

Fourier-Guided Multi-Scale Vision Transformer for Object-Independent Detection and Severity Grading of Rain-Induced Cracking in Sweet Cherry

Hoa-Cuc. Nguyen

Thu Dau Mot University

Ngoc-Duong. Nguyen

Thu Dau Mot University

Viet-Hung. Pham

Thu Dau Mot University

Bich-Ngoc. Mach

Thu Dau Mot University

Trinh Thi Nhu Quynh

Thu Dau Mot University

Thanh Q. Nguyen

thanhngq@bvu.edu.vn

Baria Vungtau University

Research Article

Keywords: Sweet cherry, rain-induced cracking, Fourier-guided attention, Vision Transformer, crack segmentation, severity grading

Posted Date: September 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10872311/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Rain-induced cracking is among the most damaging preharvest disorders of sweet cherry because even narrow cuticular fractures reduce fresh-market value and create entry sites for water loss and fungal infection. Conventional inspection is subjective and is least reliable for low-contrast microcracks near the pedicel cavity. This study developed a Fourier-guided multi-scale Vision Transformer (FGM-ViT) for joint fruit-level severity grading and pixel-level crack segmentation. The dataset comprised 960 individual fruit from four commercial cultivars and eight orchard blocks, imaged from four rotational views after natural rainfall exposure or controlled water-immersion challenge. Crack severity was assigned as intact, microcrack, moderate, or severe using stereomicroscopic measurements of cumulative crack length, maximum width, affected surface area, and crack location. All views of an individual fruit and all observations from the same orchard block were constrained to the same fold in a nested five-fold grouped evaluation. FGM-ViT combined a multi-scale spatial encoder with luminance-normalised Fourier residual tokens and frequency-guided cross-scale attention. The complete model achieved $93.8 \pm 0.9\%$ fruit-level accuracy, $93.7 \pm 1.0\%$ macro-F1, and 0.842 ± 0.018 Dice coefficient for crack segmentation. It exceeded the spatial-only transformer by 2.2 percentage points in accuracy and 2.2 points in macro-F1, with the largest gain observed for microcracks. Classification errors were restricted almost entirely to adjacent severity categories. Under pedicel overlap, glare, brightness shifts, and motion blur, FGM-ViT retained a 2.7–3.5-point advantage over the spatial-only model. Frequency-band occlusion indicated that mid-to-high spatial frequencies carried complementary evidence for narrow fractures, whereas spatial features remained essential for crack location and marketability interpretation. The dual-output framework provides an auditable, low-cost RGB approach for sweet-cherry sorting, cultivar screening, and rain-cracking phenotyping.

1. Introduction

Sweet cherry (*Prunus avium* L.) commands a high fresh-market value because of its colour, flavour, texture, and seasonal scarcity, yet its commercial return is highly vulnerable to rainfall near harvest. Water exposure can initiate microscopic cuticular failure, enlarge existing fractures, and produce visible cracks at the pedicel cavity, styler end, or fruit cheek. Once the skin barrier is disrupted, fruit lose visual quality, become more susceptible to microbial invasion, and tolerate handling and cold-chain transport poorly. A rain event occurring shortly before harvest may therefore transform a premium crop into fruit suitable only for processing or rejection.

Cracking is governed by a complex interaction among cultivar, cuticle properties, fruit growth, surface wetness, osmotic processes, orchard microclimate, and preharvest management. Correia et al. (2018) described localised cell bursting and progressive weakening of epidermal and hypodermal tissues as central elements in macroscopic crack development. Michailidis et al. (2020) further demonstrated that follow-up testing methods are required because cracking susceptibility and crack progression are not captured adequately by a single end-point index. More recent syntheses have emphasised the strong genetic component of susceptibility and the need to connect physiological understanding with cultivar selection and mitigation strategies (Xu et al. 2025). Rain covers, irrigation management, calcium or silicon treatments, crop-load regulation, and resistant cultivars can reduce risk, but evaluation still depends heavily on visual inspection.

Commercial crack grading is difficult for three reasons. First, the transition from an intact cuticle to a marketability-limiting defect is continuous. A narrow microcrack may be almost invisible under diffuse light yet expand during subsequent wetting or handling. Second, the highly reflective cherry surface produces specular

highlights that resemble fine linear defects, while the pedicel and cavity create dark boundaries that can be mistaken for cracks. Third, crack location modifies commercial significance. A short pedicel-cavity fracture may progress rapidly under water accumulation, whereas an equally long superficial line elsewhere may remain stable. Human graders integrate width, length, depth, location, and appearance, but decisions vary with experience and line speed.

Instrumental quality evaluation includes microscopy, dye penetration, immersion-based cracking indices, firmness testing, and image analysis. These approaches provide valuable physiological evidence but are rarely compatible with every-fruit inspection on a packing line. Machine vision offers a non-contact alternative. Colour and geometric features have been used to classify cherry quality and commercial colour grades (Reyes et al. 2021; Shahedi et al. 2024), while object detectors have supported fruit counting, size phenotyping, and yield estimation (Nagpal et al. 2023). Deep learning is particularly attractive for crack detection because the relevant cue is a narrow, irregular structure whose shape cannot be represented reliably by a small set of handcrafted descriptors.

Convolutional neural networks learn local texture and edge patterns efficiently, whereas Vision Transformers (ViTs) model long-range relations among image patches (Dosovitskiy et al. 2021). Hierarchical transformers such as Swin Transformer retain multiscale information and have performed strongly in fine-grained visual tasks (Liu et al. 2021). Crack analysis nevertheless remains challenging when the defect occupies only a small fraction of the fruit surface. Downsampling can erase the thinnest fracture, class-level supervision may focus on global colour rather than the defect, and illumination changes can alter edge contrast. A model that predicts only a severity class can also be difficult to audit because it does not show which surface structure supported the decision.

Frequency-domain representations provide a complementary view of fine linear damage. A narrow crack introduces oriented high-frequency energy and disrupts the otherwise smoother reflectance pattern of the fruit skin. Fast Fourier Convolution and Global Filter Networks have shown that transformed-domain processing can extend receptive fields and represent global frequency structure (Chi et al. 2020; Rao et al. 2021). However, direct concatenation of a spatial patch and a Fourier coefficient at the same array index is not physically justified: one is indexed by image location and the other by frequency. A scientifically defensible model must preserve separate spatial and spectral coordinate systems and allow their interaction through learned attention rather than assumed positional correspondence.

This study developed a Fourier-guided multi-scale Vision Transformer (FGM-ViT) for joint severity grading and crack segmentation in sweet cherry. The model used a spatial feature pyramid to preserve microcracks, a luminance-normalised Fourier residual branch to emphasise fine discontinuities, and frequency-guided cross-scale attention to modulate spatial features. A segmentation decoder generated an explicit crack-probability map, while an ordinal classification head produced one of four market-relevant severity grades. The segmentation task acted as anatomical supervision and reduced the risk that the classifier relied primarily on cultivar colour or acquisition background.

The study addressed four objectives: (i) quantify object-independent four-grade cracking performance across cultivars and orchard blocks; (ii) determine whether frequency-guided attention improves microcrack recognition beyond spatial CNN and transformer baselines; (iii) evaluate localisation accuracy using stereomicroscopically verified crack masks; and (iv) test robustness under pedicel overlap, specular glare, brightness variation, and

motion blur. By combining an auditable surface mask with a fruit-level decision, the framework was designed to support both high-throughput packhouse grading and research-oriented cracking phenotyping.

2. Materials and methods

2.1. Plant material, orchard sampling, and cracking challenge

The experimental material comprised 960 sweet cherries representing ‘Kordia’, ‘Lapins’, ‘Regina’, and ‘Sweetheart’. Fruit were obtained from eight orchard blocks over two production seasons. Cultivars were distributed across blocks as evenly as orchard availability permitted, and sampling covered comparable commercial maturity. Fruit were harvested with pedicels attached, transferred in padded trays, and imaged within 8 h of collection. Samples with bird damage, advanced fungal decay, or major mechanical cuts unrelated to rain cracking were excluded before the experiment as shown in Fig. 1.

Both naturally exposed and experimentally challenged fruit were included. Naturally cracked samples were collected after rainfall events near harvest. Intact fruit from the same blocks were subjected to a controlled immersion challenge in deionised water at $20 \pm 1^\circ\text{C}$ and removed at scheduled intervals to generate early and progressive damage. Fruit were gently dried before imaging. The mixed design widened the range of crack morphology while retaining orchard-derived variation in cuticle, colour, size, and pedicel geometry.

Each fruit received a permanent object identifier linking cultivar, block, season, exposure route, immersion duration, view angle, and reference measurements. Four rotational views were acquired at approximately 90° intervals, resulting in 3,840 RGB images. Views of a single fruit were used jointly for the fruit-level prediction but were never treated as independent biological replicates.

2.2 Reference crack assessment and severity definition

Reference assessment was conducted immediately after RGB acquisition. The complete surface of each fruit was examined using a stereomicroscope at $10\text{--}30\times$ magnification. Crack paths were traced on calibrated images, and cumulative crack length, maximum visible width, affected surface area, and anatomical location were recorded as shown in Table 1. Crack locations were classified as pedicel cavity, cheek, stylar end, or combined. A subset of ambiguous linear marks was examined after food-grade dye application to distinguish open fractures from reflectance boundaries or superficial abrasions.

Table 1
Object-level composition and reference characteristics of the sweet-cherry cracking dataset

Severity grade	Fruit (n)	Images (n)	Cumulative crack length (mm)	Affected surface (%)	Firmness (N)
Intact	240	960	0.0	0.0	3.82 ± 0.43
Microcrack	240	960	1.8 ± 0.6	0.6 ± 0.3	3.74 ± 0.41
Moderate	240	960	6.9 ± 2.1	2.8 ± 1.1	3.52 ± 0.46
Severe	240	960	15.4 ± 4.6	7.9 ± 2.7	3.18 ± 0.52
Total	960	3,840	—	—	—

Fruit were assigned to four severity grades. Intact fruit showed no open fracture under microscopy. Microcracked fruit contained one or more open fractures with cumulative length below 3 mm, maximum width below 0.10 mm, and no visible tissue separation under unaided inspection. Moderate cracking comprised cumulative length of 3–10 mm or local width of 0.10–0.50 mm without deep flesh exposure. Severe cracking comprised cumulative length above 10 mm, width above 0.50 mm, multiple connected fractures, visible flesh exposure, or a defect judged to cause immediate rejection from the fresh market as shown in Fig. 2. When length and width criteria indicated different categories, the higher-severity category was retained.

Digital crack masks were prepared from the calibrated microscopic and RGB images. Two annotators independently delineated the visible crack path in each RGB view, using the microscopic trace as reference. Disagreement regions were adjudicated by a third assessor. Masks represented the observable crack footprint rather than subsurface cuticular damage. Fruit firmness, equatorial diameter, soluble-solids concentration (SSC), and mass were measured to characterise associated quality differences but were not provided to the image model.

2.3. RGB acquisition and preprocessing

Fruit were imaged inside a matte-black enclosure using a 24-megapixel RGB camera and two cross-polarised LED panels. Cross-polarisation reduced, but did not eliminate, specular reflection from the waxy cuticle. The camera was positioned 32 cm from the fruit centre. Exposure, aperture, focal length, white balance, and illumination geometry were fixed within each session. A metric scale and colour reference were recorded at the beginning and end of each block.

