators

Three indicators are computed from the masks (Figure 10). We call them image-derived indicators: they are geometric quantities of a predicted two-dimensional mask, and none has been validated against a physical measurement of the specimen (Section 4.6). Where they are expressed in centimetres, this is a conversion at a single global scale, not a calibrated measurement. In G1, they were computed on full-resolution masks. G2 re-implemented them for speed on masks downsampled by $\sigma_d = 8$, with the contour sub-sampled to 100 points and the mask interior randomly sub-sampled to at most 1000 points; Section 3.2.3 shows the impact of this on the thickness estimate.

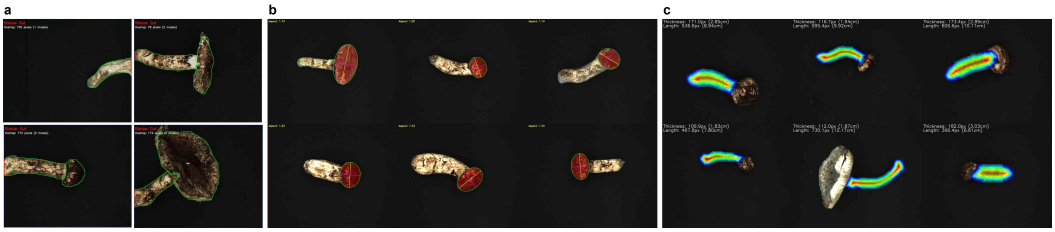


Figure 10. Indicator extraction on Camera 2 frames (G1). (**a**) Truncation gate: mask contours (green) and the count of pixels overlapping the frame border; frames marked Cut are excluded from measurement. (**b**) Cap spread ratio from a least-squares ellipse (red) fitted to the cap contour. (**c**) Stipe thickness from the distance transform (colour map) inside the interquartile band along the PCA axis, with the resulting thickness and length in pixels and centimetres.

Cap spread ratio. An ellipse is fitted to the largest external contour of the cap mask by direct least squares [33], and the ratio of the major to minor axes is taken, $r = a_{\text{major}}/a_{\text{minor}} \geq 1$. Being dimensionless, it is insensitive to calibration and rotation-invariant.

Cap area ratio. Introduced in G4, this is the share of the segmented specimen occupied by the cap,

$$\rho = \frac{|\mathcal{M}_{\text{cap}}|}{|\mathcal{M}_{\text{cap}}| + |\mathcal{M}_{\text{stipe}}|}, \tag{4}$$

a quantity that requires no contour fitting and degrades gracefully when the mask boundary is noisy—which, after the G3 field trial, was the operative concern.

Stipe thickness. This is measured over the central portion only, since both ends taper. The principal axis $\mathbf{u}$ is the leading eigenvector of the covariance of the stipe contour $\mathcal{C}$; each mask pixel is projected onto it, $\pi(\mathbf{m}) = (\mathbf{m} - \bar{\mathbf{c}})^\top \mathbf{u}$; a central band $\mathcal{M}_{\text{mid}}$ is retained—in G1, G3, and G4, the middle half of the projected extent, $\min \pi + \frac{1}{4}(\max \pi - \min \pi) \leq$

$\pi(\mathbf{m}) \leq \max \pi - \frac{1}{4}(\max \pi - \min \pi)$, and, in G2, the interquartile range of the projections of the sub-sampled pixels; and the thickness is twice the largest inscribed radius of this band from its Euclidean distance transform [34] D:

$$t_{\mathrm{px}} = 2\,\sigma_d \max_{\mathbf{m} \in \mathcal{M}_{\mathrm{mid}}} D(\mathbf{m}), \tag{5}$$

with $\sigma_d = 1$ except in G2. Using PCA for the axis makes the estimate invariant to how the specimen lies on the belt.

Stipe length. The stipe approximates a ribbon of constant width, so equating its contour perimeter P to $2\ell_{\mathrm{px}} + 2t_{\mathrm{px}}$ gives $\ell_{\mathrm{px}} = \max(0, (P - 2t_{\mathrm{px}})/2)$. This follows the curvature of a bent stipe, which a bounding box length would underestimate.

2.6.3. Truncation Gate

A specimen extending past the frame edge yields a clipped mask, and every indicator from it is invalid. With B_b, the pixels within b of any border, the specimen is flagged when the mask–border overlap reaches τ pixels. G4 additionally resolves which border was touched—left, right, or top/bottom—which is what makes the tracking of Section 2.10 possible.

2.7. Grade Rules

The rule mapping indicators to a grade took three forms, and the changes are informative in themselves.

What the rule is for. Throughout, the rule was intended to reproduce the operational decisions of the centre's inspectors from image-derived indicators, not to implement the Korea Forest Service standard; its thresholds were set from the indicator distributions of graded specimens and confirmed with the inspectors, not derived from the standard's numerical limits. Its relation to the standard is discussed at the end of this section.

G1–G2: thickness-primary, seven thresholds. The stipe thickness in centimetres selects a stipe class, and, within it, the cap spread ratio separates closed from opening specimens, with the tolerated spread growing as the stipe thins: thickness thresholds were 2.8/2.3/2.0 cm, and cap spread thresholds were 1.5/1.6/1.6/1.7. Thickness thresholds were set from the quartiles of the observed thickness distribution and confirmed with the centre's inspectors. The re-analysis shows that this distribution was that of the 90 measurable specimens of the G1 test partition: its quartiles are 2.07, 2.33, and 2.75 cm, and the per-grade means of Section 3.1.1 are those of the same 90 specimens. The rule's G1 agreement (82.42%) is therefore an in-sample figure that is optimistically biased. All seven were exposed as sliders in the operator interface.

G3: length-primary, four thresholds. The delivered G3 rule follows the structure of the standard: the stipe length in pixels is the primary discriminator and the cap spread ratio is the secondary one. Specimens with length $\ell \geq 186.3$ px are Grade 1; those with $146.3 \leq \ell < 186.3$ px are Grade 3 (opened) if $r \geq 1.5796$ and Grade 2 otherwise; shorter specimens are off-grade if $r \geq 1.4988$ and Grade 3 (growth-arrested) otherwise. The derivation of these values is not recorded.

G4: area-ratio-primary, five thresholds. The rule was restructured around the cap area ratio ρ, with the thickness in pixels as the secondary discriminator; the thin-stipe branches lead to the growth-arrested category:

$$\hat{y}_{\text{morph}} = \begin{cases} \text{Off-grade} & \rho \geq 0.4651, \\ \text{Grade 3 (opened) if } t \geq 25.58, \text{ otherwise Grade 3 (growth-arrested)} & 0.3867 \leq \rho < 0.4651, \\ \text{Grade 1} & \rho < 0.3867,\ t \geq 31.57, \\ \text{Grade 2} & \rho < 0.3867,\ 28.11 \leq t < 31.57, \\ \text{Grade 3 (growth-arrested)} & \rho < 0.3867,\ t < 28.11. \end{cases} \tag{6}$$

The submitted version gave off-grade for the two thin-stipe branches; the delivered code, and every result in this paper, uses the growth-arrested category. The derivation of the G4 thresholds is not recorded. They do not coincide with the quartiles of the G4 test partition indicators (0.345, 0.421, 0.471 for ρ; 24.8, 28.0, 30.4 px for t) or with midpoints of its per-grade means, which argues against direct fitting on the test partition but cannot exclude it. Because the G4 rule never changed a G4 grade (Section 3.4.1), this does not affect the reported G4 agreement. Two aspects changed substantively. The ellipse aspect ratio is still computed and displayed but no longer drives the decision, having been replaced by the area ratio as the primary discriminator, and thickness thresholds are pixel-domain rather than centimetre-domain, since no scale calibration was carried into the three-camera rig. Both changes trade physical interpretability for robustness, and we return to this trade in Section 4.6.

Relation to the official standard. No rule implements the Korea Forest Service specification as written. The standard is length-primary (Table 2), the G1–G2 rule is thickness-primary, and the G4 rule is area-ratio-primary, and neither uses the stipe length, for the empirical reason given in Section 3.1.1. The G3 rule is length-primary, as the standard is, but in pixels and with thresholds that are not the standard's 6 and 8 cm. The thickness and cap area ratio are proxies chosen because they are measurable from the image and discriminative on our data, not because the standard names them. We make no claim that a grade produced by Equation (6) is administratively equivalent to one assigned under the standard; the rule is an operating aid whose thresholds are set, and can be re-set, by the centre's inspectors so that its output agrees with theirs. The results therefore measure agreement with the recorded inspector decisions and say nothing about conformity to the standard. Whether the proxies agree with the standard's own quantities on the same specimens—how well the pixel thickness predicts the caliper thickness and how well the area ratio predicts an inspector's estimate of the fraction of cap opening—has not been measured; Section 3.2.3 shows that the thickness proxy is not even stable across two implementations of the same estimator (Section 4.9).

2.8. Ensemble and Arbitration

G1 used hard majority voting over dual ConvNeXt, dual ResNet-50, and the morphometric rule, with the ConvNeXt vote breaking a three-way tie.

G2 reduced to two branches, trusting agreement and otherwise deferring to the appearance branch: $\hat{y} = \hat{y}_{\text{morph}}$ if the branches agree and $\hat{y}_{\text{app}}$ otherwise; $\hat{y}_{\text{app}}$ whenever the truncation gate fires.

G4 restored three predictors but replaced voting with an asymmetric rule. If the two CNNs agree, their prediction stands. If they disagree, the morphometric prediction

is adopted only when it names one of the two lowest-priced categories; otherwise, the ConvNeXt prediction stands:

$$\hat{y} = \begin{cases} \hat{y}_{\text{cnn1}} & \hat{y}_{\text{cnn1}} = \hat{y}_{\text{cnn2}}, \\ \hat{y}_{\text{morph}} & \hat{y}_{\text{cnn1}} \neq \hat{y}_{\text{cnn2}} \ \wedge \ \hat{y}_{\text{morph}} \in \{\text{Grade 3 (growth-arrested)}, \text{off-grade}\}, \\ \hat{y}_{\text{cnn1}} & \text{otherwise.} \end{cases} \quad (7)$$

This was intended to give the rule a veto confined to the region where it was expected to be most reliable: a widely opened cap or a short, thin stipe produces a distinctive indicator pattern, whereas the distinction between Grade 1 and Grade 2 rests on a thickness difference of a few pixels. (The submitted version listed the opened-cap category instead of the growth-arrested one; the delivered code uses growth-arrested.) Section 3.4.1 shows that, on the G4 test partition, the rule's predictions of these two categories were correct only 60–65% of the time and that the veto was never invoked, because the two CNNs never disagreed. The design followed from a G3 observation, made during development, that majority voting could backfire, with two models agreeing on the same wrong answer and outvoting the one correct prediction (Section 3.3.4). Algorithm 1 states the complete per-frame procedure of the deployed system, including the tracking of Section 2.10.

Algorithm 1 G4 per-frame grading with tracking and arbitration

1. **Input:** synchronised frames $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$; thresholds $\Theta = \{\rho_5, \rho_3, t_{12}, t_{24}, t_{34}\}$; border band b, overlap threshold τ; active tracks $\mathcal{T}$
2. **Dispatch** the morphometric branch (steps 3–8) to the worker thread; run the appearance branch (step 9) on the main thread
3. **Segment:** $(\mathbf{M}_{\text{cap}}, \mathbf{M}_{\text{stipe}}, \mathbf{B}) \leftarrow \text{YOLOv8s-seg}(\mathbf{x}_1)$, with $\mathbf{B}$ the specimen bounding box
4. **Associate:** match $\mathbf{B}$ to the track in $\mathcal{T}$ of maximum IoU or register a new track in state ENTERING; delete tracks unseen for a configured number offrames
5. **Border test:** κ, side $\leftarrow \text{BorderOverlap}(\mathbf{M}_{\text{cap}} \cup \mathbf{M}_{\text{stipe}}; b, \tau)$
   if $\kappa = 1$ and side = entry: state $\leftarrow$ ENTERING; **return** (no grade yet)
   if $\kappa = 1$ and side = exit: state $\leftarrow$ LEAVING; **return** the grade stored on the track
   otherwise state $\leftarrow$ VISIBLE; if a grade is already stored, **return** it
6. **Indicators:** $\rho \leftarrow |\mathbf{M}_{\text{cap}}|/(|\mathbf{M}_{\text{cap}}| + |\mathbf{M}_{\text{stipe}}|)$; $r \leftarrow \text{EllipseRatio}(\mathbf{M}_{\text{cap}})$; $(\ell, t) \leftarrow \text{StipeMetrics}(\mathbf{M}_{\text{stipe}})$
7. **Rule:** $\hat{y}_{\text{morph}} \leftarrow \text{GradeRule}(\rho, t; \Theta)$ as in Equation (6); $\hat{y}_{\text{morph}} \leftarrow \varnothing$ if any mask or indicator is missing
8. **Join** the worker thread
9. **Classify:** $\hat{y}_{\text{cx}}, \boldsymbol{\alpha}_{\text{cx}} \leftarrow \text{TripleConvNeXt}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$; $\hat{y}_{\text{en}}, \boldsymbol{\alpha}_{\text{en}} \leftarrow \text{TripleEfficientNet}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$
10. **Arbitrate:**
    if $\hat{y}_{\text{cx}} = \hat{y}_{\text{en}}$: $\hat{y} \leftarrow \hat{y}_{\text{cx}}$
    else if $\hat{y}_{\text{morph}} \in \{\text{Grade 3 (growth-arrested)}, \text{Off-grade}\}$: $\hat{y} \leftarrow \hat{y}_{\text{morph}}$
    else $\hat{y} \leftarrow \hat{y}_{\text{cx}}$
11. **Commit:** store $\hat{y}$ on the track; send $\hat{y}$ to the PLC; display $\hat{y}$, (ρ, r, ℓ, t), $\hat{y}_{\text{cx}}$, $\hat{y}_{\text{en}}$, $\hat{y}_{\text{morph}}$ and the attention weights $\boldsymbol{\alpha}$ to the operator
12. **Output:** grade $\hat{y}$ and the measurement record for this specimen

2.9. Real-Time Deployment

Compilation. Detectors are exported through the Ultralytics pipeline to TensorRT with FP16 and a static $1 \times 3 \times H \times W$ profile. Classifiers are exported to ONNX (opset 17) with all camera inputs named explicitly and a dynamic batch axis—required for TensorRT to honour an optimisation profile—and then built with `trtexec` with min, opt, and max shapes all

fixed to batch 1, which instructs the builder to minimise batch-1 latency and removes dynamic shape checks. Classifier engines are FP32, for the reason given in Section 3.2.2.

Branch-level thread parallelism. The morphometric branch runs on a worker thread while the appearance branch runs on the main thread, so the wall-clock cost is the maximum of the two rather than their sum. Both release the Python global interpreter lock inside their CUDA and OpenCV calls. Section 3.2.1 reports why this works with two branches and failed with three.