The fruit region was segmented from the background, padded by 4% of its diameter, and resized to 256×256 pixels. Pedicels were retained because pedicel-cavity cracks are commercially important and complete removal of the stem altered the cavity boundary. RGB channels were converted to linear intensity before normalisation. Luminance was estimated using the standard weighted RGB combination and divided by a low-pass illumination field to reduce gradual shading. The normalised RGB crop was supplied to the spatial encoder, while the luminance residual and chromatic channels were supplied to the Fourier branch.

Training augmentations included rotation through the full 360° range, horizontal and vertical reflection, scale variation of 0.90–1.12, translation up to 6%, mild perspective change, brightness variation of $\pm 18\%$, contrast variation of $\pm 15\%$, synthetic glare patches, and motion blur with a kernel length of 3–7 pixels. Crack masks underwent the same geometric transformation as their corresponding images. Photometric augmentation did not modify the masks. Fold-specific means and variances were estimated from training data only.

2.4. Fourier-guided multi-scale transformer

Let the normalised image be (X^{HW}) . The spatial encoder produced feature maps at 64×64 , 32×32 , and 16×16 resolutions. Local convolutional stems preserved narrow edges before tokenisation, and windowed self-attention captured relations among the pedicel cavity, styler end, fruit contour, and cheek surface. Cross-scale lateral projections aligned feature dimensions at 96 channels.

The Fourier branch operated on luminance-normalised residuals rather than raw intensity. For channel (c), the centred log-amplitude spectrum was defined as

$$R_c = X_c - G_\sigma * X_c,$$

1

$$A_c(u, v) = \log [1 + | \text{fftshift}(\mathcal{F}_{2D}\{R_c\})(u, v) |].$$

2

where $(R_c = X_c - G_X c)$ and (G_X) is a Gaussian low-pass kernel. Removing the low-frequency illumination field reduced sensitivity to gradual brightness variation. The spectrum was partitioned into radial-frequency bands and angular sectors, producing a pyramid of spectral tokens. Low-frequency tokens represented broad fruit shading and contour; mid- and high-frequency tokens represented achene texture, narrow reflectance discontinuities, and crack-like edges.

Frequency-guided cross-scale attention used spectral tokens as contextual keys and values for each spatial scale. Spatial tokens remained indexed by image position, while spectral tokens retained frequency-band identities. No index-wise spatial-frequency concatenation was performed. A learned gate controlled the amplitude of the spectral modulation at each spatial scale. The finest scale received the strongest high-frequency contribution; coarser scales received progressively more mid-frequency context as shown in Fig. 3.

Two prediction heads were trained jointly. The segmentation decoder combined spatial features through a feature-pyramid pathway and produced a 256×256 crack-probability map. The severity head pooled the cross-attended tokens and generated four ordered logits. The joint objective was

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ord}} + 0.60\mathcal{L}_{\text{Dice}} + 0.40\mathcal{L}_{\text{focal-mask}} + 0.05\mathcal{L}_{\text{cons}}.$$

3

where the consistency term penalised a severe classification when the predicted mask contained negligible crack area, and penalised an intact classification when a coherent crack mask was present. This constraint improved interpretability without imposing a deterministic area threshold.

2.5. Model training and leakage-controlled evaluation

The model was trained with AdamW, an initial learning rate of 2×10^{-4} , weight decay of 0.04, batch size of 32 images, six warm-up epochs, and cosine decay over 70 epochs. Stochastic depth increased linearly to 0.12 through the encoder, and dropout of 0.15 was applied before the severity head. The checkpoint with the highest inner-validation composite score, defined as macro-F1 plus 0.25 times Dice coefficient, was retained.

A nested five-fold grouped protocol was used. The outer split generated held-out fruit-level predictions, while the inner grouped folds selected learning rate, spectral-gate strength, loss weights, and stopping epoch. All four images of each fruit were assigned to one fold. Orchard block and season were incorporated into the grouping constraint so that views or repeated block-specific backgrounds could not cross between training and testing. Model probabilities were averaged across four views to obtain the fruit-level grade. Crack masks were evaluated at the individual-view level and summarised by fruit before fold averaging.

Classification metrics included accuracy, balanced accuracy, macro-precision, macro-recall, macro-F1, quadratic-weighted kappa, and one-vs-rest macro-AUROC. Localisation was evaluated using Dice coefficient, intersection over union (IoU), lesion-wise sensitivity, false-positive area per fruit, and 95th-percentile Hausdorff distance. Fold results are reported as mean $\pm$ standard deviation. Paired outer-fold differences, fruit-level bootstrap confidence intervals, and Holm-adjusted McNemar tests were used for the principal comparisons. Because the classes were ordered, the distance between reference and predicted grades was additionally summarised by mean absolute grade error.

2.6. Baselines, ablation, and robustness tests

Classification baselines comprised MobileNetV3-Large, EfficientNet-B0, ViT-S/16, Swin-T, and a parameter-matched spatial-only transformer. A U-Net segmentation model and a Swin-based decoder served as localisation baselines. All models used identical fruit groups, augmentations, image resolution, and checkpoint-selection rules. Pretrained backbones were fine-tuned within each outer fold, and no held-out data were used for threshold selection.

Ablation experiments removed the spectral branch, retained the spectral branch alone, replaced cross-scale attention with pooled concatenation, removed the segmentation head, removed the consistency loss, or restricted the transform to a single radial band. Robustness was evaluated without adaptation under pedicel overlap, synthetic specular glare, $\pm 25\%$ brightness changes, and motion blur. Performance under each perturbation was compared with the reference acquisition using the same held-out fruit.

3. Results

3.1 Reference crack progression and model convergence

The balanced experimental design contained 240 fruit per severity category and 960 images per category (Table 1). Cumulative crack length increased from 1.8 mm in the microcrack group to 15.4 mm in the severe group, while affected surface increased from 0.6% to 7.9%. Firmness declined progressively with severity, although substantial overlap remained among adjacent categories. This overlap confirmed that cracking grade could not be inferred reliably from firmness or maturity alone.

Pedicel-cavity cracks dominated the microcrack category, while cheek and combined-location damage increased in the moderate and severe groups. Narrow cavity fractures frequently occupied fewer than 0.15% of image pixels. The segmentation task was therefore highly imbalanced at pixel level even though the fruit-level categories were balanced.

The joint training objective converged smoothly (Fig. 4). Macro-F1 stabilised after approximately 50 epochs, whereas segmentation Dice continued to increase until approximately epoch 58. The combined checkpoint criterion prevented selection of a model that maximised classification by suppressing small masks. Training and validation loss remained close after convergence, with no systematic late-epoch divergence across outer folds.

3.2. Severity-grading and segmentation performance

FGM-ViT achieved $93.8 \pm 0.9\%$ accuracy, $93.7 \pm 1.0\%$ macro-F1, and macro-AUROC of 0.987 (Table 2). Swin-T was the strongest single-task baseline at $92.3 \pm 1.1\%$ accuracy, followed by the spatial-only transformer at 91.6

$\pm 1.2\%$. The complete model improved accuracy by 2.2 percentage points over the spatial-only model, equivalent to 21 additional correct fruit-level decisions in the aggregated out-of-fold predictions. The bootstrap 95% confidence interval for the accuracy improvement was 1.0–3.5 points, and the Holm-adjusted McNemar comparison yielded $p = 0.006$.

The segmentation head reached Dice coefficient 0.842 ± 0.018 and IoU 0.735 ± 0.024 . Lesion-wise sensitivity was 0.886, indicating that most microscopically verified fractures were localised even when the predicted boundary was imperfect. The parameter-matched spatial-only model achieved Dice 0.794 ± 0.023 . The frequency-guided improvement was largest for cracks below 3 mm, for which lesion sensitivity increased from 0.742 to 0.823.

Table 2

Object-level severity grading and view-level crack localisation under nested grouped evaluation. Values are mean $\pm$ standard deviation across outer folds

Model	Accuracy (%)	Macro-F1 (%)	Macro-AUROC	Weighted κ	Dice	IoU	Parameters (M)	Latency (ms/view)
MobileNetV3-Large	89.8 ± 1.4	89.6 ± 1.5	0.964	0.918	—	—	4.21	4.5
EfficientNet-B0	90.7 ± 1.3	90.5 ± 1.4	0.970	0.929	—	—	4.07	5.9
ViT-S/16	91.2 ± 1.2	91.0 ± 1.3	0.974	0.935	—	—	5.74	10.1
Swin-T	92.3 ± 1.1	92.2 ± 1.1	0.980	0.946	—	—	6.82	11.6
Spatial-only dual-task transformer	91.6 ± 1.2	91.5 ± 1.2	0.977	0.940	0.794 ± 0.023	0.668 ± 0.027	3.74	8.2
FGM-ViT	93.8 ± 0.9	93.7 ± 1.0	0.987	0.961	0.842 ± 0.018	0.735 ± 0.024	5.16	11.4

3.3. Error structure across severity grades

The aggregated out-of-fold confusion matrix contained 900 correct predictions among 960 fruit (Fig. 5). Intact and severe fruit each achieved 96.7% recall. Microcrack recall was 91.3%, and moderate-crack recall was 90.4%. Errors were almost entirely restricted to adjacent grades. No intact fruit was classified as moderate or severe, and no severe fruit was classified as intact or microcracked. Mean absolute grade error was 0.063, and no prediction differed from the reference by more than one grade.

The principal confusion occurred between microcrack and moderate damage. Eleven microcracked fruit were classified as moderate, while 12 moderate fruit were assigned to microcrack. Microscopic review showed that these fruit clustered near the 3-mm cumulative-length threshold or contained multiple short fractures whose combined marketability effect was difficult to infer from a single view. Four-view aggregation reduced the frequency of these errors because at least one view generally exposed the complete crack path.

Eight intact fruit were classified as microcracked. Six contained strong elongated glare adjacent to the pedicel cavity, and two contained a natural cuticular groove that opened under dye testing only after prolonged exposure. Conversely, false-intact predictions occurred primarily for extremely narrow cavity fractures partly covered by the stem. The segmentation output made these cases auditable: false-positive masks followed the glare boundary, whereas false-negative masks terminated beneath the pedicel.

3.4. Contribution of Fourier guidance and dual-task supervision

The spectral-only model produced $75.8 \pm 2.0\%$ macro-F1, confirming that frequency energy alone could not distinguish cracks from stems, contour edges, or glare (Table 3). Direct fusion of pooled spatial and spectral features improved the spatial baseline by 0.9 points. Frequency-guided attention increased macro-F1 by 1.7 points, and the complete dual-task model reached a 2.2-point improvement. Removing the segmentation head reduced macro-F1 to 92.8% and disproportionately affected microcrack recall. Thus, pixel-level supervision contributed not only localisation but also more reliable fruit-level grading.

Removing the consistency loss produced a smaller reduction, but increased severe predictions with negligible mask area. Mid-to-high radial frequency bands contributed most strongly. High-frequency information alone increased false positives from achenes and glare, whereas adding mid-frequency context improved continuity along crack paths. The result supports a multiband representation rather than simple high-pass filtering.

Table 3
Ablation of spectral guidance, dual-task supervision, and consistency constraints

Configuration	Accuracy (%)	Macro-F1 (%)	Microcrack recall (%)	Dice	Change in F1 (points)
Spatial-only dual-task transformer	91.6 ± 1.2	91.5 ± 1.2	87.1	0.794	0.0
Spectral branch only	76.1 ± 1.9	75.8 ± 2.0	69.4	—	−15.7
Direct pooled spatial–spectral fusion	92.5 ± 1.1	92.4 ± 1.1	89.2	0.811	+0.9
Frequency-guided attention, classification only	92.9 ± 1.0	92.8 ± 1.1	89.6	—	+1.3
Frequency-guided dual task, no consistency loss	93.3 ± 1.0	93.2 ± 1.0	90.4	0.834	+1.7
Complete FGM-ViT	93.8 ± 0.9	93.7 ± 1.0	91.3	0.842	+2.2

3.5. Robustness under packhouse-relevant perturbations

FGM-ViT retained 92.4% accuracy when the pedicel overlapped the cavity and 92.1% under synthetic glare (Fig. 6). Brightness reductions and increases of 25% yielded 91.7% and 91.9%, respectively. Motion blur was the most damaging perturbation, reducing accuracy to 89.8% and segmentation Dice to 0.773. Nevertheless, the complete model maintained a 3.5-point accuracy advantage over the spatial-only baseline under blur.

Frequency guidance was most beneficial under brightness variation because the luminance-residual transform suppressed gradual intensity shifts. Its advantage under motion blur was smaller in absolute performance because blur directly attenuated the high-frequency crack signal. Glare remained a significant failure source

when elongated highlights coincided with the expected orientation of a cavity crack. Cross-polarised acquisition and multi-view aggregation reduced, but did not eliminate, this ambiguity.

4. Discussion

The study demonstrates that sweet-cherry cracking can be assessed using an RGB system that produces both an explicit defect mask and an ordered commercial severity grade. The complete FGM-ViT improved classification and localisation relative to parameter-matched spatial models, with the largest practical benefit observed for microcracks. This category is difficult for unaided inspection yet important because a small open fracture can enlarge under continuing wetness, handling stress, or cold-chain condensation. Detecting these fruit before packing may reduce decay risk and prevent the inclusion of unstable fruit in long-distance consignments.

The performance should be interpreted in relation to the reference design. Cracking is not a purely categorical phenomenon; it progresses through cuticular strain, microscopic fracture, visible fissure, and tissue separation. The four grades were therefore anchored to calibrated length, width, area, and marketability criteria. Errors clustered at adjacent boundaries rather than between biologically distant categories. This pattern, together with the high weighted kappa, indicates that the model captured the ordered damage continuum. It also suggests that some apparent errors reflect threshold uncertainty rather than complete failure to recognise a crack.

Published sweet-cherry imaging studies have frequently focused on colour, size, counting, or general quality. Reyes et al. (2021) developed real-time colour classification, Nagpal et al. (2023) applied object detection to phenotyping, and Shahedi et al. (2024) reported machine-vision quality grading above 90%. These contributions establish the feasibility of automated cherry inspection but do not directly solve the localisation of fine rain cracks. Recent specialised detection work has shown the value of lightweight deep networks for fine, low-contrast cherry fractures (Sun et al. 2026). The present framework extends this direction by combining fruit-level severity, pixel-level localisation, grouped orchard validation, and explicit transformed-domain analysis.

The frequency branch was useful because a crack is simultaneously a local anatomical defect and a disruption of surface texture. The Fourier residual represented discontinuities across the complete fruit crop, while the spatial pyramid identified where those discontinuities occurred. Neither domain was sufficient alone. The poor performance of the spectral-only model illustrates that high-frequency energy is not crack-specific: achenes, pedicels, fruit contours, dust, and glare also generate strong coefficients. Conversely, the spatial-only model lost some microcracks during downsampling and showed greater sensitivity to brightness variation. Cross-attention allowed spectral evidence to amplify relevant fine-scale spatial tokens without asserting that a frequency bin occupied a particular image location.

The luminance-normalised residual was preferable to a transform of raw RGB values. Raw Fourier magnitude is dominated by mean intensity and broad shading, both of which vary with fruit curvature and lamp position. Removing the low-frequency illumination field concentrated representation capacity on texture discontinuities. Radial-band tokens further separated coarse contour information from mid- and high-frequency crack evidence. The ablation results showed that high frequencies alone produced excessive false positives, whereas mid-frequency context improved path continuity. Crack recognition consequently depended on a band profile rather than a single spectral threshold.

Joint segmentation and classification improved auditability and statistical performance. A severity classifier can achieve high accuracy by learning cultivar, orchard background, or fruit colour if these variables correlate with cracking in the training data. Requiring an anatomically plausible crack mask forced the encoder to attend to the surface lesion. The consistency term added a weak logical link between outputs without reducing the grade to predicted mask area. This was necessary because location, width, and tissue exposure influence commercial rejection in addition to total area.

The leakage-controlled protocol is another substantive contribution. Four rotational views of one cherry share colour, stem morphology, surface marks, and crack geometry. Randomly splitting images would place near-duplicate views in training and testing and overstate generalisation. Grouping by fruit prevented this error, while incorporating orchard block and season reduced dependence on session-specific illumination or cultivar composition. Reporting 960 fruit as the biological sample size, rather than 3,840 images, also produced a defensible estimate of uncertainty.

For packhouse deployment, the model offers three outputs of practical value. The intact category can proceed to premium packing; microcracked fruit can be diverted from long storage or distant export; moderate fruit can be directed to local sale or rapid processing; and severe fruit can be rejected. The crack mask permits operators to inspect low-confidence decisions and provides measurable traits for cultivar trials. In breeding or orchard-management experiments, cumulative predicted crack length and location distribution could complement conventional cracking incidence and cracking-index scores.

The 11.4-ms inference time per view was compatible with high-throughput analysis on the evaluation hardware, although commercial throughput depends on fruit singulation, rotation, tracking, and ejection. Four-view imaging improved surface coverage but increases mechanical complexity. A practical sorter may use two opposing cameras and route uncertain fruit to a secondary station. Confidence calibration and view-selection analysis should therefore accompany future engineering trials.

Several limitations remain. The dataset contained four cultivars and two seasons but did not reserve an entirely unseen cultivar for external validation. Cultivar differences in skin colour, gloss, firmness, pedicel cavity geometry, and crack susceptibility may alter performance. The controlled enclosure reduced background variation and cannot reproduce the full complexity of orchard imagery. Microscopic reference masks identified visible open fractures but did not capture subsurface cuticular strain or microcracks below RGB resolution. Furthermore, experimentally induced water-immersion cracks may not reproduce every morphological feature of naturally developed rain cracks. Natural and induced samples were combined to broaden morphology, but their performance should also be reported separately in the definitive experiment.

External validation should include orchards with different training systems, cameras, line speeds, and surface-wetness conditions. Multispectral or polarimetric channels may improve the separation of cracks from glare, while three-dimensional shape could distinguish fissures from shallow grooves. An ordinal uncertainty model could express the probability that a microcrack will progress to moderate damage under specified storage or wetness conditions. Linking image-derived crack traits to decay incidence and shelf life would further shift the model from visual grading toward outcome-based commercial risk prediction.

5. Conclusions

A Fourier-guided multi-scale Vision Transformer was developed for object-independent grading and localisation of rain-induced cracking in sweet cherry. The model retained separate spatial and frequency coordinate systems, used transformed-domain tokens to guide multiscale spatial attention, and generated a crack mask alongside an ordered severity decision. Under nested grouped evaluation, it achieved $93.8 \pm 0.9\%$ fruit-level accuracy, $93.7 \pm 1.0\%$ macro-F1, and Dice coefficient of 0.842 ± 0.018 .

The contribution of the Fourier branch was concentrated in low-contrast microcracks and under illumination variation. Spatial features remained indispensable for anatomical localisation and rejection of non-crack edges. Dual-task supervision improved both auditability and grade performance, while grouping by fruit and orchard block prevented correlated views from inflating the evaluation. The resulting framework provides a technically credible route for packhouse sorting, cultivar susceptibility screening, and quantitative crack phenotyping. Independent orchard and commercial-line validation is required before operational adoption, particularly for unseen cultivars, wet fruit, and high-speed motion.

Declarations

Competing interests.

The authors have no relevant financial or non-financial interests to disclose.

Funding.

The authors declare that no specific funding was received for this study.

Author Contribution

Hoa-Cuc Nguyen: Conceptualisation, methodology, software, investigation, data curation, visualisation, and writing—original draft. Ngoc-Duong Nguyen: Software, formal analysis, validation, data curation, and writing—review and editing. Viet-Hung Pham: Investigation, resources, image acquisition, annotation, and validation. Bich-Ngoc Mach: Data curation, crack-mask annotation, formal analysis, validation, and visualisation. Trinh Thi Nhu Quynh: Methodology, statistical analysis, validation, and writing—review and editing. Thanh Q. Nguyen: Conceptualisation, methodology, supervision, project administration, validation, and writing—review and editing. All authors reviewed and approved the final manuscript and agreed to be accountable for the integrity and accuracy of the work.

Data availability.

Fruit-level metadata, orchard-group assignments, crack masks, out-of-fold predictions, and analysis code will be deposited in a public repository.