Measurements were taken on Windows 11 with Python 3.11.7, PyTorch 2.7.1+cu128, CUDA 12.9, TensorRT 10.10.0.31, and onnxruntime-gpu 1.19.2, on an NVIDIA GeForce RTX 3060 (NVIDIA Corporation, Santa Clara, CA, USA); the G3 field trial used an RTX 4090 workstation, and the G4 measurement console reported an RTX 4080.

What the timing figures measure. Three types of timing figures appear in this paper, and Table 8 states what each includes. Model latency is the time of one model's forward pass on a preprocessed input at batch 1. Per-stage latency is the time of one stage of the pipeline at batch 1, used to locate bottlenecks. Inference throughput, to which every FPS figure for a complete generation refers, is the reciprocal of the mean wall-clock time per specimen when the complete grading pipeline is run offline over the stored test images, from the stored image pair or triple to the final grade. Only for G2 was this time itemised, and it includes image loading, preprocessing, and result assembly (Section 3.2.1); for G1, G3, and G4, the evaluation records do not itemise these stages. None of the three includes camera acquisition and transfer, the interface, serial communication, or actuation, and none is the rate at which specimens can pass along the belt, which is set by the belt speed and specimen spacing and was not measured. Whether a grade reaches the rejector in time depends on the full chain, from the moment that a specimen enters the field of view to the moment that the flip-bar moves, compared with the time that the belt takes to carry the specimen from the camera to the rejector; none of these intervals, or the belt speed or camera-to-rejector distance, was recorded at the field trials or in the belt-level experiment of Section 3.4.2; Section 4.3 states the condition that they must satisfy.

Table 8. Timing quantities reported in this paper and what each includes. None includes camera acquisition, the interface, serial communication, PLC scan, or actuation.

Quantity	Where Reported	Includes	Hardware
Model latency	Backbone FPS (Section 3.1.2); segmentation comparison (Section 3.2.4); YOLO 12.3 ms and classifier 28 and 21.4 ms (Section 3.2.2); YOLO26m 105 FPS (Section 3.2.4)	One model's forward pass on a preprocessed input at batch 1. Excludes image loading, preprocessing, the other branches, the rule, and result assembly.	RTX 3090 Ti for the backbone FPS (G1 report); otherwise as in this section
Per-stage latency	Per-stage profile and YOLO branch 14.3 ms (Section 3.2.1)	One pipeline stage at batch 1; the YOLO branch includes CPU mask processing where stated. Excludes image loading, preprocessing, and result assembly.	RTX 3060
Inference throughput	8.5, 30.64 (32.6 ms), 30.5, 55 (18.2 ms) and 30.16 FPS; Table 3; Sections 3.2.1 and 3.5; t_{inf} in Equation (8)	Mean wall-clock time per specimen for the complete pipeline run offline on stored test images, from the stored images to the final grade: segmentation, indicators, classification, rule, and arbitration. In G2, it also includes image loading, preprocessing, and result assembly; in G1, G3, and G4, these stages are not itemised in the records.	RTX 3060, except 55 FPS on the G4 console (RTX 4080)

2.10. Object Tracking for Frame Boundary Truncation

Through G3, the truncation gate simply discarded frames containing a clipped specimen. On a moving belt, this created a systematic failure: a specimen leaving the frame at

the conveyor exit was truncated in every remaining frame, so no grade was ever emitted for it. G4 added an IoU-based tracker in the manner of SORT [16], without the Kalman motion model, since belt motion is uniform and detections are dense. Each specimen is registered from the bounding box returned by the segmentation model, and successive frames are associated through the bounding box IoU. A track occupies one of three states:

- ENTERING—clipped at the entry border; grading is deferred;
- VISIBLE—wholly inside the frame; the grade is computed once, on first entry to this state, and stored on the track;
- LEAVING—clipped at the exit border; the stored grade is re-emitted without recomputation.

The border actually touched is determined from mask–border pixel overlap with a 5-pixel threshold, and tracks not re-detected for a configured number of frames are deleted to bound memory. Besides eliminating the dropped specimen failure, this makes the grading cost per-specimen rather than per-frame, contributing to the throughput reported in Section 3.4.

2.11. PLC Integration

G4 connected the grading result to the sorting hardware. Grades are written to a PLC register over a serial link (38,400 baud, 8N1) as an XGT protocol ASCII frame, and the PLC drives a flip-bar rejector that diverts the specimen to the appropriate lane. The system acknowledges each write, so a failed transmission is detectable rather than silent. How the PLC schedules the actuation relative to the specimen's arrival at the rejector (by fixed delay, encoder count, or a downstream sensor) is outside the delivered software and was not characterised.

2.12. Operator Interface

The interface is a Gradio web application, chosen over Streamlit due to lower runtime overhead, native slider widgets, and single-option link sharing, all of which suit a collection centre without server infrastructure.

Through G2, it had a parameter tab and an image viewer tab over a stored dataset (Figure 11). From G3, it became a live console: camera connection by IP or serial number with per-camera exposure, gain, and frame rate; the grade thresholds as directly editable numeric fields with increment buttons; per-model checkpoint paths, so a deployment can be re-pointed without code changes; the three live camera streams refreshed by 200 ms polling; and a results panel showing each branch's prediction, the ensemble grade, and the morphometric values behind it. Threshold edits take effect in the inference loop immediately, with no save step and no retraining. Truncated frames raise an explicit notice, so the operator knows that a judgement was based on incomplete evidence.

2.13. Evaluation Protocol

Grading performance is top-1 agreement with the recorded inspector grade on each generation's own test set. Segmentation is evaluated by mAP@0.5 and mAP@0.5:0.95 over the two part classes and by the rate of usable masks. Latency is measured per specimen at batch 1 with a per-stage decomposition. Because neighbouring grades differ in price far less than distant ones, the off-by-one accuracy and mean absolute rank error track the economic cost of a misgrade more faithfully than top-1. Only top-1 was recorded during development; the re-analysis (Section 2.2) provides per-specimen predictions, from which the per-class precision, recall, F1, confusion matrix, off-by-one accuracy, mean absolute rank error, Cohen's κ, and quadratic-weighted κ are reported for every generation and every component (Section 3.5). Ranks for the rank-aware metrics follow the standard's

price order: Grade 1, Grade 2, Grade 3 (both categories at the same rank), off-grade. A rule abstention counts as an error. Paired comparisons on the same specimens use the exact McNemar test [35].

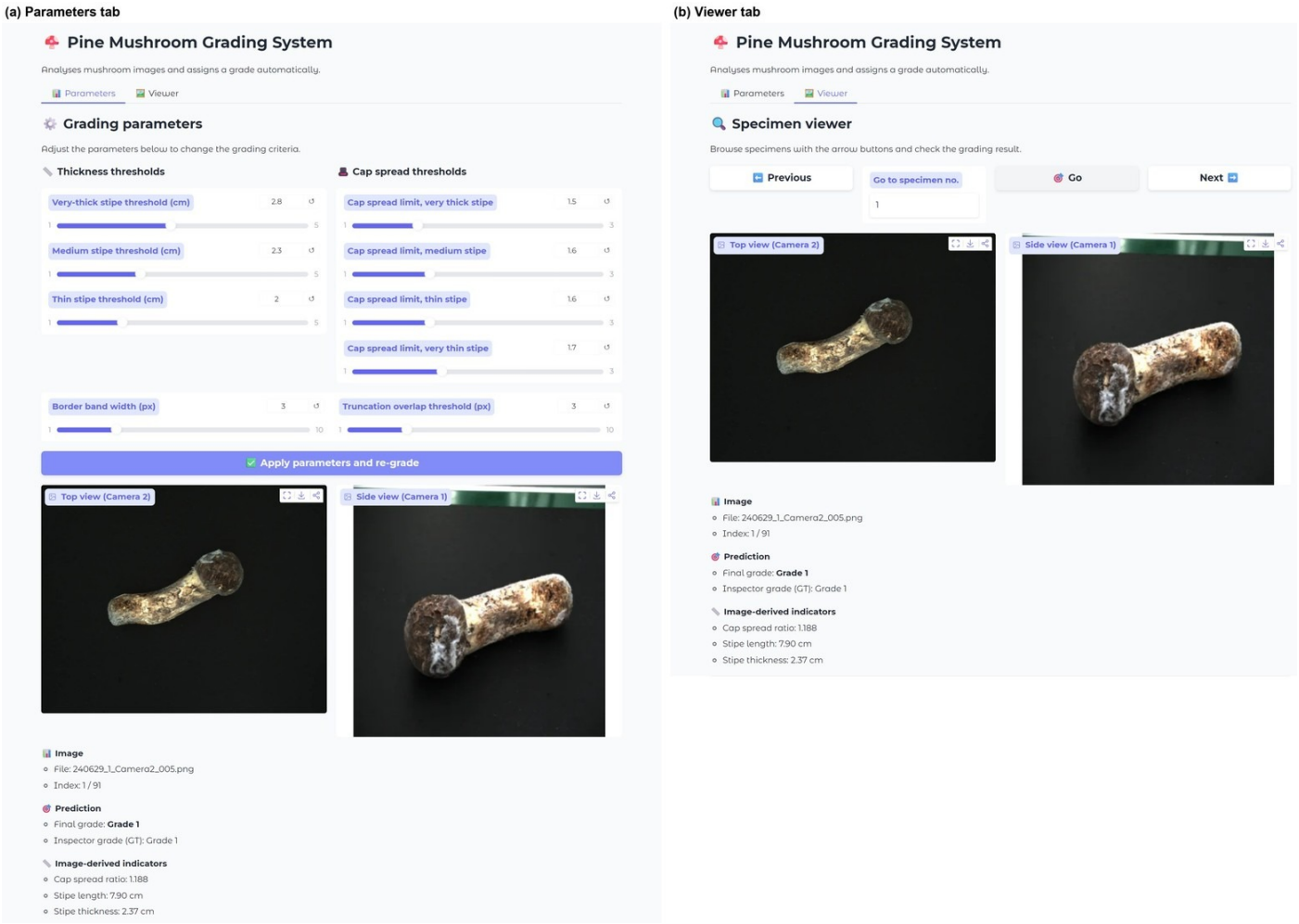


Figure 11. The G2 operator interface, shown with its user-facing labels translated into English (the deployed interface is in Korean). (**a**) Parameter tab: the seven grade thresholds and two truncation parameters as sliders, with the re-graded result of the current specimen shown below. (**b**) Viewer tab: both views, the recorded inspector grade and the predicted grade, and the measured indicators for the specimen under review.

Uncertainty. Every accuracy is a proportion k/n on a small n, and we report each with its exact count and a 95% Wilson score interval [36,37], which, unlike the normal approximation, remains valid when k is close to n. For every generation, we also report Cohen's κ [38] over the five categories and the quadratic-weighted κ over the four price ranks [39]; both are agreement between the system and the single recorded inspector grade, and neither is an inter-rater statistic, for the reason given in Section 2.4. No seed repeats (Section 2.5) were performed in any generation, so the intervals reflect test-set sampling only and not training variance.

3. Results

3.1. Generation 1: What a Bench Ensemble Achieves

3.1.1. Morphological Signature of the Grades

Table 9 reports the indicators per grade on the G1 test set, establishing both the basis of the rule and one clear negative result.

Table 9. Image-derived morphometric indicators by grade, G1 test set ($n = 91$; 90 with a complete measurement). Cap spread is a dimensionless axis ratio; length and thickness are in centimetres at the global scale of 60 px cm^{-1}. The G1 thickness thresholds were set from this distribution, so the rule is in-sample on this set (Section 2.7).

Grade	Cap Spread		Length (cm)		Thickness (cm)	
	Mean	Median	Mean	Median	Mean	Median
Grade 1	1.33	1.32	8.81	8.94	2.85	2.89
Grade 2	1.67	1.68	7.92	7.64	3.04	3.07
Grade 3 (growth-arrested)	1.34	1.35	7.66	7.65	2.31	2.33
Grade 3 (opened cap)	1.69	1.70	8.50	8.86	2.06	2.12
Off-grade	2.11	2.07	9.64	9.35	1.98	1.87

The thickness falls almost monotonically from Grade 2 (3.04 cm) through Grade 1 (2.85 cm), growth-arrested (2.31 cm), and opened cap (2.06 cm) to off-grade (1.98 cm); cap spread separates the closed-cap grades (1.33, 1.34) from the open-cap ones (1.67, 1.69, 2.11). The two are complementary rather than redundant: the thickness alone cannot separate Grade 1 from Grade 2 or growth-arrested from opened cap, and cap spread alone cannot separate Grade 1 from growth-arrested. Their conjunction can, which is the structure that the G1 rule encodes.

The length behaves differently (Figure 12). Its grade means—8.81, 7.92, 7.66, 8.50, 9.64 cm—are neither monotonic nor well separated relative to spread, and the largest belongs to the off-grade class. The length was therefore excluded from the G1 rule, even though it is the primarycriterion of the official standard (Table 2). A plausible explanation is that the length is largely determined by where the stipe was cut at collection rather than by maturity; if so, no measurement improvement would make it discriminative. This is worth stating because the length remained in the interface, and, in G3, it was reportedly promoted to the primary discriminator to match the standard—a decision that our G1 data do not support.

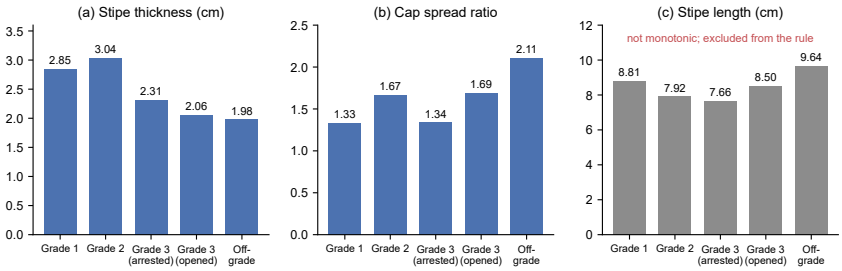


Figure 12. Mean morphometric indicators by grade on the G1 test set ($n = 91$). (**a**) Stipe thickness falls almost monotonically across the price order. (**b**) Cap spread ratio separates closed from open caps. (**c**) Stipe length is non-monotonic and was excluded from the rule.

3.1.2. Backbone and Fine-Tuning

Table 10 compares the backbones and fine-tuning regimes for the dual-view classifier.

Accuracy does not increase with capacity: ConvNeXt-Tiny (95.60%) beats Small (85.71%) and Base (91.21%), and ResNet-50 (94.51%) beats ResNet-101 and -152. With 373 training specimens, this is the expected regime, and it is convenient for deployment—the most accurate backbone is also the fastest of its family. Full fine-tuning beats head-only training everywhere, by 18.7 points for ConvNeXt-Tiny: ImageNet features alone do not separate these grades, which rest on a fine texture and boundary detail. The same small-is-better pattern recurs in the G3 and G4 screenings (Figure 13b,c), reported in their own sections.

Table 10. Dual-view classifier accuracy and throughput on the G1 test set. "All" fine-tunes the whole network; "Classifier only" trains the fusion head with the backbone frozen. FPS is model latency of the dual-view classifier alone, measured on an RTX 3090 Ti as recorded in the G1 report (Table 8).

Backbone	Fine-Tuning	Accuracy (%)	FPS
ConvNeXt-Tiny	All	95.60	38.30
	Classifier only	76.92	38.30
ConvNeXt-Small	All	85.71	27.93
	Classifier only	83.52	27.93
ConvNeXt-Base	All	91.21	26.54
	Classifier only	80.22	26.54
ResNet-18	All	92.31	54.95
ResNet-50	All	94.51	37.45
ResNet-101	All	90.11	24.75
ResNet-152	All	91.21	18.69

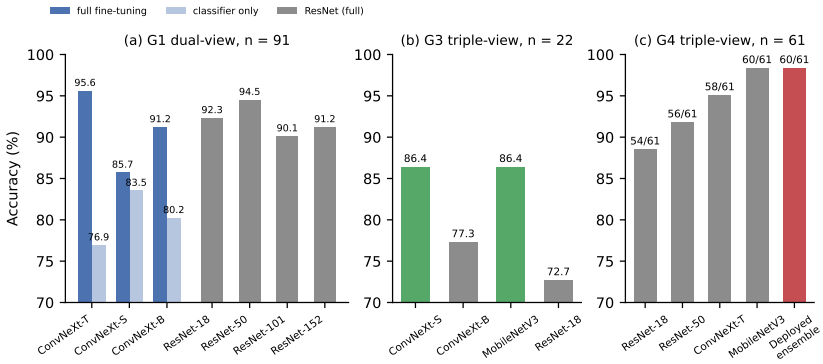


Figure 13. Backbone screenings across generations. (**a**) G1 dual-view, full fine-tuning versus classifier-only, plus the ResNet family. (**b**) G3 triple-view on the filtered 78/22 split. (**c**) G4 triple-view on the cleaned 235/61 split, with the deployed ensemble; labels give correct counts out of 61.

Figure 14 shows representative G1 outputs with the per-branch predictions and the measurements overlaid.

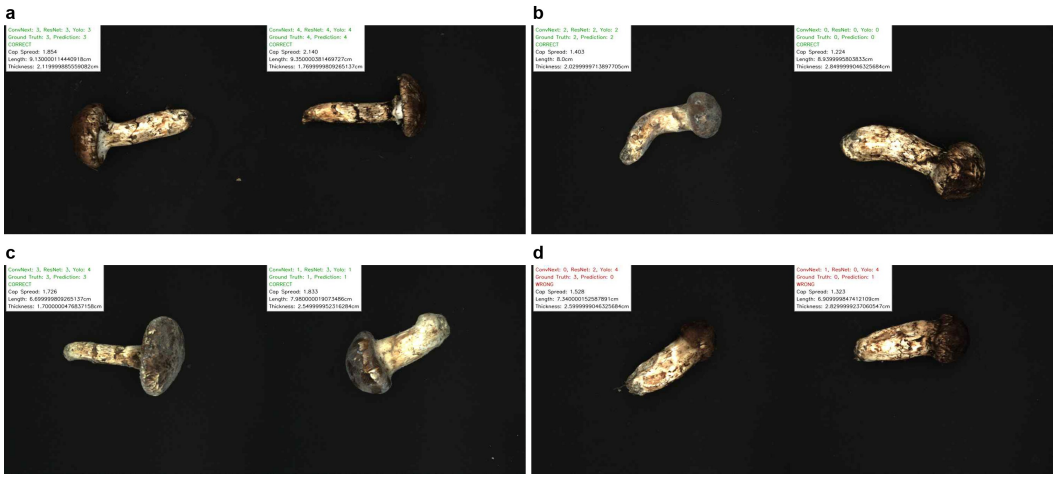


Figure 14. G1 outputs on Camera 2 frames. Each panel lists the ConvNeXt, ResNet, and morphometric predictions; the recorded inspector grade; the ensemble prediction; and the cap spread, length, and thickness. (**a**–**c**) Correctly graded specimens. (**d**) Two failure cases reported for G1: an off-grade specimen read as opened cap and a Grade 2 specimen with an approximately 8 cm, evenly thick stipe read as Grade 1. The on-screen overlays give grades as the G1–G2 code indices of Table 2 (0 = Grade 1, 1 = Grade 2, 2 = Grade 3 growth-arrested, 3 = Grade 3 opened cap, 4 = off-grade). Green overlay text marks a correct prediction and red text a wrong one.

3.1.3. What the Physical Rule Captures

Applied to the extracted indicators with no learned classifier, the G1 rule reached 82.42% (75 of 91; 95% CI 73.3–88.9%) agreement with the recorded inspector grades, against 95.60% (87 of 91; 89.2–98.3%) for the appearance branch alone. Both were scored on the same 91 specimens; the re-run gives the paired comparison: the rule was correct where ConvNeXt was wrong on two specimens, and the converse held on 14 (exact McNemar $p = 0.004$). Because the rule's thresholds were set on this test set (Section 2.7), 82.42% is an upper estimate of what the rule would achieve on new specimens. The figure describes only this implemented rule, which differs from the official standard and uses unvalidated image-derived indicators; it therefore cannot establish how much of inspector judgement the standard explains. The paired comparison also fixes the correct expectation for the hybrid: the morphometric branch cannot be the primary decider, and its value lies in producing auditable measurements and an agreement signal.

The full G1 ensemble reached 96.70% (88 of 91; 90.8–98.9%) at 8.5 FPS, exceeding its best single member by 1.1 points—one specimen, so the three predictors' errors are at most partly independent—but roughly four times too slow for the conveyor. The paired test confirms that the gain is not distinguishable from zero (one specimen gained, none lost; $p = 1$); ResNet-50 alone scored 86/91, and ConvNeXt and ResNet-50 do not differ ($p = 1$). The ensemble's three errors—a Grade 2 read as Grade 1, an opened-cap specimen read as growth-arrested, and an off-grade specimen read as opened cap—are all within one price rank, so its off-by-one accuracy is 100% (Section 3.5).

3.2. Generation 2: The Real-Time Barrier

3.2.1. Why Three Models Could Not Reach Real Time

Table 11 gives the per-stage latency of the G1 ensemble at batch 1.

Run sequentially, the components sum to 42.71 ms (23.4 FPS), already past the 33.33 ms budget. The instructive column is the parallel one: overlapping (ConvNeXt + ResNet) with (YOLO + mask processing) on one GPU slowed the first group by 45% (28.36 to 41.10 ms) and the second by 26% (14.35 to 18.04 ms, the YOLO branch including mask processing), for a net gain of 3.8%—far from the $\max(\cdot)$ behaviour that thread-level overlap is supposed to deliver. Three heavy networks contending for the same CUDA cores and memory bandwidth at batch 1 do not overlap; they interleave and interfere. We removed ResNet-50 rather than tuning the scheduler.

Table 11. Per-stage latency at batch 1 on an RTX 3060 for the G1 three-model ensemble. Naive three-way parallelism degrades both branches through GPU resource contention. In the parallel configuration, the YOLO branch includes CPU mask processing, so its sequential reference is $12.34 + 2.01 = 14.35$ ms. The throughput row is the reciprocal of the stage total and covers these stages only; it is not comparable with the 8.5 FPS inference throughput of the complete pipeline (Table 8).

Stage	Sequential (ms)	Parallel, 3 Models (ms)	Change
ConvNeXt + ResNet (PyTorch FP32)	28.36	41.10	+44.9%
YOLO branch with mask processing	14.35	18.04	+25.7%
YOLO (TensorRT FP16)	12.34	—	—
Mask processing (CPU)	2.01	—	—
Total	42.71	41.10	−3.8%
Throughput	23.4 FPS	24.3 FPS	

With two branches, the picture changes completely (Figure 15): the compiled ConvNeXt takes 21.4 ms and the YOLO branch 14.3 ms, and overlapping them gives $\max(21.4, 14.3) = 21.4$ ms. The inference throughput (Table 8) is 30.64 FPS (32.6 ms per specimen); the difference between 21.4 and 32.6 ms lies in image loading, preprocessing, and result assembly, which were not overlapped in this implementation.

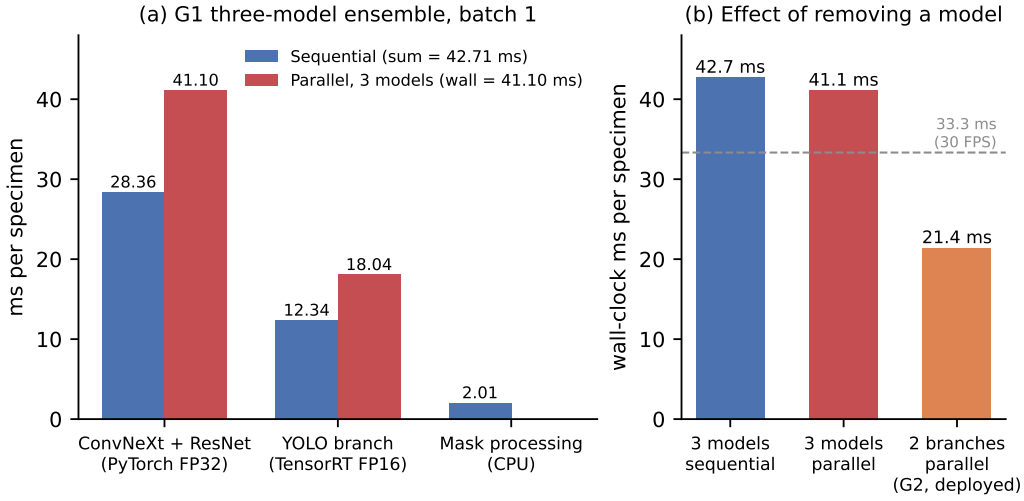


Figure 15. Latency at batch 1 on an RTX 3060. (**a**) Per-stage times of the G1 three-model ensemble, sequential versus naively parallel: both GPU branches slow down under contention. (**b**) Wall-clock time per specimen for the three configurations; only removing a model brings the pipeline under the 33.3 ms budget.

3.2.2. Precision Is Per-Architecture, Not Per-Pipeline

FP16 compilation behaved very differently for the two networks. The YOLOv8n detector, a conventional convolutional architecture, converted with no measurable accuracy loss and ran in 12.3 ms. The ConvNeXt classifier did not: FP16 dropped it below the 93% target. We did not isolate the cause, and we state the candidate explanation as a hypothesis. ConvNeXt inherits Transformer-style components—LayerNorm [40] and GELU—in which intermediate activation magnitudes span a wider range than in a convolution-plus-BatchNorm stack and can approach the limits of the FP16 representable interval ($\pm$65,504) or lose precision in the reduction inside the normalisation; this is a known sensitivity of LayerNorm-based networks under reduced precision and is consistent with what we saw. However, the observation is equally consistent with factors that we did not control for: the TensorRT builder's per-layer precision choices, the ONNX export graph and how normalisation was folded, the static batch-1 optimisation profile, the specific TensorRT and driver versions, and preprocessing parity between the PyTorch reference and the engine. A diagnosis would require a layer-wise comparison of activations between the FP32 and FP16 engines and a build with LayerNorm and Softmax pinned to FP32 while the rest of the network runs in FP16, which is the standard remedy if the hypothesis is correct; neither was performed. What the experiment does establish is narrower and, for deployment, sufficient: under one build configuration, FP16 was accuracy-neutral for one architecture and not for the other, so the precision policy had to be verified per network rather than assumed per pipeline.

Because batch-1 real-time grading values accuracy and stability over raw throughput, the classifier was kept at FP32 and accelerated structurally instead: batch-1-specialised compilation, layer fusion, and kernel auto-tuning took it from about 28 ms in PyTorch to 21.4 ms, a 23.5% reduction with no accuracy change. This was sufficient, since, after ResNet removal, the two branches were close enough in cost that a further classifier speed-up

would not have lowered the pipeline maximum. G3 repeated the experiment on its own models and reached the same conclusion, retaining FP16 for MobileNetV3 and FP32 for ConvNeXt; measured against the PyTorch reference, the mean logit deviation was about 0.001 for ConvNeXt FP32 and 0.02 for MobileNetV3 FP16, with predictions matching in 100% of cases for both.

3.2.3. What the G2 Figure Measures and an Indicator Drift

G2 reached 85 of 91 (93.41%; 95% CI 86.4–96.9%) on the G1–G2 test partition. The re-run shows that this figure is the appearance branch alone. The G2 arbitration returns the rule's grade when the two branches agree and the CNN's grade otherwise, so its output equals the CNN's on every specimen by construction; on the 91 test specimens, the two were identical. The G2 classifier is the same checkpoint as the G1 ConvNeXt (identical file), and the loss from 87 to 85 specimens comes from running it at 416×416 instead of the 384×384 used in G1, not from the change in decision logic.

The re-implemented indicator code changed more than the speed. On the same 91 specimens, the G2 rule agreed with the inspectors on 50 (54.95%; 44.7–64.8%, two abstentions) against 75 for the G1 rule (exact McNemar $p < 10^{-5}$). Figure 16 traces the cause. Cap spread agreed closely between the two implementations ($r = 0.96$, mean difference -0.04) and length moderately ($r = 0.84$, -0.52 cm), but the thickness did not agree at all: $r = -0.32$, mean difference -0.60 cm, 95% limits of agreement -2.38 to $+1.19$ cm, against a G1 mean of 2.41 cm. The downsampling by $\sigma_d = 8$ quantises the distance transform to steps of 0.27 cm, and the random sub-sampling of interior pixels makes the estimate non-deterministic: over 20 repeated runs on the same images, the thickness of 28 of 91 specimens varied (median range 0.37 cm, maximum 1.31 cm) and the rule's grade changed for three. Because the arbitration never allows the rule to change a grade, none of this reached the reported accuracy, which is precisely why it went unnoticed. It illustrates two points that recur below: an image-derived indicator is not a measurement until it has been validated, and the same named quantity can lose its meaning across two implementations of one estimator.

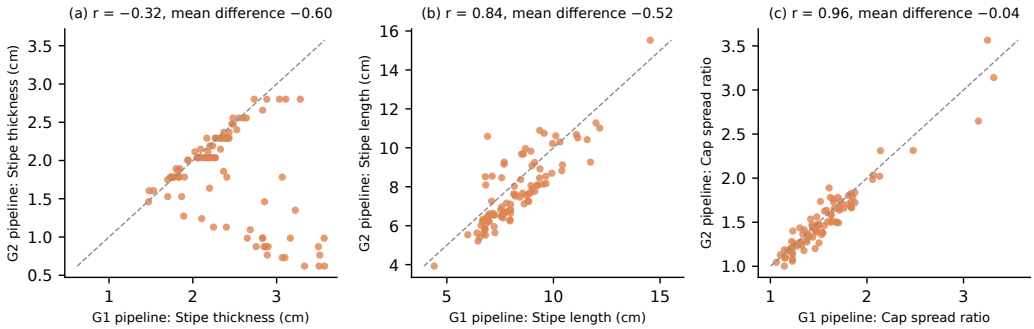


Figure 16. Indicator agreement between the G1 and G2 implementations on the same 91 Camera 2 images. (**a**) Stipe thickness: the two implementations are negatively correlated. (**b**) Stipe length. (**c**) Cap spread ratio. Dashed lines mark equality.