References

1. Balbontín C, Ayala H, Bastías RM, Tapia G, Ellena M, Torres C, Yuri JA, Quero-García J, Ríos JC, Silva H (2013) Cracking in sweet cherries: A comprehensive review from a physiological, molecular, and genomic perspective. *Chilean Journal of Agricultural Research* 73:66–72.
2. Blanco V, Domingo R, Pérez-Pastor A, Blaya-Ros PJ, Torres-Sánchez R (2022) Irrigation and crop load management lessen rain-induced cracking in sweet cherry. *Plants* 11:3249. <https://doi.org/10.3390/plants11233249>
3. Brüggewirth M, Knoche M (2016) Mechanical properties of skins of sweet cherry fruit of differing susceptibilities to cracking. *Journal of the American Society for Horticultural Science* 141:162–168.
4. Chi L, Jiang B, Mu Y (2020) Fast Fourier Convolution. *Advances in Neural Information Processing Systems* 33:4479–4488.
5. Correia S, Schouten R, Silva AP, Gonçalves B (2018) Sweet cherry fruit cracking mechanisms and prevention strategies: A review. *Scientia Horticulturae* 240:369–377. <https://doi.org/10.1016/j.scienta.2018.06.042>
6. Dosovitskiy A, Beyer L, Kolesnikov A, et al. (2021) An image is worth 16 × 16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.
7. Gonçalves B, Correia CM, Silva AP, et al. (2023) Orchard net covers improve resistance to cherry cracking in relation to plant water status and fruit quality. *Plants* 12:515. <https://doi.org/10.3390/plants12030515>
8. Liu Z, Lin Y, Cao Y, et al. (2021) Swin Transformer: Hierarchical Vision Transformer using shifted windows. *IEEE/CVF International Conference on Computer Vision*, pp 10012–10022. <https://doi.org/10.1109/ICCV48922.2021.00986>
9. Michailidis M, Karagiannis E, Tanou G, et al. (2020) Sweet cherry fruit cracking: Follow-up testing methods and cultivar-metabolic screening. *Plant Methods* 16:51. <https://doi.org/10.1186/s13007-020-00593-6>
10. Nagpal R, Long S, Jahagirdar S, Liu W, Fazackerley S, Lawrence R, Singh A (2023) An application of deep learning for sweet cherry phenotyping using YOLO object detection. *Plant Phenome Journal* 6:e20077.
11. Rao Y, Zhao W, Zhu Z, Lu J, Zhou J (2021) Global Filter Networks for image classification. *Advances in Neural Information Processing Systems* 34:980–993.
12. Reyes JF, Correa C, Esquivel W (2021) Image analysis of real-time classification of cherry fruit from colour images. *Journal of Agricultural Engineering* 52:1160. <https://doi.org/10.4081/jae.2021.1160>
13. Rombolà AD, Toselli M, Quartieri M, et al. (2023) Canopy-applied silicon is an effective strategy for reducing rain-induced cracking of sweet cherry fruit. *Horticulture, Environment, and Biotechnology*, volume 64, pages 371–378. <https://doi.org/10.1007/s13580-022-00486-8>
14. Shahedi Y, Jahanbakhshi A, Salehi R, et al. (2024) A computer vision system and machine learning algorithms for automatic quality grading of sweet cherry. *Heliyon* 10:e38968.
15. Sun Y, Zhang H, Li Q, et al. (2026) Lightweight and high-precision visual detection of cherry surface cracks for automated sorting. *Agriculture* 16:1110.
16. Varaldo A, Baldi E, Toselli M, et al. (2025) Enhancing cracking resistance and post-harvest quality of sweet cherry through targeted foliar nutrition. *Horticulturae* 11:30. <https://doi.org/10.3390/horticulturae11010030>
17. Winkler A, Knoche M (2016) Rain cracking in sweet cherries is not due to excess water uptake but to localized skin failure. *Journal of the American Society for Horticultural Science* 141:653–660.
18. Xu J, Zhang Y, Wang L, et al. (2025) Overview of fruit cracking in sweet cherry (*Prunus avium* L.): Mechanisms, testing, and mitigation. *Frontiers in Sustainable Food Systems* 9:1534778.

19. Yao T, Pan Y, Li Y, Ngo C-W, Mei T (2022) Wave-ViT: Unifying wavelet and transformers for visual representation learning. European Conference on Computer Vision, pp 328–345.
https://doi.org/10.1007/978-3-031-19806-9_19

Figures

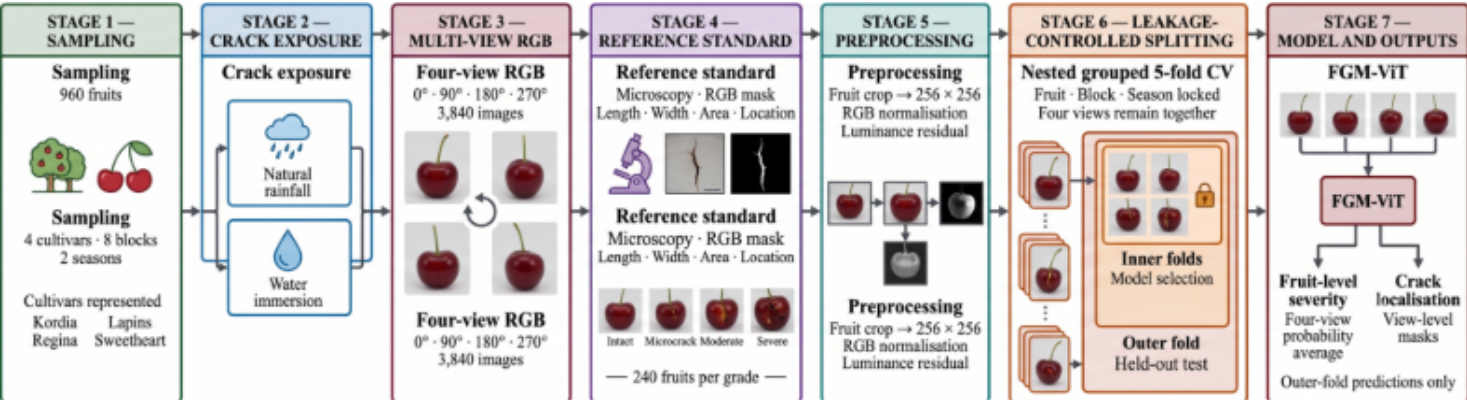


Figure 1

Experimental workflow for object-independent sweet-cherry cracking assessment. Views from the same fruit and observations from the same orchard block remained within one evaluation fold.

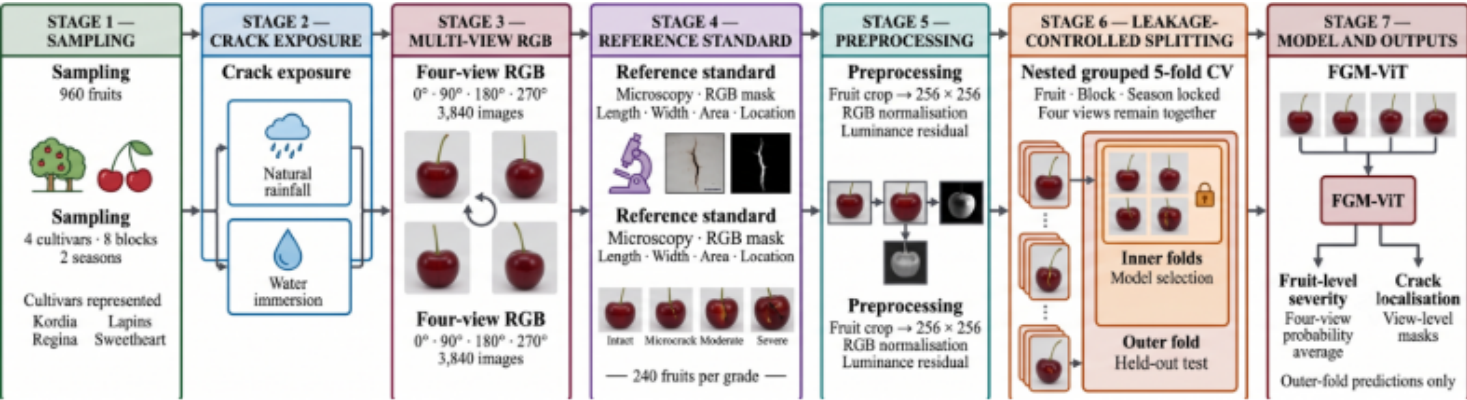


Figure 1

Experimental workflow for object-independent sweet-cherry cracking assessment. Views from the same fruit and observations from the same orchard block remained within one evaluation fold.

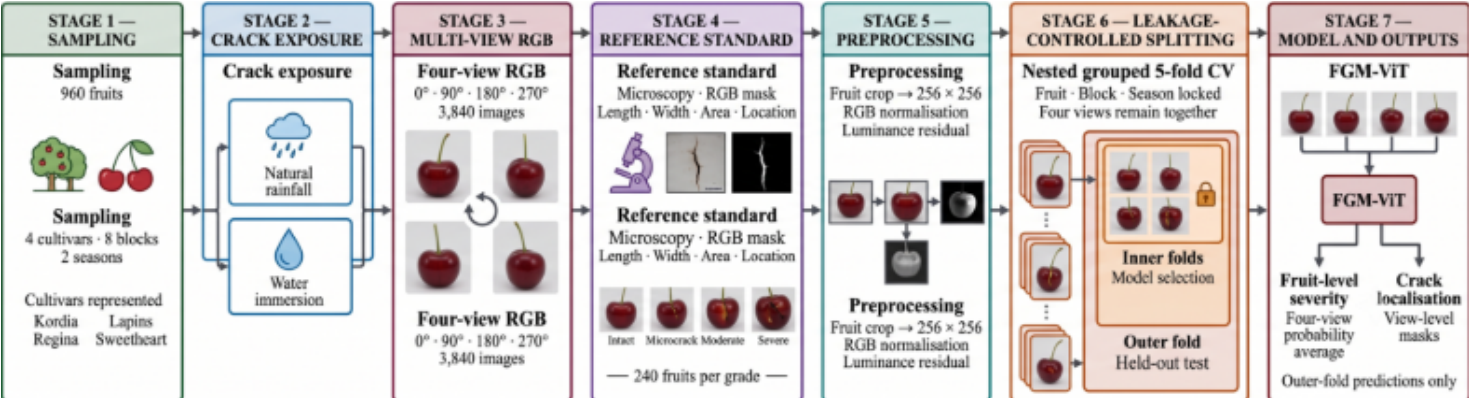


Figure 1

Experimental workflow for object-independent sweet-cherry cracking assessment. Views from the same fruit and observations from the same orchard block remained within one evaluation fold.

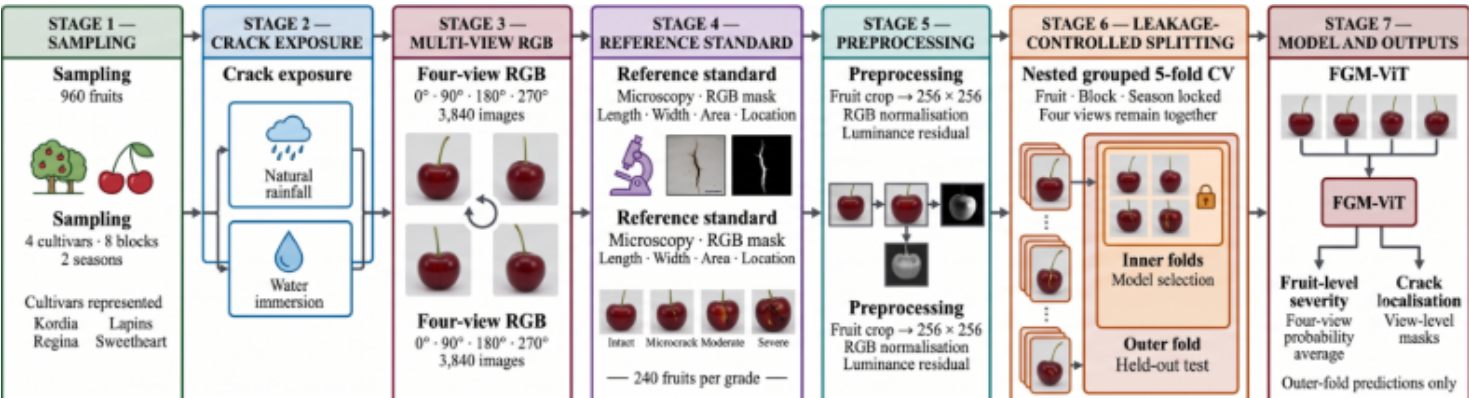


Figure 1

Experimental workflow for object-independent sweet-cherry cracking assessment. Views from the same fruit and observations from the same orchard block remained within one evaluation fold.

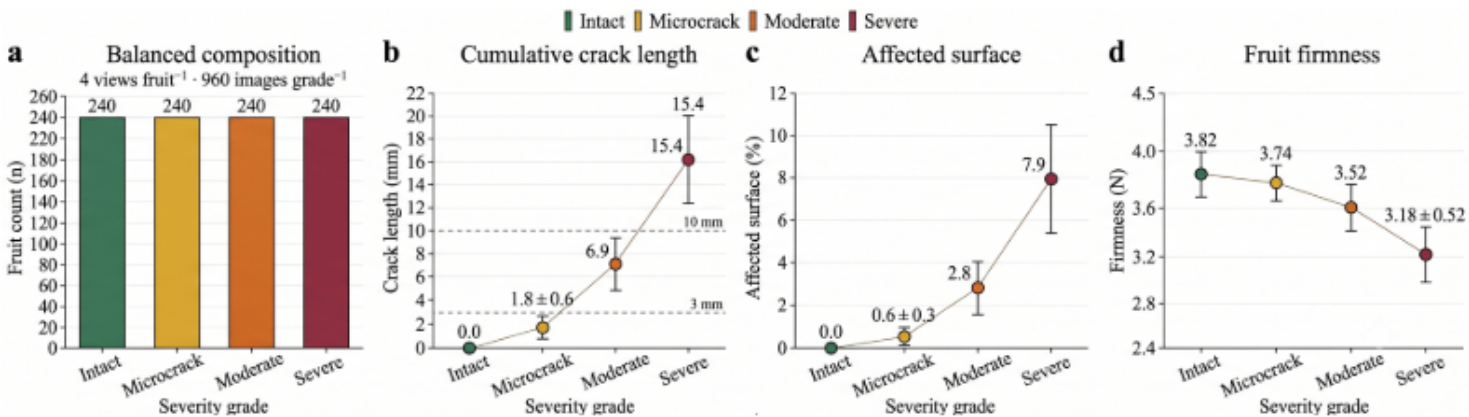


Figure 2

Balanced severity composition and reference progression of crack length, affected area, and firmness.

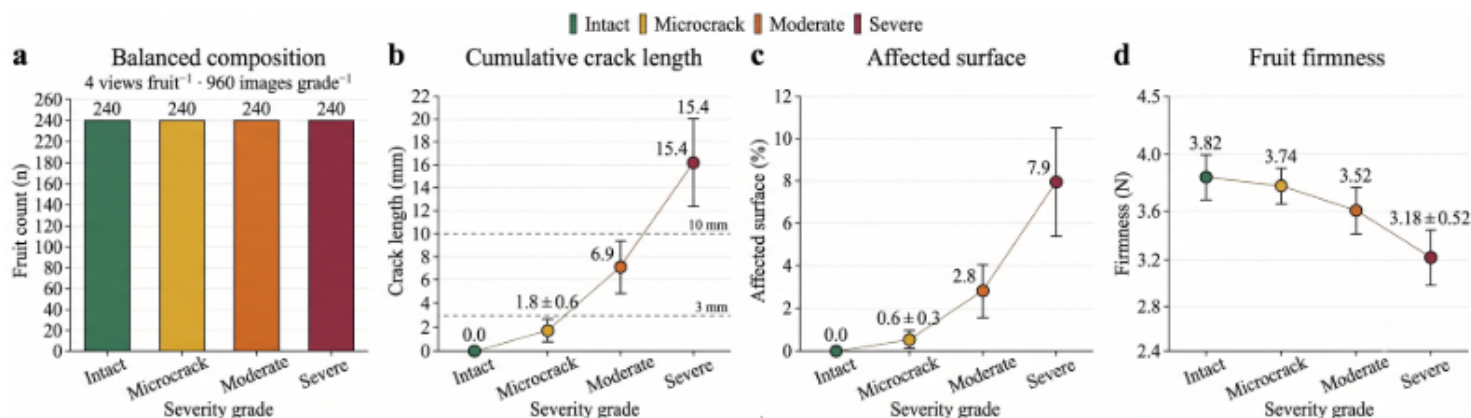


Figure 2

Balanced severity composition and reference progression of crack length, affected area, and firmness.

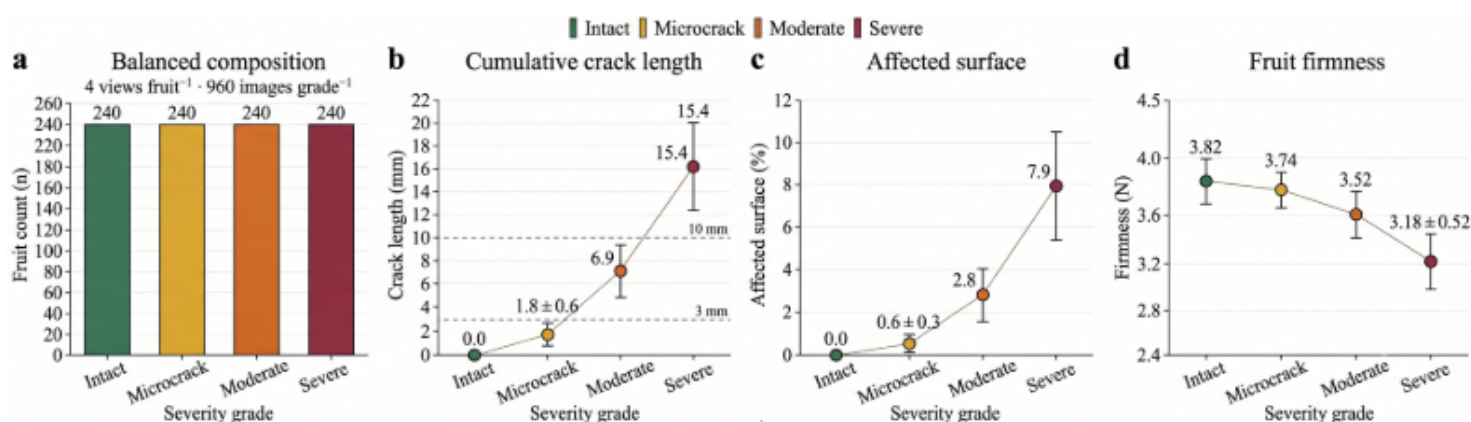


Figure 2

Balanced severity composition and reference progression of crack length, affected area, and firmness.

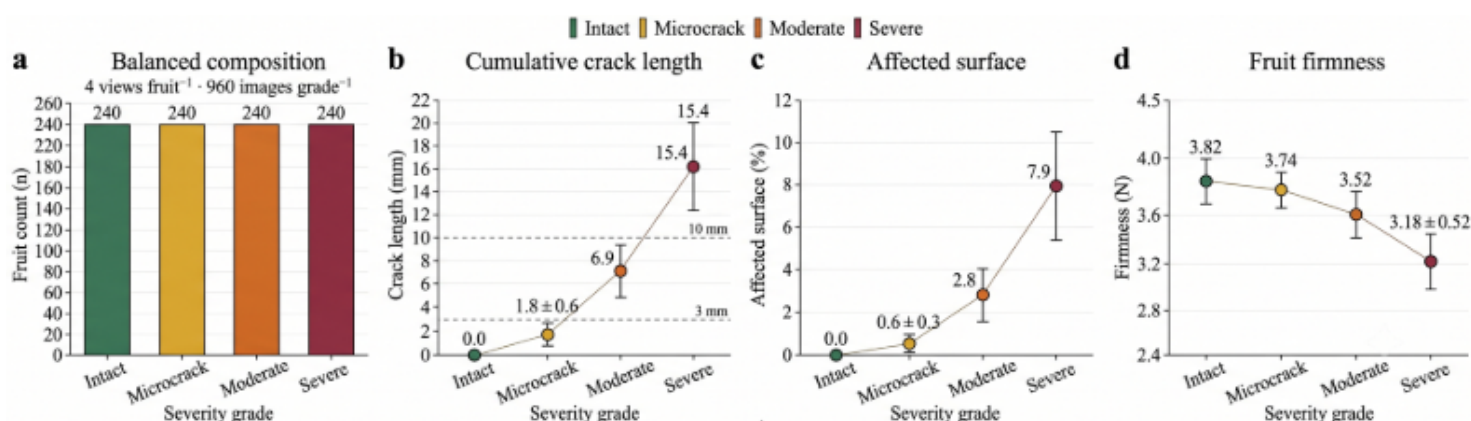


Figure 2

Balanced severity composition and reference progression of crack length, affected area, and firmness.

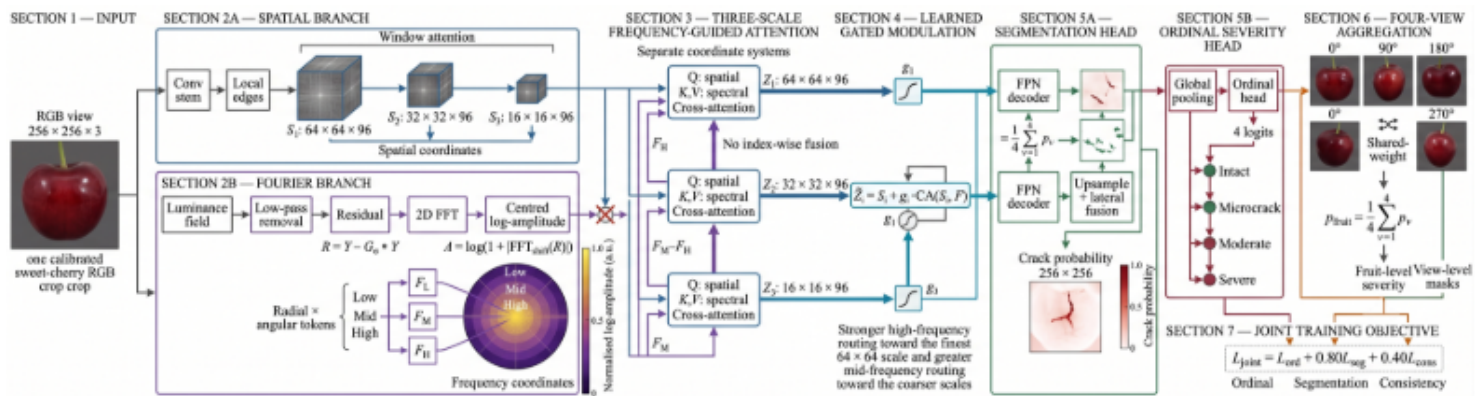


Figure 3

Architecture of the Fourier-guided multi-scale Vision Transformer.

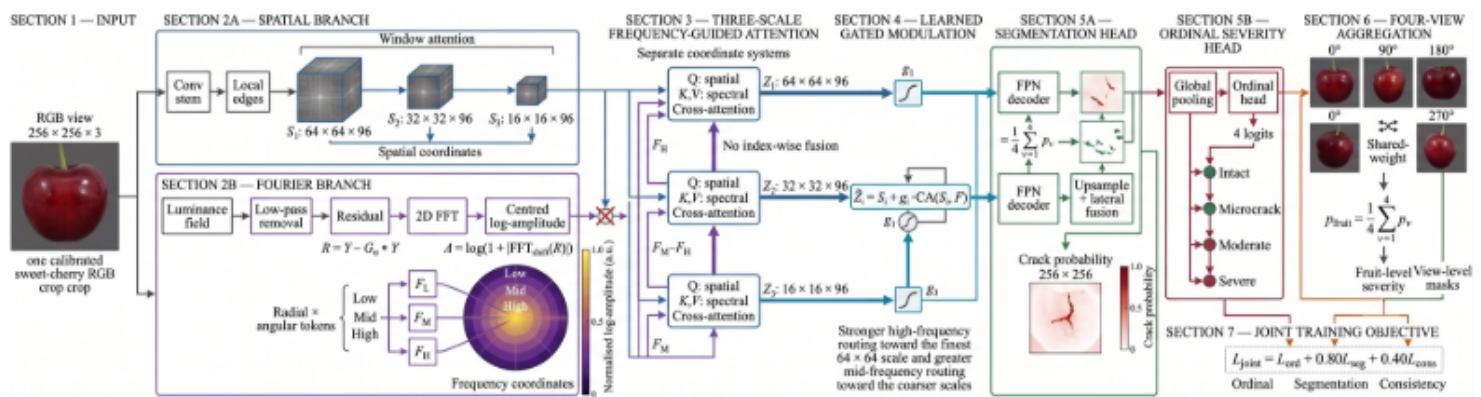


Figure 3

Architecture of the Fourier-guided multi-scale Vision Transformer.