3.2.4. Segmentation Model Selection

Table 12 compares the segmentation candidates evaluated in G2 and G3.

Table 12. Segmentation candidates. Upper block: G2 comparison at 416 × 416. Lower block: the G3-era comparison on three-camera data, reporting inference time and model cost; YOLO26m was the interim G3 choice, and the delivered G3 and G4 pipelines use the YOLOv8 family. Deployed choices in bold. [a] Mask mAP@0.5.

Gen.	Model	mAP@0.5	mAP@0.5:0.95	FPS	Params/GFLOPs
G2	YOLOv12n	0.985	0.941	135.1	—
	YOLOv12s	0.990	0.949	62.9	—
	YOLOv12m	0.985	0.945	64.1	—
	YOLOv8n	**0.990**	**0.959**	**87.0**	—
	YOLOv8s	0.990	0.955	76.9	—
	YOLOv8m	0.955	0.762	53.8	—
G3	**YOLOv8n**	**0.981** [a]	—	**5.1 ms**	**3.26 M/11.3 G**
	YOLO11n	0.990 [a]	—	5.4 ms	2.83 M/9.6 G
	YOLO26n	0.990 [a]	—	9.0 ms	3.04 M/9.0 G
	YOLO26m	0.988 [a]	—	10.4 ms	26.9 M/121.2 G

At mAP@0.5, the task is easy for every candidate—one object of two parts on a uniform background—so the discriminating metric is the boundary precision, which propagates directly into the measurements (Figure 17). In G2, YOLOv8n led in the mAP@0.5:0.95 at 0.959, i.e., 1.8 points above YOLOv12n, while fast enough not to be the bottleneck. Scaling up was harmful: YOLOv8m collapsed to 0.762, which we attribute to overfitting a 373-specimen training set. At the interim stage, G3 used YOLO26m, which reached a mask mAP@0.5 of 0.988 at 105 FPS on the three-camera data; the lower block of Table 12 is the comparison that then argued for returning to the YOLOv8 family, and the delivered G3 pipeline uses YOLOv8n-seg: YOLOv8n's mask mAP is 0.009 below the best candidate while it is the fastest at 5.1 ms, whereas YOLO26m's 0.7-point mask gain costs 10.7× the arithmetic and, at the time, lacked a stable TensorRT export path. G4 adopted YOLOv8s (trained at 640 × 640), trading some throughput for boundary quality once tracking had reduced grading from per-frame to per-specimen. These figures reflect the pixel-level mask quality on images with good masks; they do not say how often a usable mask is produced on new material, which Section 4.3 reports.

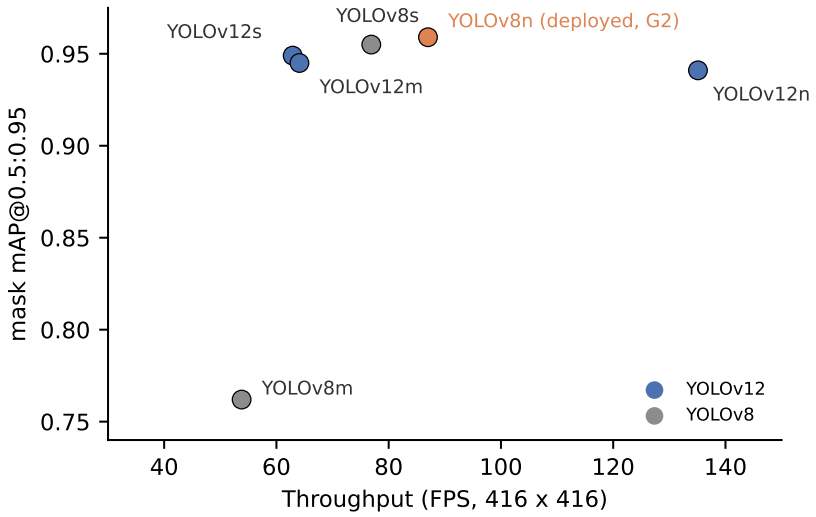


Figure 17. G2 segmentation candidates: mask mAP@0.5:0.95 against throughput at 416 × 416. YOLOv8n was adopted; YOLOv8m overfits the 373-specimen training set.

3.3. Generation 3: The Field Encounter

3.3.1. Performance Collapse on New Data

Re-evaluating the G2 dual-network architecture on newly collected three-camera data produced the result in Table 13.

Table 13. G2-style dual-network accuracy by camera pair, and triple-network accuracy, on the raw G3 test set (40 specimens).

Camera Combination	Accuracy (%)
Camera 1–2	55.0
Camera 1–3	67.5
Camera 2–3	65.0
Camera 1–2–3 (triple network)	70.0

Against a value above 90% on G2 data, this is a drop of more than 20 points (Figure 18). With 40 specimens, the intervals are wide—70.0% is 28 of 40, 95% CI 54.6–81.9%, and 55.0% is 22 of 40, 39.8–69.3%—and, because the two evaluations use different collections, they are not a paired comparison. Even so, the upper limit of the best combination lies below the lower limit of the G2 result (86.4%), a separation that is consistent with substantial distribution shift rather than with small-sample fluctuation alone. Two causes were identified: the G3 training set is about 60% smaller in specimens, and it contains outlier samples whose influence on generalisation is disproportionate at this scale. The third camera helps—70.0% for the triple network against 55.0% for the worst pair—confirming that additional viewpoints carry real information, but it does not rescue the absolute level.

This is the single most important measurement in the study for interpreting all others. The same architecture that scored above 90% on the dataset that it was developed with scored 70% on data collected a few months later under a nominally similar protocol. Whatever the G1 and G2 headline numbers measured, it was not generalisation to the next batch of mushrooms.

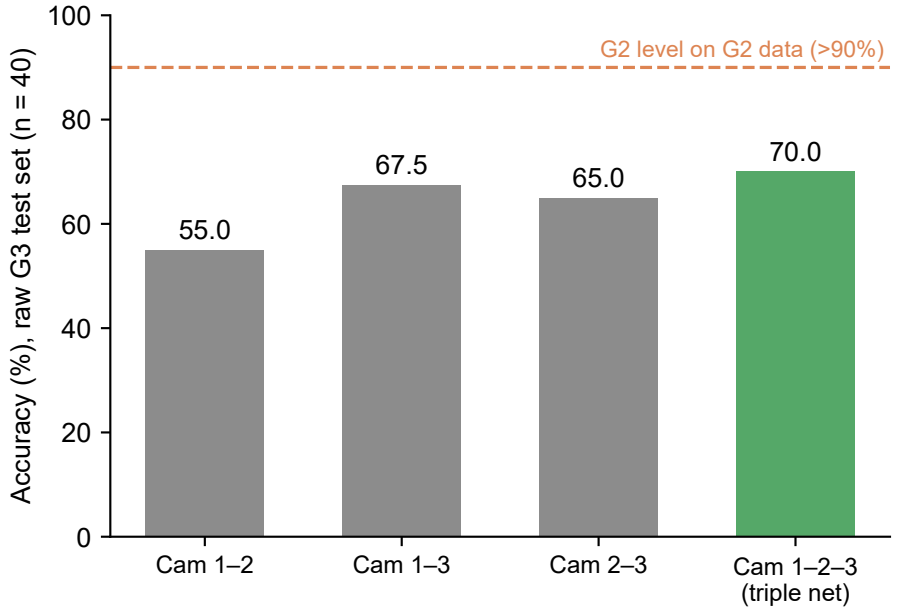


Figure 18. Accuracy of the G2-style architecture on the raw G3 test set by camera combination. Every combination is more than 20 points below the level that the same architecture reached on G2 data.

3.3.2. Feature Space Analysis

To ask whether the grades are visually separable at all, DINOv3 [41] features were extracted from segmented regions and projected with t-SNE [42] under three conditions: whole specimen, cap only, and stipe only (Figure 19). No condition produced clearly separated per-grade clusters, confirming that the inter-grade visual differences are subtle. Cluster cohesion was ordered as whole > cap only > stipe only, with stipe only being the most dispersed. Two design consequences followed: combining cap and stipe features is better than either alone, which supports the part-based decomposition, and the stipe carries less grade-discriminative visual signal than the cap, which is consistent with the G1 finding that the stipe length does not separate grades.

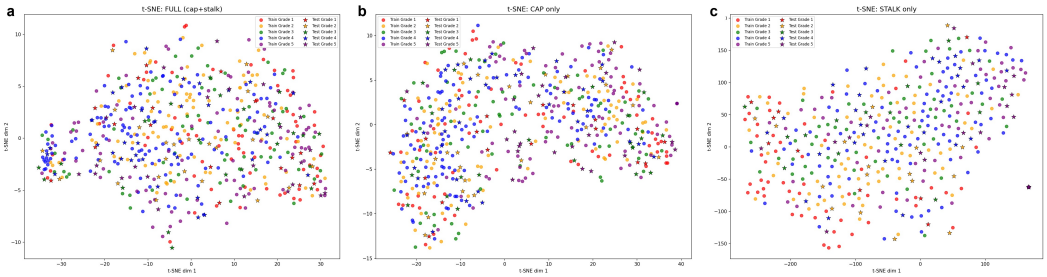


Figure 19. t-SNE projections of DINOv3 features of G3 specimens, coloured by grade (circles: training, stars: test). (**a**) Whole specimen. (**b**) Cap only. (**c**) Stalk only. No condition separates the grades; the stalk-only embedding is the most dispersed.

3.3.3. Backbone Selection Under Data Scarcity

On the filtered 78/22 split, ConvNeXt-Small and MobileNetV3 tied at 86.4% (19 of 22; 95% CI 66.7–95.3%) while ConvNeXt-Base (77.3%, 17 of 22) and ResNet-18 (72.7%, 16 of 22) trailed by two and three specimens (Figure 13b). At this sample size, the ordering is suggestive rather than established—all four intervals overlap—but it is the same ordering as G1 and G4. The G1 lesson—smaller is better at this data scale—held, and the two small models were carried forward as the G3 ensemble.

3.3.4. Qualitative Behaviour and a Voting Pathology

Qualitative review found the ensemble correcting morphometric errors as designed: a specimen whose cap was hemispherical (aspect ratio 1.312) was mis-scored as Grade 1 by the rule and correctly returned to Grade 2 by the two CNNs; a specimen with an irregular, widely opened cap (aspect ratio 1.529) and almost no stipe was mis-scored as off-grade by the rule and correctly graded by the CNNs.

It also found the opposite. In one case, a camera angle distortion pushed the aspect ratio to 1.929, far above the off-grade median, and the morphometric branch and MobileNetV3 both returned off-grade, while ConvNeXt alone was correct at Grade 2 (Figure 20d). Majority voting adopted the two-to-one wrong answer. This is the pathology that motivated the asymmetric arbitration of G4 (Section 2.8): with correlated errors, a vote is not a safeguard. The examples of Figure 20 were recorded during development, with intermediate checkpoints. With the delivered checkpoints, the pathology does not occur on the filtered test set: the vote corrected two ConvNeXt-S errors and introduced none (exact McNemar $p = 0.5$).

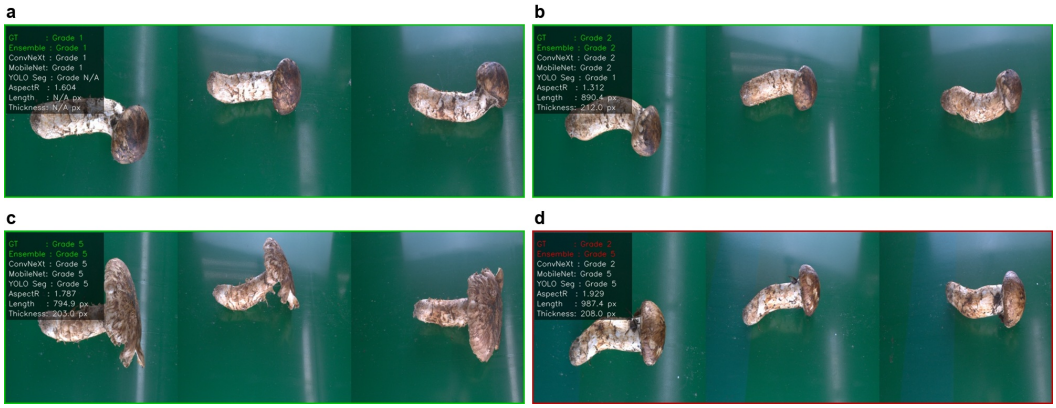


Figure 20. G3 outputs on three-camera triples, with the recorded inspector grade; the ensemble grade; the ConvNeXt, MobileNetV3, and morphometric predictions; and the measured aspect ratio, length, and thickness. (**a**) All three agree on Grade 1. (**b**) The rule mis-scores a hemispherical cap (aspect ratio 1.312) as Grade 1 and is outvoted correctly. (**c**) All three agree on off-grade. (**d**) The voting pathology: a camera-angle distortion (aspect ratio 1.929) leads the rule and MobileNetV3 to off-grade, outvoting ConvNeXt's correct Grade 2. The on-screen overlays give grades as the G3–G4 code indices of Table 2 (1 = Grade 1, 2 = Grade 2, 3 = Grade 3 opened cap, 4 = Grade 3 growth-arrested, 5 = off-grade). Green frames and text mark a correct ensemble grade and red a wrong one; GT is the recorded inspector grade. Panels were recorded during development with intermediate checkpoints.

The complete G3 pipeline reached 95.45% (21 of 22; 95% CI 78.2–99.2%) at 30.5 FPS on the filtered test set. The interval is the widest in the paper and the test set is the most heavily curated; the figure shows that the pipeline worked on the specimens that survived filtering, while achieving little else. The single error is a growth-arrested specimen read as Grade 2. The length-primary rule of Section 2.7 alone agreed on 15 of 22 (68.2%), abstaining on three that the truncation gate flagged, and it never returned the opened-cap category correctly (0 of 3); its agreement with the inspectors does not differ significantly from ConvNeXt-S on 22 specimens ($p = 0.34$).

3.3.5. The Field Trial That Failed

On 18 March 2026, the system was installed at the collection centre—a workstation with three Basler cameras over the conveyor (Figure 21a)—and run on real specimens. The specimens had been frozen for storage and thawed for the trial, while all training data comprised fresh material. Freezing and thawing changes the surface texture and colour and deposits frost residue, and, under this shift, the segmentation model failed to produce usable cap and stipe masks. This is an observation made by the operators on the live display; no count of specimens, masks, or grades was kept at the trial, so the trial itself yields no accuracy and no failure rate. Section 3.3.6 quantifies the same shift offline.

The consequence was structural rather than gradual. With no masks, no indicators could be computed, so the morphometric branch produced nothing, and the fallback path delivered a grade from the CNN classifiers alone. The system continued to emit grades and gave no indication that it had lost half its evidence. Everything that made the design defensible—the measurements in physical units, the auditable rule, and the agreement signal between independent branches—was silently unavailable, in the first situation where a real payment might have depended on it.

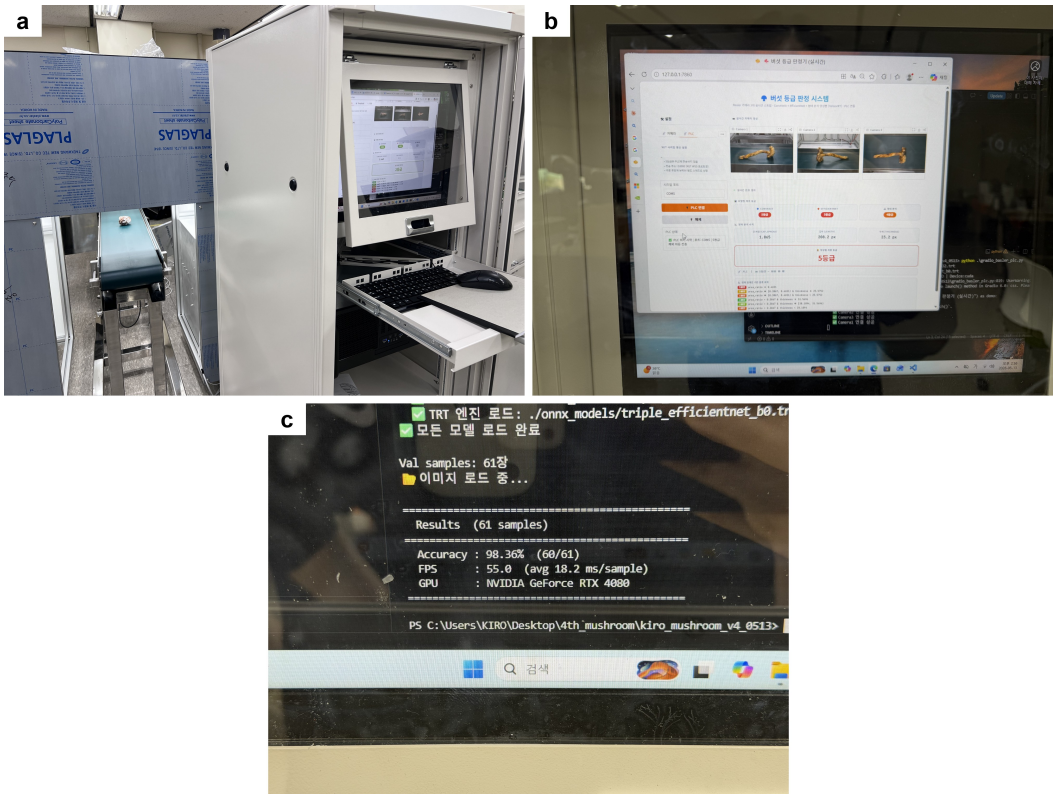


Figure 21. Field trials at the collection centre. (**a**) 18 March 2026: three-camera rig and workstation installed over the conveyor. (**b**) 13 May 2026: the live operator interface grading specimens on the belt. (**c**) The measurement console at the second trial: 60 of 61 specimens correct, 55.0 FPS, 18.2 ms per sample, RTX 4080. The interface in (**b**) is the deployed Korean version (its English version is shown in Figure 11); in (**c**) the Korean console messages read 'all models loaded' and 'loading images', and the taskbar field reads 'search'.

Two properties of this failure deserve emphasis. It was asymmetric in what it made visible: the classification branch kept producing plausible outputs, while the segmentation branch failed in a way that the operators could see. Section 3.3.6 shows that the asymmetry was one of observability, not of robustness: on frozen–thawed material, the classifiers degraded more than the masks did, but their failure produced confident grades rather than missing ones. Moreover, it was silent: nothing in the pipeline distinguished "the rule abstained because the specimen was truncated" from "the rule abstained because segmentation collapsed under a domain shift".

3.3.6. Forward Evaluation: The G3 Pipeline on the G4 Test Split

Because the trial kept no counts, we applied the delivered G3 pipeline—with the checkpoints, segmentation model, rule, and vote unchanged—to the 61-specimen G4 test split, which was collected after G3 was frozen and includes frozen–thawed specimens. This is the only like-for-like measurement of the G3 system on later material and the closest available substitute for the per-condition analysis that the storage records do not allow (Section 2.4). The G3 and G4 collections share no specimens: the most similar cross-collection image pair has ConvNeXt embedding cosine similarity of 0.826, against 0.968 for re-photographed duplicates within one collection.

The G3 pipeline agreed with the inspectors on 20 of 61 specimens (32.8%; 95% CI 22.3–45.3%), against 95.45% on its own filtered test set. Its members scored 26/61 (ConvNeXt-S, 42.6%), 13/61 (MobileNetV3, 21.3%), and 8/61 (rule, 13.1%). Segmentation degraded but did not vanish: a cap mask was produced for 60 of 61 specimens and a stipe mask for 40, so the rule had usable masks for 39 (63.9%) and abstained on 22. The classifiers,

by contrast, never abstained, and they failed by returning a wrong grade. On this evidence, the field observation of Section 3.3.5 was correct in direction and wrong in emphasis: the measured branch abstained on a third of the specimens, while the learned branch lost more than half of its accuracy without any outward sign. The disagreement between the two CNNs would have flagged 32 specimens, including 19 of the 41 errors; this is the only operational signal that the G3 design offered, and it was not used. Two qualifications apply. The G4 test split mixes fresh and frozen–thawed specimens in unrecorded proportions and was collected in a different period, so the drop combines storage and seasonal shift and cannot be attributed to either, and the G4 split was cleaned by the G4 criteria rather than the G3 ones, so part of the drop may reflect specimens that the G3 filter would have removed.

3.4. Generation 4: Recovering on the Frozen–Thawed Collection

G4 responded on four fronts: frozen–thawed specimens were collected and merged into the training set and all models retrained; the dataset was cleaned of cross-camera pair mismatches and label errors, yielding 296 specimens; camera–attention fusion replaced uniform concatenation; and IoU tracking replaced frame-dropping. Table 14 gives the backbone screening and the deployed result.

The deployed ensemble—triple-view ConvNeXt and EfficientNet, arbitrated against the morphometric branch—graded 60 of 61 specimens correctly (98.36%; 95% CI 91.3–99.7%) at 55 FPS (18.2 ms/specimen) on the field workstation, namely an NVIDIA GeForce RTX 4080 (Figure 21c), and 30.16 FPS on the RTX 3060 reference machine. A second field trial on 13 May 2026 confirmed end-to-end operation: three-camera capture, live visualisation, grading, and PLC-driven flip-bar rejection (Figure 21b). This trial was a functional demonstration: it demonstrated that the chain works, but it did not measure how well. The 55 FPS figure on the console is the inference throughput on stored triples (Section 2.9), not a line rate. No count of specimens run, grades emitted, rejector diversions, or mis-diversions was kept at this trial; the belt-level experiment of Section 3.4.2, carried out afterwards, provides these counts. The 98.36% figure remains a laboratory figure on stored triples rather than a measurement of the line. Two operational defects were logged at the trial and remain open: the empty belt between specimens was itself sometimes graded, so an empty-belt gate is needed ahead of the pipeline, and the path from live capture through the browser interface to the PLC added latency that the per-specimen inference figure does not include and that was not measured. Section 4.3 sets out how both should be closed, and Section 4.8 gives the operating metrics that a third trial should record.

Table 14. Triple-view single-backbone screening on the cleaned G4 test set (61 specimens), with the deployed ensemble for comparison. Every entry is an exact count out of 61, so the accuracy column has granularity of 1.64 points; the 95% Wilson intervals of all six rows overlap. The screening rows were recorded during development; the EfficientNet-B0 row is the delivered ensemble member evaluated alone in the re-analysis. All checkpoints were selected on this test set (Table 7).

Configuration	Correct/61	Accuracy (%)	95% CI (%)
ResNet-18	54	88.52	78.2–94.3
ResNet-50	56	91.80	82.2–96.4
ConvNeXt-Tiny	58	95.08	86.5–98.3
MobileNetV3	60	98.36	91.3–99.7
EfficientNet-B0 (deployed member)	60	98.36	91.3–99.7
Deployed: ConvNeXt + EfficientNet + morph	60	98.36	91.3–99.7

The screening column deserves comment, because, at this sample size, it says less than it appears to. MobileNetV3 as a single backbone also scored 60 of 61, matching the deployed ensemble exactly; the two configurations are separated by no test specimen at all. The ConvNeXt-Tiny row (58) is taken from the development screening; the delivered ConvNeXt-T checkpoint scores 60 of 61 alone (Section 3.4.1), so the screened and delivered models are different training runs. They are development runs, not seed repeats of one configuration, and the records do not show whether their settings differed, so the two-specimen difference cannot be read as training variability. Nothing in this table establishes that the ensemble is better than its best single member, and, with 61 specimens, nothing could: the entire spread from worst to best is six specimens. The ensemble was retained for the reason given in Section 2.8—two independent classifiers plus a measurement branch give an agreement signal and a constrained veto that a single network cannot—and not on the strength of this comparison.

The submitted version stated that EfficientNet, half of the deployed ensemble, had not been evaluated as a single backbone. The re-analysis evaluates the delivered member alone: it scores 60 of 61, with the same single error as the ensemble. The ensemble members were nevertheless fixed before the screening rather than chosen by it, and no other pair was trained, so we make no claim that ConvNeXt and EfficientNet are the best available pair, only that the resulting system performs as reported.

Four further cautions are attached to the 98.36% figure. Both classifiers were checkpointed at the epoch of best accuracy on these 61 specimens (Table 7), so the figure is biased upwards by model selection. It is measured on 61 specimens, so a single additional error moves it by 1.64 points and the 95% interval reaches down to 91.3%; a reader who desires one number should take the lower limit. It is measured on a test set from which cross-camera mismatches, label-ambiguous specimens, and duplicates were deliberately removed—a defensible choice for training, but one that makes the test set easier than the belt by an amount that we have not measured, because no uncurated set was held back. Moreover, it is an internal split of the same collection that the models were trained on; Section 3.3.1 shows what happened the last time an internal split figure of this kind met the next collection. The appropriate reading is that the observed frozen–thawed gap was mitigated on the material at hand—an internal split of a collection that now includes frozen–thawed specimens—and that the pipeline works end to end, not that the domain gap is solved in general, not that grading is now solved to within 1.64%, and not that next season's harvest will be graded at this rate.

The re-run gives the per-specimen predictions directly. The single mis-graded specimen was a Grade 1 specimen that both CNNs predicted as Grade 2 (the morphometric branch returned Grade 3 (opened) for it, so the arbitration rule left the CNN agreement in place); the test partition contains 11, 9, 13, 14, and 14 specimens of the five grades. Table 15 and Figure 22 give the breakdown. Every class other than the two involved has precision and recall of 1.00; the Grade 1 recall is $10/11$ and the Grade 2 precision is $9/10$. Because the error is between adjacent price ranks, the off-by-one accuracy is 100% and the mean absolute rank error is $1/61 = 0.016$. Chance-corrected agreement with the recorded inspector grade is $\kappa = 0.979$ over the five categories and quadratic-weighted $\kappa = 0.992$ over the four price ranks. These are high because the single error is adjacent and the classes are balanced; we report them so that a future inter-rater study (Section 4.8) has a system-versus-inspector figure to set beside the inspector-versus-inspector one, which is the comparison that actually bounds what a grader can achieve.

Table 15. Per-class performance of the deployed G4 configuration on the 61-specimen test set, from the re-run of the delivered models; the single error is a Grade 1 specimen predicted as Grade 2.

Grade	Support	Precision	Recall	F1
Grade 1	11	1.000	0.909	0.952
Grade 2	9	0.900	1.000	0.947
Grade 3 (opened cap)	13	1.000	1.000	1.000
Grade 3 (growth-arrested)	14	1.000	1.000	1.000
Off-grade	14	1.000	1.000	1.000
Macro average	61	0.980	0.982	0.980

The one error is instructive. The specimen's cap area ratio, 0.4519, lay above the $\rho_3 = 0.3867$ boundary, so the rule read the cap as opened and, with a thickness of 28.0 px, above the $t_{34} = 25.58$ px boundary, returned Grade 3 (opened); both CNNs read the same specimen as Grade 2. The inspector had graded it as Grade 1 on the basis that the cap was not in fact opened. A cap that is closed but broad projects a large cap area from the front view, which is exactly the failure mode that the area ratio was known to carry when it replaced the ellipse ratio; the two CNNs erred in the same direction, which is also the pattern that Section 3.3.4 identifies. An adjacent-grade error on a broad-capped Grade 1 specimen is the least costly kind, but it is not an accident of sampling, and it is the case that a future cap opening estimator should target.

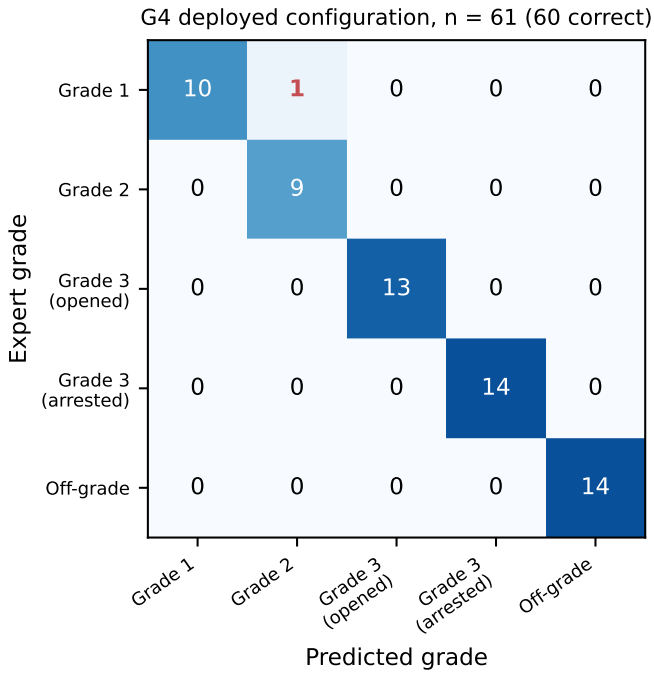


Figure 22. Confusion matrix of the deployed G4 configuration on the 61-specimen test set. Rows are recorded inspector grades, columns are predicted grades; the single off-diagonal entry is the Grade 1 specimen predicted as Grade 2.

3.4.1. Component and View Ablation

Table 16 and Figure 23 decompose the deployed configuration on the same 61 specimens. Each component was run alone; the fusion weights were replaced by uniform weights at inference; and each camera image was fed to all three inputs in turn, which isolates what each view supports on its own.

Table 16. Ablation of the deployed G4 configuration on the 61-specimen test split, from the delivered checkpoints at inference (no retraining). "Camera *k* only" feeds the camera-*k* image to all three inputs. *p*: exact McNemar test against the deployed ensemble.