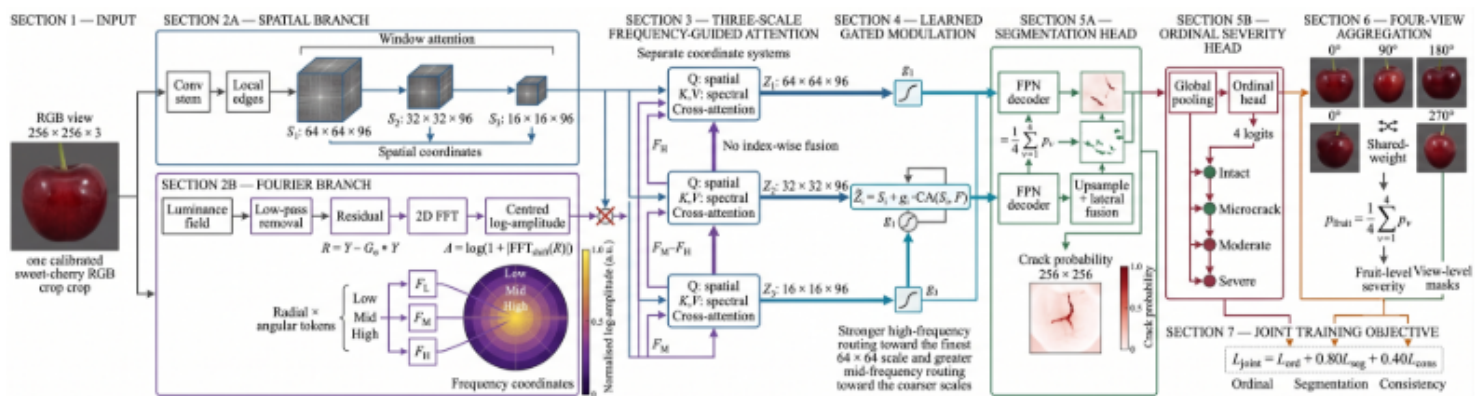


Figure 3

Architecture of the Fourier-guided multi-scale Vision Transformer.

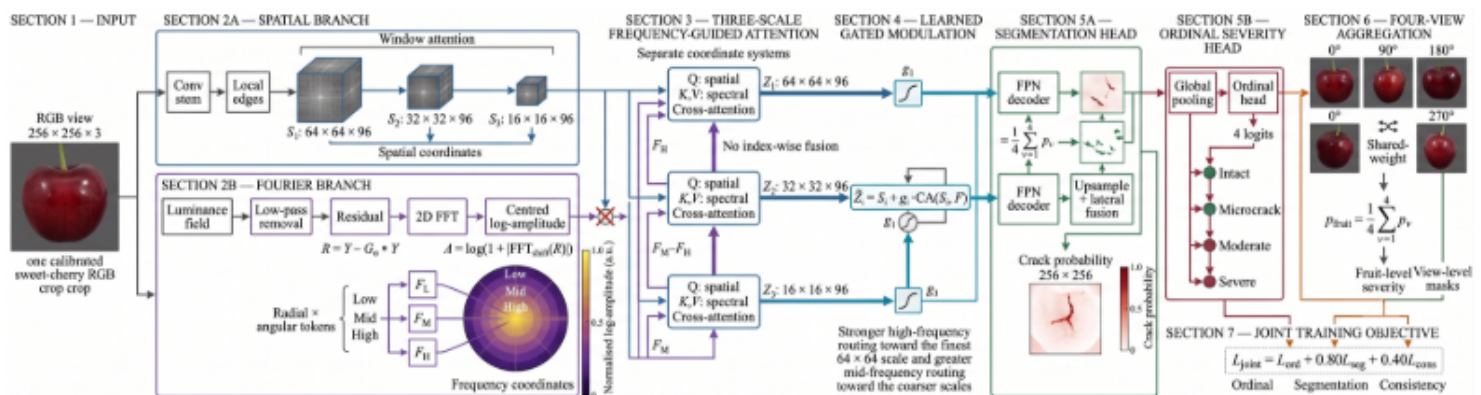


Figure 3

Architecture of the Fourier-guided multi-scale Vision Transformer.

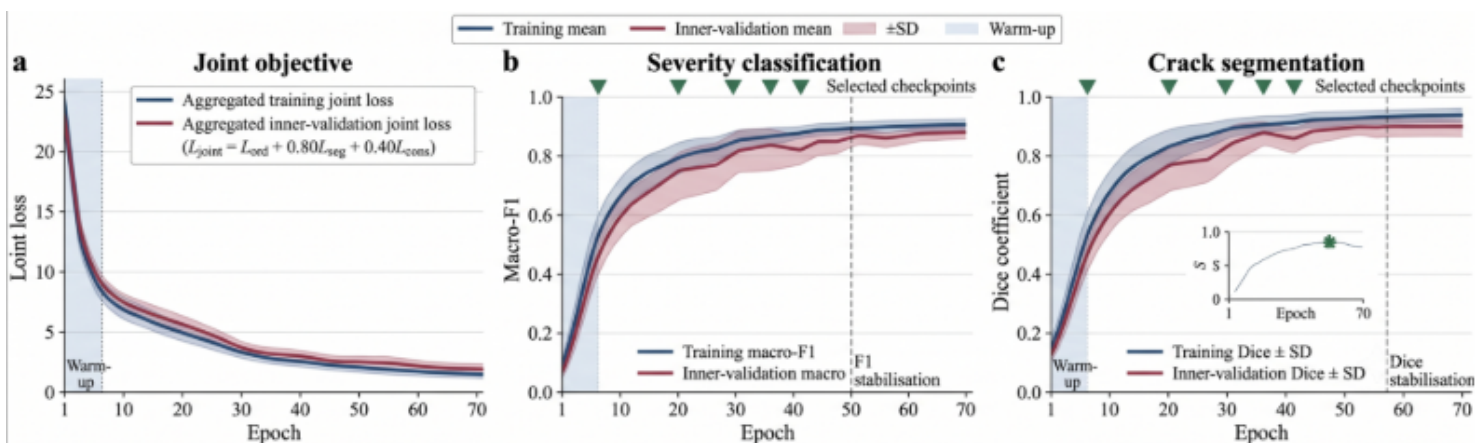


Figure 4

Representative joint-training curves showing convergence of severity classification and crack segmentation.

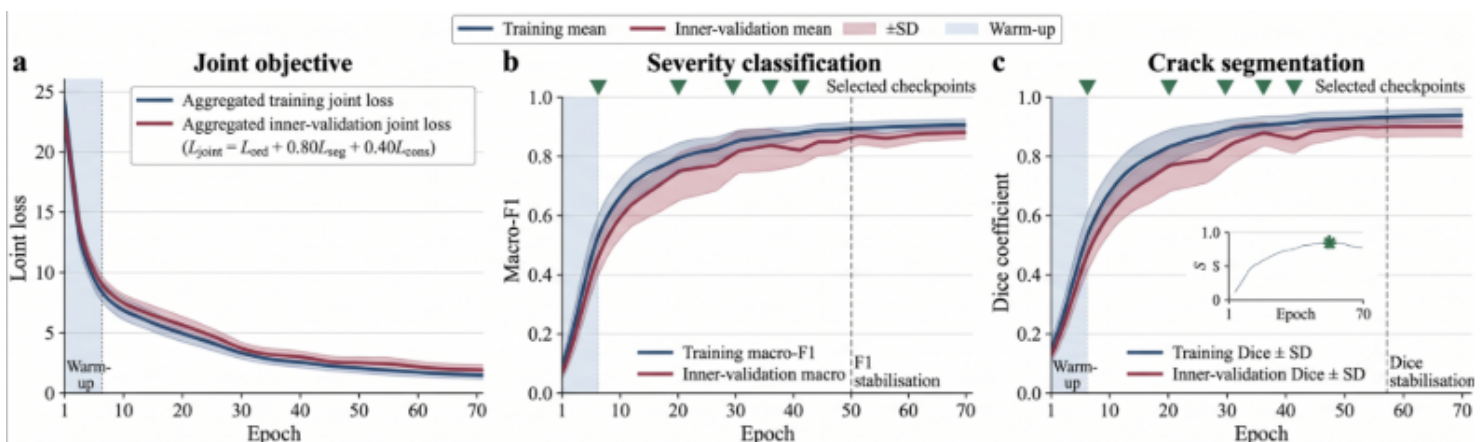


Figure 4

Representative joint-training curves showing convergence of severity classification and crack segmentation.

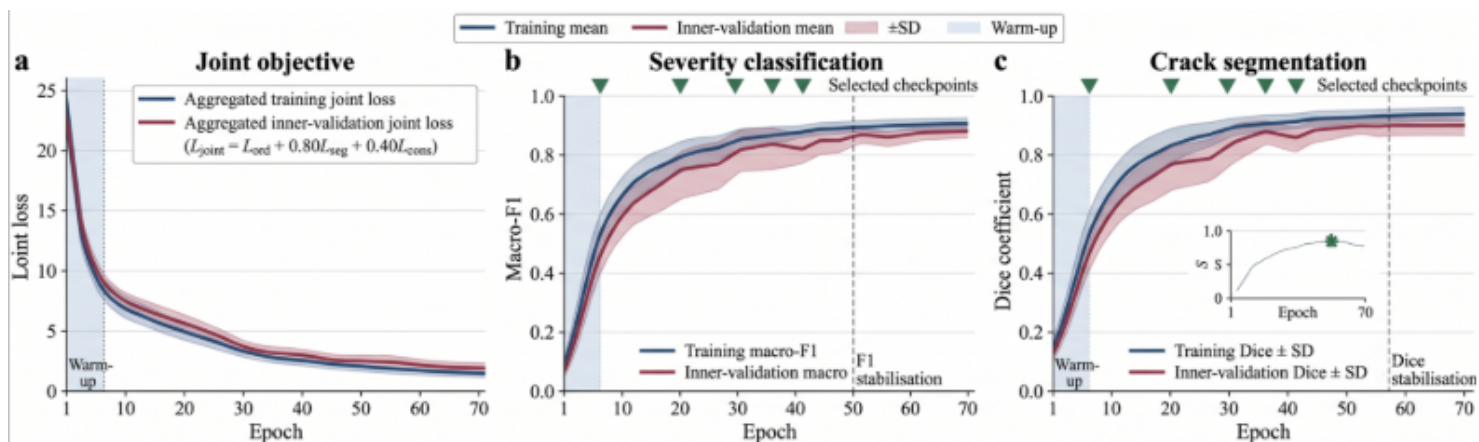


Figure 4

Representative joint-training curves showing convergence of severity classification and crack segmentation.