Configuration	Correct/61	95% CI (%)	*p*
Deployed ensemble (ConvNeXt-T + EfficientNet-B0 + rule)	60	91.3–99.7	—
ConvNeXt-T alone	60	91.3–99.7	1
EfficientNet-B0 alone	60	91.3–99.7	1
Morphometric rule alone	32	40.2–64.5	$<10^{-8}$
ConvNeXt-T, uniform view weights	60	91.3–99.7	1
EfficientNet-B0, uniform view weights	53	76.2–93.2	0.016
ConvNeXt-T, camera 1/2/3 only	55/36/55	—	—
EfficientNet-B0, camera 1/2/3 only	36/21/60	—	—

Three findings bear on what the ensemble is. First, the two CNNs returned the same grade for all 61 specimens, including the single error, so the arbitration of Equation (7) never reached its second or third case and the rule never changed a grade; removing either CNN, or the rule, leaves the output unchanged. Second, the rule alone agreed on 32 of 61 (52.5%), and its predictions of the two categories in its veto set were correct 9 of 15 times (growth-arrested) and 11 of 17 times (off-grade); had the veto been invoked, it would have been invoked with a 60–65% chance of being correct. Third, the camera–attention fusion did not do what it was designed to do. The ConvNeXt weights were nearly constant across specimens (0.376, 0.289, and 0.336 for cameras 1–3, standard deviation below 0.002), so they amount to a fixed weighted average, and replacing them with uniform weights changes no prediction. The EfficientNet weights collapsed onto Camera 3 for every specimen (weight 1.000; the other two below 10^{-14}), so the EfficientNet member is in effect a single-view classifier, which is why it scores 60 with Camera 3 alone and degrades under uniform weights. Camera 2 alone supports the weakest grading in both networks.

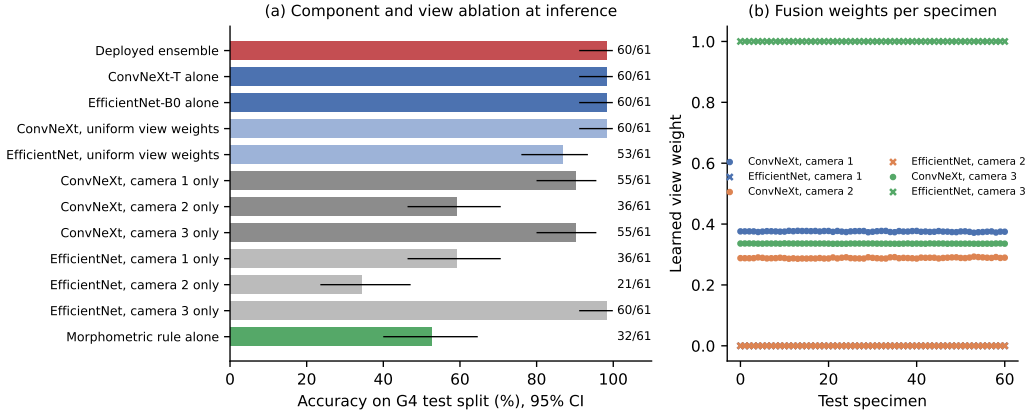


Figure 23. G4 ablation. (**a**) Agreement with the recorded inspector grade for each component and view configuration, with 95% Wilson intervals. (**b**) Learned camera weights per test specimen: the ConvNeXt weights (dots) are nearly constant, and the EfficientNet weights (crosses) place all mass on Camera 3. The EfficientNet weights of Cameras 1 and 2 are both zero, so their markers coincide.

The appropriate description of the deployed G4 configuration is therefore a ConvNeXt classifier with a near-fixed view average, accompanied by a second classifier that duplicates its outputs on this test split and a rule that displays measurements but did not affect any grade. The ensemble provides redundancy and an agreement signal in principle; on this split, it provided neither, because the members never disagreed. The forward evaluation of Section 3.3.6 shows that, on more challenging material, disagreement does occur and does flag errors.

3.4.2. Belt-Level Sorting Experiment

To measure what the field trials did not, the deployed G4 configuration was run on the laboratory conveyor sorter of Figure 1. The specimens were the 61 physical specimens whose images formed the G4 test split (Table 6), each with the inspector grade recorded at intake. They were placed on the infeed belt one at a time, imaged in the tunnel, graded, and diverted by the flip-bar lanes on the PLC's command; for every specimen, we recorded the inspector grade and the lane that it physically reached.

Table 17 gives the result. All 61 specimens received a grade and none passed without one. Sixty (98.36%; 95% CI 91.3–99.7%) reached the lane of their inspector grade. The one wrong-lane specimen was the specimen misgraded in the image-level evaluation (Section 3.4): inspector Grade 1, sent to the Grade 2 lane, and an adjacent-rank error caused by the grade rather than by the actuation. No specimen was diverted at the wrong time or on the wrong specimen's grade. The correct lane rate on the belt is therefore identical to the agreement on the stored images: in this run, the physical line added no error of its own.

The result has three limits. First, the specimens are the G4 test split, on which the G4 checkpoints were also selected (Table 7); the experiment shows that the line reproduces the image-level grades, not that these grades generalise to unseen specimens. Second, it is a single run on one laboratory installation with specimens fed one at a time, not the collection centre's line at its working throughput. Third, the belt speed, the imaging-to-rejector distance, the time stamps of the decision chain, and the empty-belt false trigger rate were not recorded in this run, so the timing margin required by Equation (8) remains unmeasured.

Table 17. Belt-level sorting experiment on the laboratory conveyor sorter of Figure 1, with the 61 specimens of the G4 test split. Lane accuracy is agreement between the lane reached and the inspector grade recorded at intake.

Quantity	Value
Specimens fed	61
Specimens graded and committed to the PLC	61
Specimens missed (no grade emitted)	0
Reached the correct lane	60 (98.36%; 95% CI 91.3–99.7%)
Wrong lane: wrong grade, adjacent rank/two or more ranks	1/0
Wrong lane: correct grade, mistimed, or wrong-specimen actuation	0
Image-level agreement on the same specimens (Section 3.4)	60 (98.36%)

3.5. Cross-Generation Summary

Table 18 reports the same metrics for every generation and every component, computed from the re-run, and Table 19 shows the per-class performance of each deployed configuration. Three regularities hold across generations. The rule alone is always the weakest component, by 13 to 46 points, and, in G1, G2, and G4, the difference from the CNN is significant on the paired test. The rank-aware metrics move with the top-1 for the CNNs, whose errors are almost all to an adjacent rank (off-by-one accuracy 95.5–100% on their own splits), but not for the rules, whose errors reach two ranks in every generation. Moreover, no ensemble is significantly better than its best member.

Figure 24 places the four generations on a common accuracy–throughput plane.

Table 18. Common metrics for every generation and component, each on its own test split, from the re-run of the delivered models. Abst.: rule abstentions, counted as errors in top-1 and off-by-one. Off-1: share of specimens graded within one price rank. MARE: mean absolute rank error over graded specimens. κ: Cohen's κ over five categories; κ_q: quadratic-weighted κ over four price ranks. F1: macro-F1 over five categories. The last block applies the delivered G3 pipeline to the G4 test split (Section 3.3.6).

Gen.	Component	Correct/*n*	Acc. (%) [95% CI]	Abst.	Off-1 (%)	MARE	κ	κ_q	F1
G1	Deployed ensemble (vote)	88/91	96.70 [90.8, 98.9]	0	100.0	0.022	0.957	0.984	0.963
	ConvNeXt-T	87/91	95.60 [89.2, 98.3]	0	98.9	0.055	0.943	0.946	0.942
	ResNet-50	86/91	94.51 [87.8, 97.6]	0	100.0	0.033	0.928	0.977	0.946
	Rule	75/91	82.42 [73.3, 88.9]	1	93.4	0.200	0.782	0.738	0.812
G2	Deployed (=ConvNeXt-T)	85/91	93.41 [86.4, 96.9]	0	97.8	0.088	0.914	0.906	0.923
	Rule	50/91	54.95 [44.7, 64.8]	2	84.6	0.517	0.420	0.339	0.460
G3	Deployed ensemble (vote)	21/22	95.45 [78.2, 99.2]	0	100.0	0.045	0.941	0.981	0.942
	ConvNeXt-S	19/22	86.36 [66.7, 95.3]	0	95.5	0.182	0.824	0.875	0.831
	MobileNetV3	19/22	86.36 [66.7, 95.3]	0	95.5	0.182	0.821	0.893	0.855
	Rule	15/22	68.18 [47.3, 83.6]	3	72.7	0.368	0.734	0.637	0.683
G4	Deployed ensemble; ConvNeXt-T; EfficientNet-B0 (identical outputs)	60/61	98.36 [91.3, 99.7]	0	100.0	0.016	0.979	0.992	0.980
	Rule	32/61	52.46 [40.2, 64.5]	0	90.2	0.525	0.398	0.605	0.494
G3 on G4	Deployed G3 ensemble	20/61	32.79 [22.3, 45.3]	0	95.1	0.705	0.178	0.503	0.349
	ConvNeXt-S	26/61	42.62 [31.0, 55.1]	0	96.7	0.607	0.306	0.606	0.405
	MobileNetV3	13/61	21.31 [12.9, 33.1]	0	85.2	0.885	0.028	0.203	0.214
	Rule	8/61	13.11 [6.8, 23.8]	22	47.5	0.974	0.029	0.148	0.132

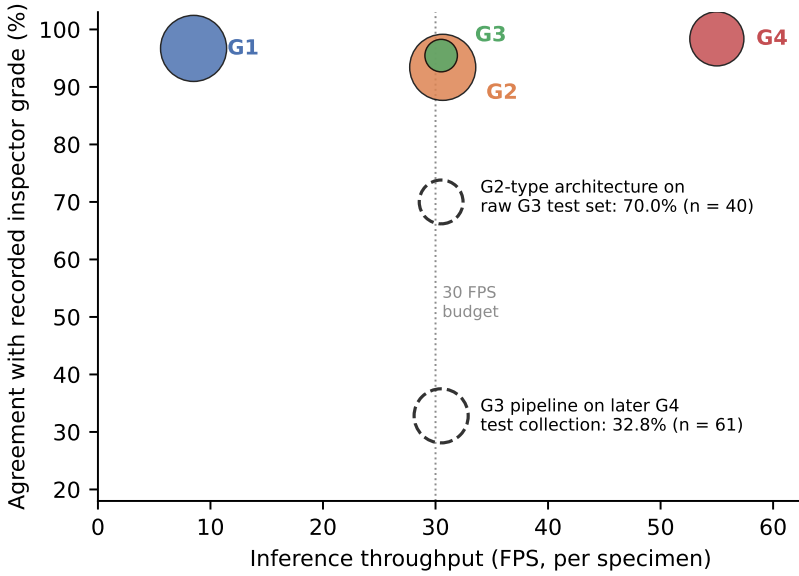


Figure 24. Accuracy and throughput across the four generations; marker area is proportional to the size of the test set that each was measured on. No line joins the generations, because each was evaluated on different data. The hollow markers are the two like-for-like re-evaluations on later data: the G2-style architecture on raw G3 data (70.0%, $n = 40$) and the delivered G3 pipeline on the G4 test split (32.8%, $n = 61$), both plotted at the throughput of the architecture evaluated. FPS is inference throughput (Section 2.9).

Table 19. Per-class precision/recall of each deployed configuration on its own test split (support in parentheses).

Grade	G1 (91)	G2 (91)	G3 (22)	G4 (61)
Grade 1	0.91/1.00 (10)	0.82/0.90 (10)	1.00/1.00 (4)	1.00/0.91 (11)
Grade 2	1.00/0.94 (17)	0.94/0.94 (17)	0.83/1.00 (5)	0.90/1.00 (9)
Grade 3 (opened cap)	0.95/0.95 (20)	0.95/0.95 (20)	1.00/1.00 (3)	1.00/1.00 (13)
Grade 3 (growth-arrested)	0.97/1.00 (31)	0.94/0.97 (31)	1.00/0.67 (3)	1.00/1.00 (14)
Off-grade	1.00/0.92 (13)	1.00/0.85 (13)	1.00/1.00 (7)	1.00/1.00 (14)
Macro-F1	0.963	0.923	0.942	0.980

4. Discussion

4.1. Our Own Numbers Are Not a Progress Curve

The four headline accuracies rise, fall, and rise again—96.70, 93.41, 95.45, 98.36—and it would be easy to present them as a development trajectory. They are not one. They were measured on four different test sets of 91, 40 (or 22 filtered), and 61 specimens, under two different camera configurations, with curation policies that grew progressively more aggressive. Because the test sets differ, their 95% intervals (Table 3) are not a formal basis for comparing generations at all, but their width—8 to 21 points each, all four overlapping—illustrates that sample sizes of this order would not support a claim of monotonic improvement even on a common test set. The one measurement that is comparable across generations is the re-evaluation of the G2 architecture on G3 data: above 90% became 70.0%, a separation that is far larger than either interval. The forward evaluation added in this revision is a second one, and it is larger: the delivered G3 pipeline, 95.45% on its own filtered split, agreed with the inspectors on 32.8% (95% CI 22.3–45.3%) of the G4 test split (Section 3.3.6). Each generation's headline is also optimistic for a reason that is internal to it—thresholds fitted on the test set in G1, checkpoints selected on the test set in G3 and G4 (Section 2.5)—so the headlines are best read as upper bounds on four different tasks.

We report this prominently because the alternative is a familiar and misleading genre. A system that scores 96.70% on the data that it was developed with has not demonstrated that it will grade next season's harvest at 96.70%, and our own sequence shows the size of the gap. For deployment work in agriculture, where the material varies by season, region, storage, and handling, we would argue that a held-out split from the development collection is a weak claim and that the informative experiments are re-evaluations on data collected later and elsewhere. This is the agricultural instance of a general observation: even a fresh test set drawn by the original protocol costs ImageNet classifiers several points [21], and, in agricultural imaging, where the development collection is small and the field varies by season and handling, the loss is routinely far larger [22,23].

4.2. Domain Shift Damaged Both Branches, but Only One Visibly

The frozen–thawed failure is the most transferable finding here, and the re-analysis changes how it should be stated. Freezing changed the surface texture and colour; the classification branch kept producing plausible grades; the segmentation branch visibly lost masks. Measured offline on the G4 test split (Section 3.3.6), the masks were usable for 39 of 61 specimens, while the G3 classifiers fell from 86.4% to 42.6% and 21.3%. The branch that failed visibly was the less damaged one; the branch that kept emitting grades was the more damaged one, and nothing in its output said so. A hybrid system built precisely so that a learned component and a measured component could check each other lost much of the measured component, and it did so without signalling it.

The general lesson is that branch-level robustness is not uniform, that a branch that abstains is easier to monitor than one that does not, and that a pipeline should therefore

distinguish its abstention reasons. "No measurement, because this specimen is truncated" is a routine condition; "no measurement, because segmentation has stopped working across the whole batch" is an alarm. A monitor that distinguishes them is lightweight relative to the vision pipeline—a running rate of mask production failures, compared with a threshold—and would have converted a silent degradation into an operator-visible one; its threshold, false alarm behaviour, and PLC interlock policy would still need validation before it could be trusted on the line.

4.3. Toward Failure-Aware Operation

G4 did not implement this monitor, and the second field trial exposed a related gap when the empty belt was graded as a specimen. We state here, as a design rather than a result, what a failure-aware version of the pipeline should contain, because the G3 incident shows that the absence of these components is not a minor omission but the condition under which the system's central guarantee lapses without notice.

Failure Modes and Their Frequencies

Table 20 lists the failure modes encountered across the four generations, the frequency with which each was measured, and its effect on the grade sent to the rejector. Frequencies are from the re-run on stored test splits; the line-level modes were observed at the trials but not counted, and they are counted only where the table cites the belt experiment.

The table makes plain that the failure modes that change a specimen's lane silently—classifier shift and correlated CNN error—are the ones that the present system cannot detect, while the one that it detects most reliably, a missing mask, has no effect on the lane at all.

The G3 failure would have been caught by a segmentation quality monitor: a running rate of frames in which no cap or no stipe mask is produced, together with the distribution of mask confidence, of the cap and stipe areas, and of the area ratio ρ, each compared with its range on the training collection. The frozen–thawed material shifted every one of these at once. Cheaper still is an input quality monitor—per-camera brightness, saturation, sharpness, and background occupancy and the timestamp skew between the three frames—which catches lens condensation, illumination drift, a dirty enclosure window, and a dropped camera before they reach the models. Between the two sits a drift detector on the appearance branch itself, such as the distance of a fused descriptor from the training-set embedding distribution, which is the only one of the three that would notice a shift that leaves masks intact but moves the classifier away from the data that it was trained on.

The empty belt error is a different failure—one of scope rather than of robustness: the pipeline was never given a way to say that there is nothing to grade. Any of a foreground occupancy gate on the imaging region, a background subtraction check against the empty belt, a lightweight detector ahead of the segmentation model, or a photo-interrupter at the cabinet entry would close it, and the latter is the cheapest and the most reliable, because it does not depend on a learned model behaving under conditions that it was not trained on.

These monitors are only useful if their outputs change what the system does. We propose four operating states in place of the present two: a normal grade, committed to the PLC; a low-confidence grade, committed but flagged on the interface, for the case where the two CNNs disagree and the rule has no veto; a request for manual inspection, with no PLC commit, for the case where a monitor is out of range on the current specimen; and a line alarm, with grading suspended, for the case where a monitor has been out of range across a batch. The G3 incident belongs in the fourth state and was handled in the first. The ensemble's disagreement signal, whose accuracy value Section 4.5 finds thin, is exactly the input that the second state needs, which is the better justification for keeping it.

Table 20. Failure modes, measured frequencies, and effects on specimen-level sorting. "n.c.": observed but not counted.

Failure Mode	Frequency	Effect on the Grade Sent to the Rejector	Detection and Mitigation
No usable mask (truncation, missing cap or stipe)	G1 1/91; G2 2/91; G3 3/22; G4 0/61; G3 segmentation on G4 split 22/61 (stipe lost 21, cap lost 1)	Rule abstains; the grade comes from the CNNs. In G2–G4, the output is unchanged, because the rule can only alter a grade when the CNNs disagree	Rate of abstentions by reason, against its training-set range; line alarm when it exceeds a threshold
Classifier shift (later collection, frozen–thawed material)	G3 pipeline on G4 split: 41/61 wrong, no abstention; G3 ConvNeXt-S 35/61 wrong	Confident wrong grade; wrong lane	CNN disagreement flagged 19 of 41 errors (and 13 correct specimens); embedding drift monitor; periodic re-grading of a sample by an inspector
Correlated CNN error	G4: 1/61 (Grade 1 read as Grade 2 by both); belt experiment: the same specimen, sent to the Grade 2 lane (1/61); G3 on G4: 22 of 41 errors with both CNNs agreeing	Wrong lane, unflagged; adjacent rank in G4	Not detectable by agreement; cap opening estimator for the broad-cap case
Indicator implementation drift	G2 thickness vs G1: $r = -0.32$; 3/91 grades changed across 20 repeated runs	None on the output (arbitration), but the displayed thickness is unreliable	Deterministic estimator; caliper validation
Empty belt graded as a specimen	G4 trial: n.c.; belt experiment: n.c.	A grade without a specimen is sent to the PLC; whether the rejector then acts on the wrong specimen depends on the PLC's actuation scheduling, which was not characterised	Entry photo-interrupter or occupancy gate before grading
Tracking error (missed association, duplicate, or lost track)	Not measured; the recorded trial video shows one to two specimens in view at a time	A duplicate track emits two grades for one specimen; a lost track either re-grades or drops a specimen	Log of track births, deaths, and grades per specimen, reconciled with an entry count
Late decision	Belt experiment: 0 of 61 specimens diverted late or on another specimen's grade; chain not timed	The grade reaches the PLC after the specimen has passed the rejector	Timing chain measurement (below)

Finally, the latency figure of Section 3.4 stops at the classifier output. What the rejector needs is the interval from the specimen's entry into the cabinet to the flip-bar actuation and the position error that this interval produces at the belt speed; the chain passes through the tracker, the browser interface, and the serial link and was not timed. Writing the chain out makes the requirement explicit. With t_{acq} being the time from the specimen becoming fully visible to the frame being in memory, t_{inf} being the inference time, t_{ui} and t_{ser} being the interface and serial transmission, t_{plc} being the PLC scan, and t_{act} being the flip-bar response, the grade arrives in time if

$$t_{\mathrm{acq}} + t_{\mathrm{inf}} + t_{\mathrm{ui}} + t_{\mathrm{ser}} + t_{\mathrm{plc}} + t_{\mathrm{act}} < \frac{d}{v}, \tag{8}$$

where d is the distance from the point of grading to the rejector and v is the belt speed, and the jitter σ_t of the left-hand side displaces the actuation by $v\,\sigma_t$, which must stay within the rejector's positional tolerance. Only the inference throughput of the complete pipeline on stored triples was measured (18.2 ms per specimen on the G4 console, 33.2 ms on the RTX 3060; Table 8), which stands in for t_{inf}; on the line, reading a stored image is replaced

by t_{acq}. In the belt-level experiment of Section 3.4.2, none of 61 specimens fed one at a time was diverted at the wrong time or on another specimen's grade; the single wrong-lane specimen followed its own, incorrect grade (Table 17), which is a grading error rather than a timing error. This shows that the condition held on the laboratory sorter at this spacing; the terms themselves were not time-stamped, so the margin is unknown. At the centre's belt speed, spacing, and geometry, the chain must be timed term by term. Reporting it, with the belt speed and the rejector's positional tolerance, is what transforms a per-specimen inference time into a claim about the line.

4.4. Deployment Engineering

Two results run against common expectation. First, thread-level parallelism across models is not additive capacity: at batch 1 on a single GPU, three networks contended so severely that overlap gained 3.8% where the ideal predicts 34%. Reducing the model count was the effective response and should be considered before scheduler tuning. Second, mixed-precision policy is per-model rather than per-pipeline: FP16 was free for the convolutional detectors and unacceptable for the LayerNorm-based classifiers in the same system, twice, in two independent generations. We have not established why—the LayerNorm/GELU range explanation of Section 3.2.2 is the natural hypothesis, not a finding—and the practical recommendation does not depend on it: verify the FP16 engine of every network against its FP32 reference on the task metric before deploying it, and treat a mixed build with normalisation layers pinned to FP32 as the first step to try when a network fails this check. Batch-1-specialised FP32 compilation recovered 23.5% of classifier latency at no accuracy cost, which was enough.

The tracking result is more mundane but was worth more operationally than either. Discarding frames containing a truncated specimen is correct per frame and wrong per specimen: on a belt, a specimen leaving the field of view is truncated in every remaining frame, so it silently receives no grade at all. Moving state from the frame to the tracked object fixed a systematic loss at the conveyor exit and, as a by-product, made the grading cost per-specimen rather than per-frame.

4.5. What the Ensembles Actually Bought

Every generation ensembled, and, across the four, the evidence that ensembling raised the accuracy is thin; the re-run makes it thinner. G1's three-member vote exceeded its best single member by one specimen ($p = 1$). G2's two-branch arbitration is output-equivalent to its appearance branch by construction and was identical to it on all 91 specimens; the 2.19-point drop from G1 that the submitted version attributed to the arbitration comes from the input resolution (Section 3.2.3). G3's vote corrected two ConvNeXt-S errors and introduced none ($p = 0.5$) with the delivered checkpoints; the case in which two correlated errors outvoted the correct prediction occurred during development. G4's two CNNs returned identical grades on all 61 specimens, and the rule never altered a grade (Section 3.4.1).

We nonetheless kept an ensemble in every generation, and we believe that this was appropriate—but for reasons that are not accuracy-related. Two independent classifiers plus a measurement branch produce an agreement signal that the operator can act on, a constrained veto, and a fallback when one branch stops working, which is exactly the situation that Section 3.3.5 describes. What the sequence does argue against is the reflex of reporting an ensemble gain as the contribution. On test sets of this size, the gain is usually within one or two specimens of nothing, and the honest justification for the added cost is redundancy rather than accuracy. We therefore describe the ensembles as redundancy mechanisms, not as accuracy improvements.

If this is the justification, then the top-1 accuracy is the wrong metric on which to compare ensemble configurations, and the comparison should be scored on what redundancy buys: the fraction of errors on which the members disagree, which is the fraction that a disagreement flag could route to manual inspection; the accuracy of the ensemble when one member is withheld, which is its resilience to the G3 failure; the calibration of its confidence; and the latency that each member adds. The re-run provides the first two. On their own splits, the members almost never disagree: in G4, they do not agree once, so the flag caught none of the one error; in G3, it occurred on 4 of 22 specimens. On the later material of the forward evaluation, they disagreed on 32 of 61 specimens, and the flag caught 19 of the 41 errors, at the cost of flagging 13 correct grades. Withholding any member leaves the G4 output unchanged. On this evidence, the redundancy is real but conditional: it is invisible on an internal split and becomes useful only under the shift that it is meant to guard against—which is an argument for evaluating it on shifted data and for the G4 members, which are near-copies of each other on the internal split and one of which sees only one camera, to be re-examined on such data.

4.6. Interpretability Eroded Under Deployment Pressure

The system's original justification was that its intermediate quantities were physical: an inspector could check "stipe thickness 2.15 cm" with a caliper. This property was weakened generation by generation. G1 reported centimetres against a calibrated scale. G4 reports the pixel-domain thickness, because no scale calibration was carried to the three-camera rig, and its primary discriminator is a cap area ratio—this is robust but further from anything that an inspector measures by hand than the ellipse ratio that it replaced. Each individual change was locally justified by robustness. The cumulative effect is that the interface now shows numbers that the operator cannot independently verify.

We flag this as a design failure mode rather than a solved problem, because it happened without anyone deciding that it should. A system justified by auditability should treat its calibration as a requirement to be maintained across hardware changes, not as a convenience to be dropped when the rig changes. Restoring a scale calibration for the three-camera rig, and reporting the thickness in centimetres again, would cost little and would restore the property that the design was built on. The attention weights introduced in G4 were meant to partly compensate—an operator could at least see which camera drove a decision—but they explain the model, not the mushroom; in the delivered models, they explain little even of the model, being nearly constant in one network and collapsed onto one camera in the other (Section 3.4.1). A second erosion was invisible until the re-analysis: the G2 re-implementation of the thickness estimator produced values that did not agree with the G1 implementation on the same images (Section 3.2.3), and, because the G2 arbitration never allows the rule to change a grade, nothing downstream exposed it.

Calibration alone is not enough, because a number in centimetres is only auditable if it is also correct, and no generation checked this. The claim that the system measures the stipe thickness rests entirely on the construction of the estimator in Section 2.6; a reviewer, or a collector disputing a grade, is entitled to ask for the evidence. The experiment is small: 30 to 50 specimens spanning the grades, each measured by a caliper at the mid-stipe and by the rig from each camera, would give the mean error and bias of the thickness estimate, its agreement limits in the form of a Bland–Altman plot [43], and its repeatability across the three cameras and across repeated placements on the belt. The same specimens, with an inspector's estimate of the fraction of the cap opened, would show how well the area ratio ρ—which the standard does not mention—tracks the quantity that the standard does. Until this table exists, the safer description of the morphometric branch is that it produces

consistent, inspectable numbers that the inspectors have tuned thresholds against, not that it produces measurements.

4.7. Practical Value at the Collection Centre

A system designed for repeatable grading could reduce the operational effect of the inter-inspector variability that collectors experience as unfairness, produces a record attachable to each transaction, and shortens the interval between arrival and dispatch for a product whose value decays within days—provided that its agreement with a multi-inspector reference and its stability across seasons are validated, which Section 4.8 sets out and this paper has not done. With PLC integration, the result now drives physical sorting rather than a screen. The adjustment interface is what makes this durable: a grading system whose boundaries only its developers can change will drift out of agreement with commercial practice and be abandoned, whereas one that a centre manager can re-calibrate in a morning stays in use. Seasonal flexing of grade boundaries is a requirement to support, not a defect to engineer away.

4.8. What Would Be Needed to Make a Commercial Claim

The argument of this paper is that internal held-out accuracy under-specifies a deployed grader. It would be inconsistent to end on an internal held-out accuracy. We therefore state the evaluations that, in our judgement, a claim of commercial readiness would require, in the order in which they would most change what is known; the belt-level experiment of Section 3.4.2 covers part of the operating metrics on a laboratory sorter; none of the others has been performed, and the next harvest season is the first opportunity for most of them.

An independent test collection. Specimens collected in a later season, on a different date, and ideally at a second centre or on a second rig, graded before any model sees them and never used for training, threshold tuning, or curation decisions. The G3 re-evaluation (Section 3.3.1) is the only experiment of this kind in the paper, and it is the one whose result mattered most; a few hundred specimens would give an interval that is narrow enough to be useful.

An uncurated belt set. A run of consecutive specimens taken from the line as it operates, with nothing removed, on which the system's grade, its abstentions, and its monitor states are all recorded. The accuracy on this set, reported together with the abstention rate, is the operating figure; the curated figure is the ceiling.

An inter-rater study. Two or three inspectors grading the same specimens independently and blind, with Fleiss' κ or pairwise weighted κ [39,44] reported alongside the system's agreement with each and with their consensus. This sets the ceiling that any grader, human or automatic, can reach on this task and identifies the borderline specimens, which should then be analysed as a subset rather than removed. Because adjacent errors cost less than distant ones, the same study should report a cost-weighted confusion matrix under a price matrix agreed with the centre, so that a Grade 1-to-Grade 2 error and a Grade 1-to-off-grade error are not counted alike.

Measurement validation. The caliper study of Section 4.6, on the calibrated rig.