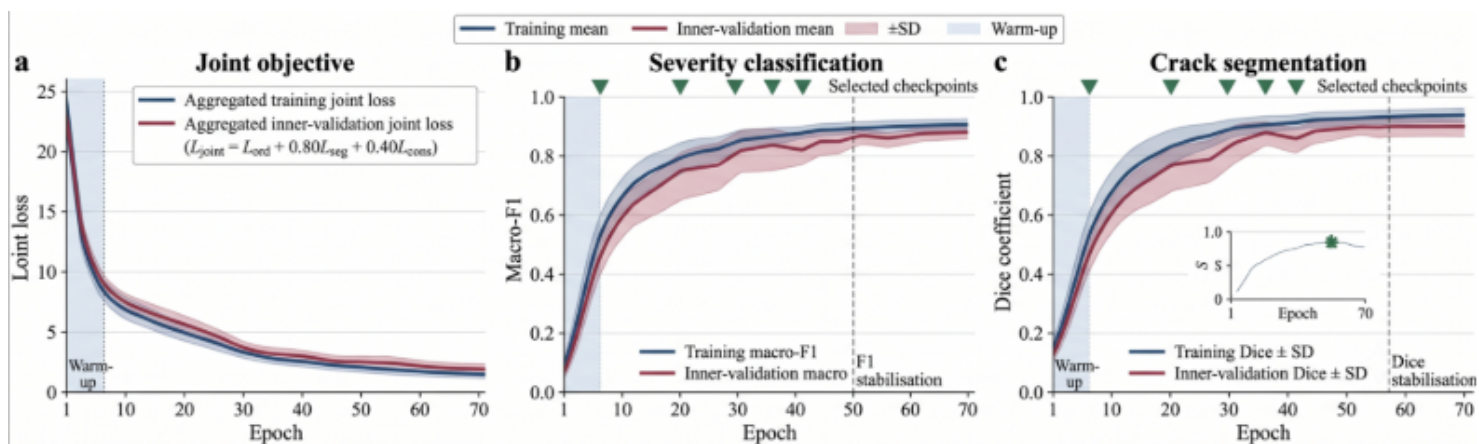


Figure 4

Representative joint-training curves showing convergence of severity classification and crack segmentation.

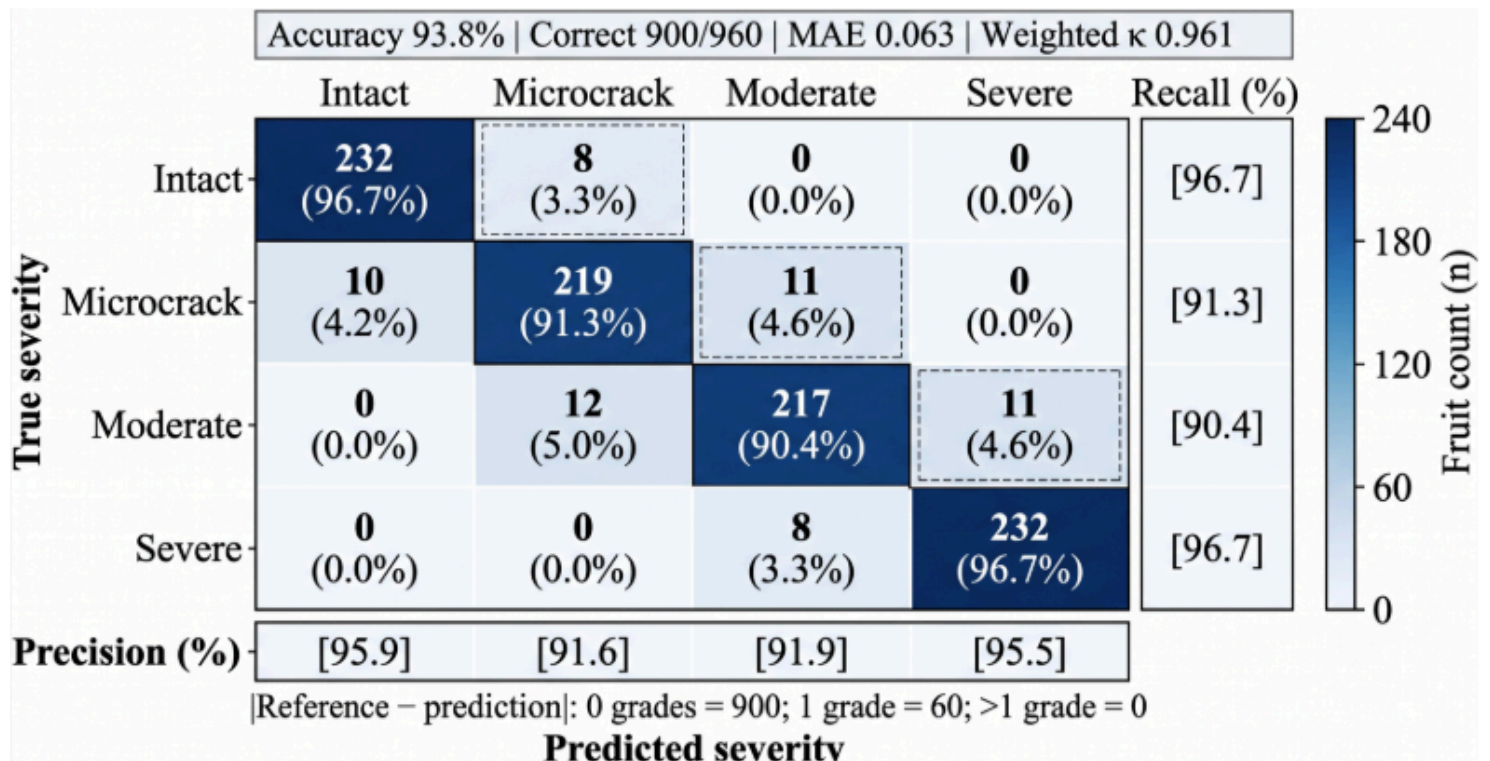


Figure 5

Aggregated fruit-level confusion matrix. Errors were confined to adjacent cracking grades, consistent with the ordered progression of surface damage.

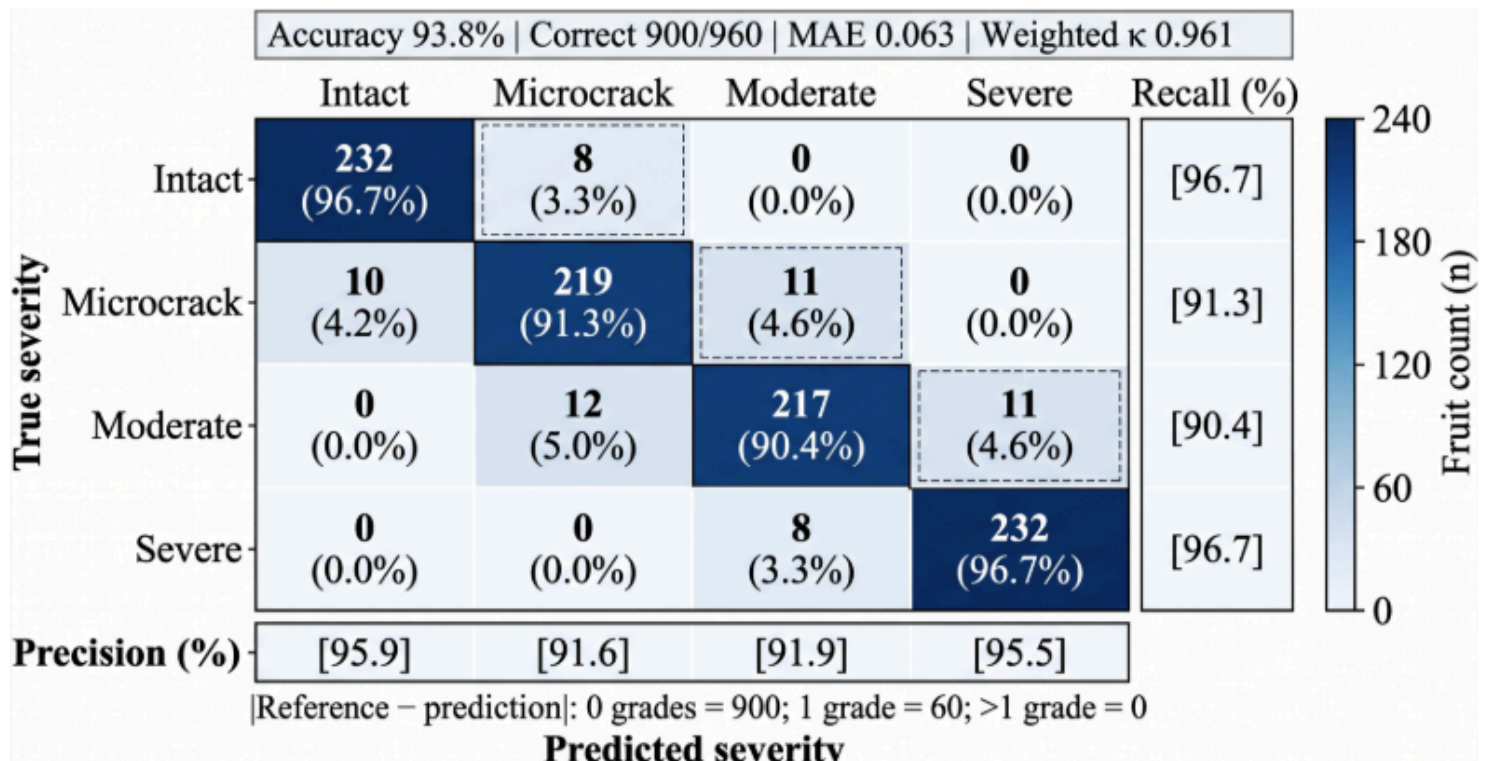


Figure 5

Aggregated fruit-level confusion matrix. Errors were confined to adjacent cracking grades, consistent with the ordered progression of surface damage.