Operating metrics. From a trial in which every specimen is logged: the fraction of specimens that receive a grade, the fraction diverted to the correct lane, the empty-belt false trigger rate, and the trigger-to-actuation latency with its variance, against the belt speed and the rejector's positional tolerance.

A model selection protocol. A three-way split by collection session—training, validation for checkpoint and threshold selection, and a test partition touched once—with each

model trained under several random seeds and the spread across seeds reported. None of the generations had a validation partition (Section 2.5).

An ensemble study. Every pair among the screened backbones, trained under the protocol above, scored on the redundancy metrics of Section 4.5 on shifted data and not only on the top-1.

Only the first of these would change the architecture; the rest would change what we are entitled to say about it. We list them because the alternative—publishing the 98.36% figure and leaving the reader to discount it—is the practice that this paper argues against.

4.9. Limitations and Generalizability

Test set sizes: 91, 40, 22, and 61 specimens. At $n = 61$, a single error moves the accuracy by 1.6 points, and the 95% intervals in Tables 3 and 14 are 8 to 21 points wide. No accuracy in this paper should be read to two decimal places, our own tables notwithstanding; the intervals are the honest statement. No delivered configuration was retrained with different seeds (development runs are not seed repeats; Section 2.5), so the intervals also omit training variance, which, at 235 training specimens, is unlikely to be small.

Model selection on the test set: The G3 and G4 checkpoints were selected at the epoch of the best test accuracy, and the G1 rule thresholds were set from the test-set distribution, with no separate validation partition. Every headline is optimistically biased by an amount that we cannot estimate.

Session-level leakage: Splits were drawn at the specimen level within each pooled collection, so collection sessions are shared between training and test partitions (Section 2.4), and session-specific conditions may have helped the classifiers.

Storage condition unrecorded: The fresh and frozen–thawed specimens in G4 cannot be separated, so accuracy by storage condition is unknown.

Curated test sets: G3 and G4 test sets had outliers and label-ambiguous specimens removed. This is defensible for training and questionable for evaluation, and it inflates the later figures relative to belt conditions by an unmeasured amount. No uncurated set was held back, so the inflation cannot be estimated from our data.

Single-annotator ground truth: Every reference grade is one inspector's judgement, recorded once. Inter-inspector agreement on this task is not known, so it is not known whether the residual errors of the system are errors at all or how far above 98% agreement with one inspector it is possible in principle to go. Label-ambiguous specimens were removed from the G4 test set rather than adjudicated, which removes precisely the cases on which inspectors would disagree.

Divergence from the official standard: The deployed rule does not use the stipe length, the standard's primary criterion, and uses instead two proxies that the standard does not name. The proxies were chosen for discriminative power on our data and their thresholds are set by the centre's inspectors; we claim operational agreement with these inspectors, not equivalence to the standard, and neither has been formally tested.

No measurement validation: The system reports image-derived indicators, but they were never validated against calipers or against inspector estimates of cap opening in any generation. The claim that these are measurements rather than plausible-looking numbers rests on the estimators' construction alone, and the disagreement between the G1 and G2 thickness implementations shows that this construction is not enough.

Limited operating-level evaluation: The field trials recorded no specimen counts. The belt-level experiment (Section 3.4.2) measured the correct lane rate (60/61) and missed specimens (0/61), but on the 61 specimens of the G4 test split, in a single run, on one laboratory installation, without timing or empty-belt measurements; it does not establish

performance on unseen specimens, at the centre's working throughput or across seasons. The 55 FPS figure is the inference throughput and not the line throughput.

FP16 cause not isolated: The reduced-precision failure of the ConvNeXt classifier was observed under one build configuration and its cause was not diagnosed; the LayerNorm/GELU explanation is a hypothesis consistent with the observation, and alternative explanations in the export and build path were not excluded.

Re-analysis rather than original records: The per-class, rank-aware, and ablation results come from re-running the delivered checkpoints in PyTorch, not from the deployed TensorRT engines; the headline counts were reproduced exactly, but, for G1, G2, and G4, the engines themselves were not re-run.

Data leakage in an earlier G4 split: A preliminary image-level split placed duplicate photographs of the same specimen on both sides of the partition and scored 100%. It was caught and the split was re-performed at the specimen level, but we note it because the same defect is easy to introduce in any dataset where specimens are re-imaged, and it would not have been detected by accuracy alone.

Ensemble composition not selected systematically: The deployed G4 pair was fixed before the backbone screening, no other pair was trained, and the two members returned identical grades on the test split, so we cannot say whether another pair would perform better. Combined with the observation of Section 4.5 that none of our four ensembles demonstrably beat its best single member, the composition of the final ensemble should be treated as an engineering choice that we have not validated.

Single site: All data come from one conveyor and one collection centre. Cross-site generalisation is untested, and the G3 result suggests that we should expect a drop.

Single-specimen assumption: Touching or overlapping specimens remain out of scope, although tracking and instance segmentation together provide the natural mechanism for extending to them.

The architecture—a learned appearance branch over several views, a part-segmentation branch producing measurable indicators, and an editable rule reconciling them—is not specific to pine mushroom. It suits any produce whose grade specification is stated morphologically, which covers most of the non-timber forest products handled by the same centre. Porting requires new segmentation annotations and a new threshold table, not a new architecture. The deployment findings—GPU contention at batch 1, per-architecture precision policy, per-specimen rather than per-frame state, and asymmetric branch robustness under domain shift—are properties of the deployment pattern rather than of the crop.

5. Conclusions

We have described four development cycles of an automated grading system for pine mushroom, from a two-camera bench dataset to a three-camera conveyor installation with PLC-driven physical sorting. The final system is a single-site prototype whose end-to-end operation has been demonstrated: it grades 60 of 61 specimens of a curated internal held-out set correctly (98.36%, 95% CI 91.3–99.7%) at 55 FPS inference throughput, a figure that is biased upwards by checkpoint selection on the same set. Fed through a laboratory conveyor sorter, the same 61 specimens reached the lane of their inspector grade in 60 cases, so the line reproduced the image-level result. Its two classifiers agree on every test specimen, its fusion weights are nearly fixed, and its rule changed no grade, so the figure is, in effect, that of a single ConvNeXt classifier. It has not been evaluated on an independent later collection, against more than one inspector, on an uncurated belt run, or with its measurements checked against calipers, and we do not claim that it is ready for commercial use. The more useful results are the ones obtained on the way there.

A threshold rule on image-derived indicators, fitted on the same specimens, agreed with 82.42% of the recorded inspector grades; this figure describes the implemented rule and not the explanatory power of the official standard. The same rule re-implemented for speed fell to 54.95%, because its thickness estimate no longer agreed with the original—an instance of how quickly an unvalidated indicator loses its meaning. Real-time operation was reached by removing a model rather than by parallelising more strongly, because three networks on one GPU at batch 1 contended rather than overlapped. Precision policy proved architecture-dependent under our build conditions, with FP16 free for convolutional detectors and unusable for LayerNorm-based classifiers in the same pipeline, in two independent generations; we have offered a hypothesis for why and have not tested it. A first field trial failed in a way that we had not anticipated: frozen and thawed specimens shifted the domain enough that the segmentation branch visibly lost masks while the classification branch kept emitting grades. Measured offline, the classifiers were the more damaged branch—the G3 pipeline agreed with the inspectors on 32.8% of the later G4 test split—so the system silently degraded into a black box that was also wrong, which is worse than the black box that the design existed to avoid.

Most importantly, our own accuracy figures do not form a progress curve, and we have said so with the evidence: re-evaluated on the next season's data, an architecture that had scored above 90% scored 70.0%, and the delivered G3 pipeline, reaching 95.45% on its own split, scored 32.8% on the next collection. For deployment work on agricultural and forest products, we suggest that a held-out split from the development collection is a weak claim, that re-evaluation on later data collected elsewhere is the informative experiment, and that a system justified by the interpretability of its measurements must defend this interpretability actively—because ours eroded, one locally reasonable decision at a time, until the numbers on the operator's screen were no longer ones that the operator could check.

The work that follows is set by Section 4.8. In the next harvest season, we intend to hold back an independent test collection and an uncurated belt run, graded before any model sees them; to have the same specimens graded blind by two or three inspectors so that inter-rater agreement bounds what any grader can achieve; to restore scale calibration on the three-camera rig and validate the thickness and cap opening against calipers and inspector estimates; to instrument the centre's line so that the belt-level accuracy, missed specimen rate, false trigger rate, and trigger-to-actuation latency are recorded; to implement the segmentation quality, input quality, and empty belt monitors of Section 4.3 with the four-state operating policy and measure their false alarm rates; to select models and thresholds on a session-level validation partition, with repeated seeds; and to complete the backbone and pair screening so that the deployed ensemble's composition rests on measurement rather than on an early engineering choice. Extension to multiple specimens per frame and to other non-timber forest products processed on the same line follows once these measurements exist.

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

Funding: This research was funded by the R&D Program for Forest Science Technology (Project No. RS-2023-KF002445), provided by the Korea Forest Service (Korea Forestry Promotion Institute), Republic of Korea.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. The re-analysis scripts and the per-specimen predictions of all four generations are available from the corresponding author on request. Further inquiries can be directed to the corresponding author.

Acknowledgments: The four development contracts were carried out with the Department of Information and Communication Engineering, Hanbat National University. The authors thank the inspectors and staff of the collection and distribution centre for grading the specimens and for hosting the field trials.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yamanaka, T.; Yamada, A.; Furukawa, H. Advances in the cultivation of the highly-prized ectomycorrhizal mushroom *Tricholoma matsutake*. *Mycoscience* **2020**, *61*, 49–57.
2. van Gevelt, T. Community-based management of *Tricholoma matsutake* (S. Ito and S. Imai) Singer: A case study of South Korean mountain villages. *Int. J. Commons* **2014**, *8*, 134–154.
3. Kamilaris, A.; Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* **2018**, *147*, 70–90.
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.
5. Ismail, N.; Malik, O.A. Real-time visual inspection system for grading fruits using computer vision and deep learning techniques. *Inf. Process. Agric.* **2022**, *9*, 24–37.
6. Yin, H.; Yi, W.; Hu, D. Computer vision and machine learning applied in the mushroom industry: A critical review. *Comput. Electron. Agric.* **2022**, *198*, 107015.
7. Wu, Y.; Sun, Y.; Zhang, S.; Liu, X.; Zhou, K.; Hou, J. A size-grading method of antler mushrooms using YOLOv5 and PSPNet. *Agronomy* **2022**, *12*, 2601.
8. Wang, L.; Dong, P.; Wang, Q.; Jia, K.; Niu, Q. Dried shiitake mushroom grade recognition using D-VGG network and machine vision. *Front. Nutr.* **2023**, *10*, 1247075.
9. Shui, Q.; Miao, F.; Liu, S.; Cao, L.; Jiang, H.; Lu, J. A fast and real-time machine vision evaluation system for size grading of *Agaricus bisporus* based on video. *Food Sci. Nutr.* **2026**, *14*, e71328.
10. Sculley, D.; Holt, G.; Golovin, D.; Davydov, E.; Phillips, T.; Ebner, D.; Chaudhary, V.; Young, M.; Crespo, J.-F.; Dennison, D. Hidden technical debt in machine learning systems. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 28, pp. 2503–2511.
11. Paleyes, A.; Urma, R.-G.; Lawrence, N.D. Challenges in deploying machine learning: A survey of case studies. *ACM Comput. Surv.* **2022**, *55*, 114.
12. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
13. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8, Version 8.0.0. Available online: https://github.com/ultralytics/ultralytics (accessed on 17 September 2026).
14. Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E. Multi-view convolutional neural networks for 3D shape recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 945–953.
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
16. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the IEEE International Conference on Image Processing, Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468.
17. Jacob, B.; Kligys, S.; Chen, B.; Zhu, M.; Tang, M.; Howard, A.; Adam, H.; Kalenichenko, D. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2704–2713.
18. Micikevicius, P.; Narang, S.; Alben, J.; Diamos, G.; Elsen, E.; Garcia, D.; Ginsburg, B.; Houston, M.; Kuchaiev, O.; Venkatesh, G.; Wu, H. Mixed precision training. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
19. Torralba, A.; Efros, A.A. Unbiased look at dataset bias. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 1521–1528.

20. Wang, M.; Deng, W. Deep visual domain adaptation: A survey. *Neurocomputing* **2018**, *312*, 135–153.
21. Recht, B.; Roelofs, R.; Schmidt, L.; Shankar, V. Do ImageNet classifiers generalize to ImageNet? In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 5389–5400.
22. Barbedo, J.G.A. Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. *Comput. Electron. Agric.* **2018**, *153*, 46–53.
23. Hu, X.; Chen, S.; Duan, Q.; Ahn, C.K.; Shang, H.; Zhang, D. Domain adaptation for big data in agricultural image analysis: A comprehensive review. *arXiv* **2025**, arXiv:2506.05972.
24. Ahn, H.; Kim, Y.; Jee, S.C.; Kang, J.-S.; Chung, H.-J.; Choi, D.-G. Artificial intelligence-based pine mushroom grading automation system. *J. Korean Next Gener. Comput. Soc.* **2025**, *21*, 7–19. https://doi.org/10.23019/kingpc.21.4.202508.001. (In Korean)
25. Ahn, H.; Jang, H.; Jee, S.C.; Kim, J.; Chung, H.-J.; Choi, D.-G. Automated matsutake quality grading system using dual-camera multi-view analysis and morphological feature ensemble. *J. Korean Next Gener. Comput. Soc.* **2026**, *22*, 85–100. https://doi.org/10.23019/kingpc.22.3.202606.006. (In Korean)
26. Wang, A.; Liu, L.; Chen, H.; Lin, Z.; Han, J.; Ding, G. YOLOE: Real-time seeing anything. *arXiv* **2025**, arXiv:2503.07465.
27. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
28. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
30. Tan, M.; Le, Q.V. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
31. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; Le, Q.V.; Adam, H. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
32. 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.
33. Fitzgibbon, A.; Pilu, M.; Fisher, R.B. Direct least square fitting of ellipses. *IEEE Trans. Pattern Anal. Mach. Intell.* **1999**, *21*, 476–480.
34. Felzenszwalb, P.F.; Huttenlocher, D.P. Distance transforms of sampled functions. *Theory Comput.* **2012**, *8*, 415–428.
35. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157.
36. Wilson, E.B. Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.* **1927**, *22*, 209–212.
37. Brown, L.D.; Cai, T.T.; DasGupta, A. Interval estimation for a binomial proportion. *Stat. Sci.* **2001**, *16*, 101–133.
38. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46.
39. Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* **1968**, *70*, 213–220.
40. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
41. Siméoni, O.; Vo, H.V.; Seitzer, M.; Baldassarre, F.; Oquab, M.; Jose, C.; Khalidov, V.; Szafraniec, M.; Yi, S.; Ramamonjisoa, M.; et al. DINOv3. *arXiv* **2025**, arXiv:2508.10104.
42. van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
43. Bland, J.M.; Altman, D.G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1986**, *327*, 307–310.
44. Fleiss, J.L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **1971**, *76*, 378–382.

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.