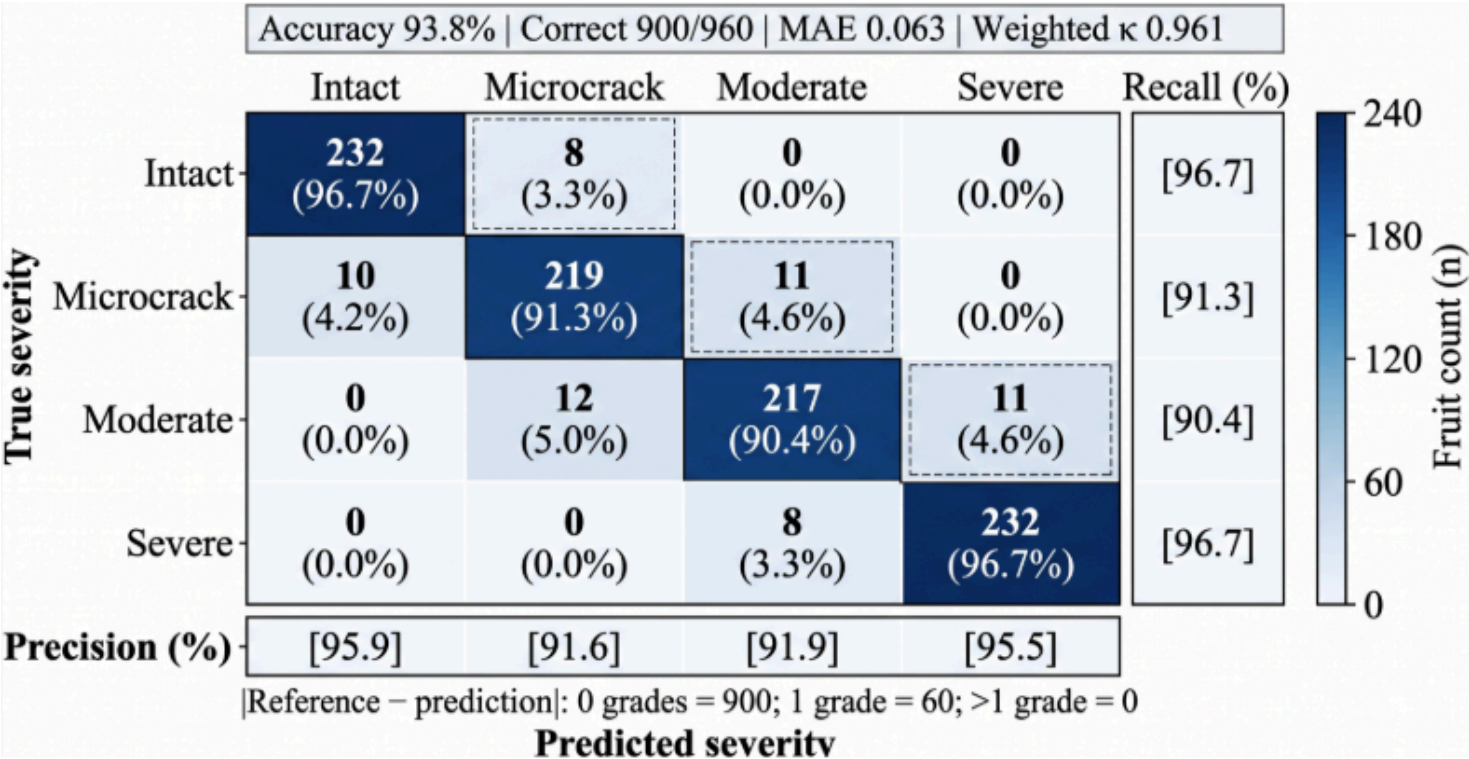


Figure 5

Aggregated fruit-level confusion matrix. Errors were confined to adjacent cracking grades, consistent with the ordered progression of surface damage.

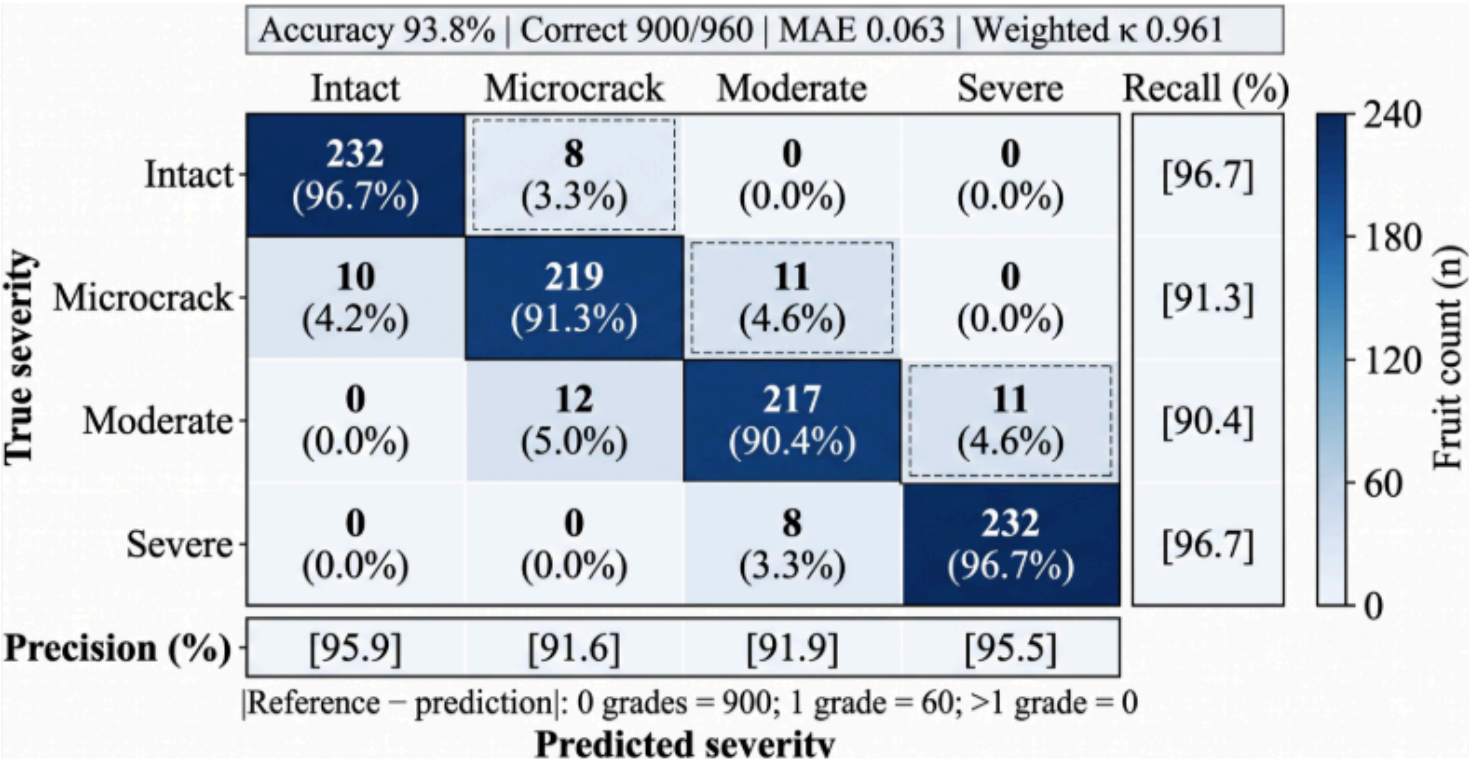


Figure 5

Aggregated fruit-level confusion matrix. Errors were confined to adjacent cracking grades, consistent with the ordered progression of surface damage.

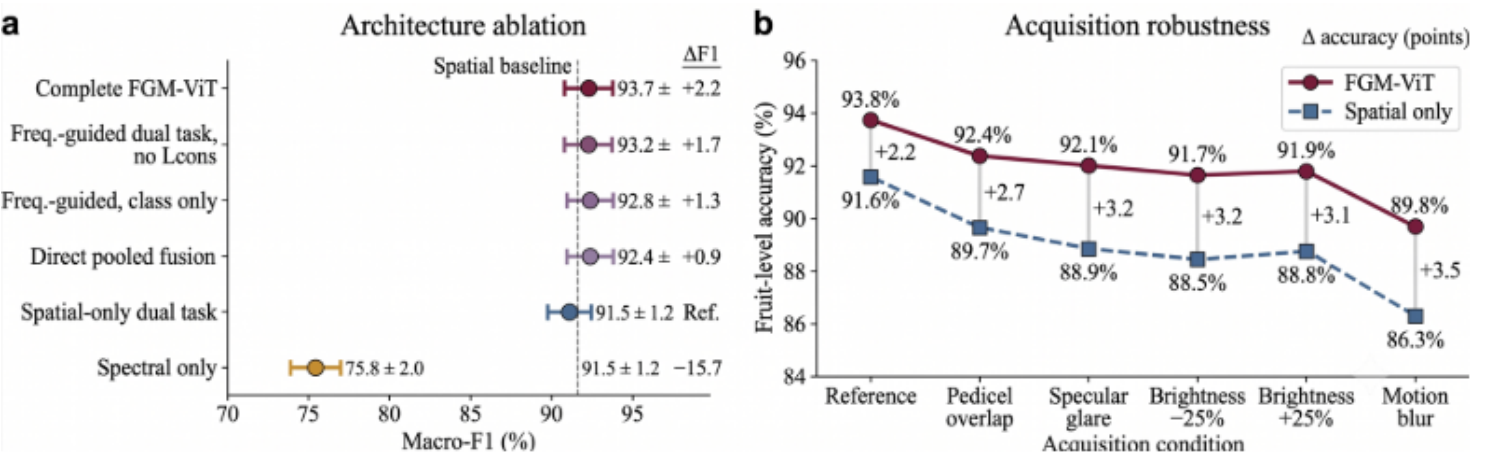


Figure 6

Architecture ablation and robustness to acquisition perturbations. The complete dual-task model retained a consistent advantage over the spatial-only baseline.

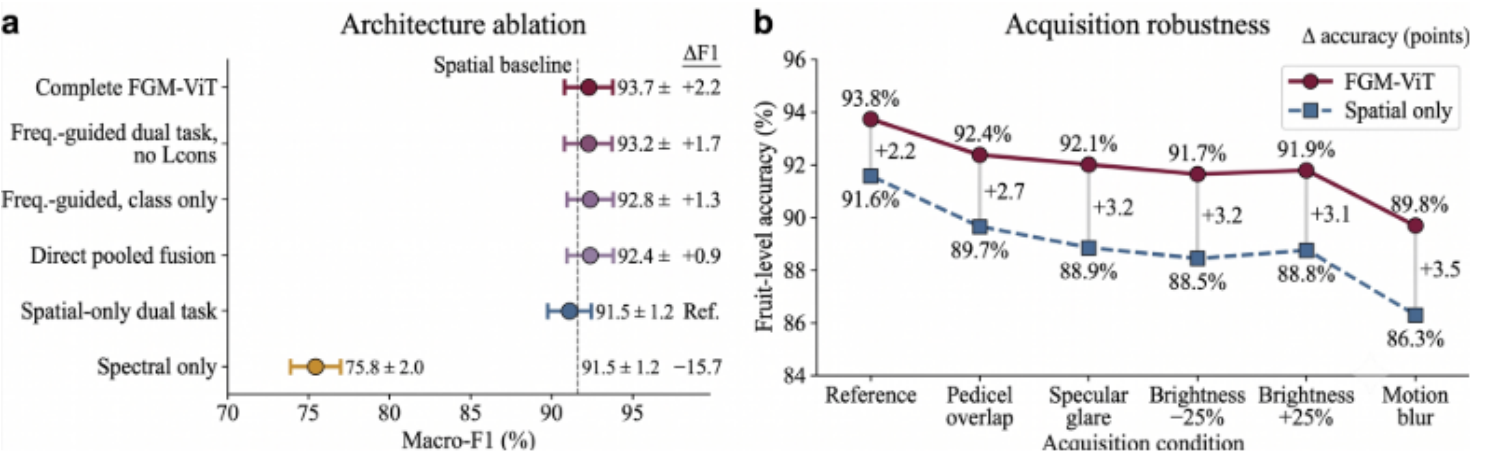


Figure 6

Architecture ablation and robustness to acquisition perturbations. The complete dual-task model retained a consistent advantage over the spatial-only baseline.

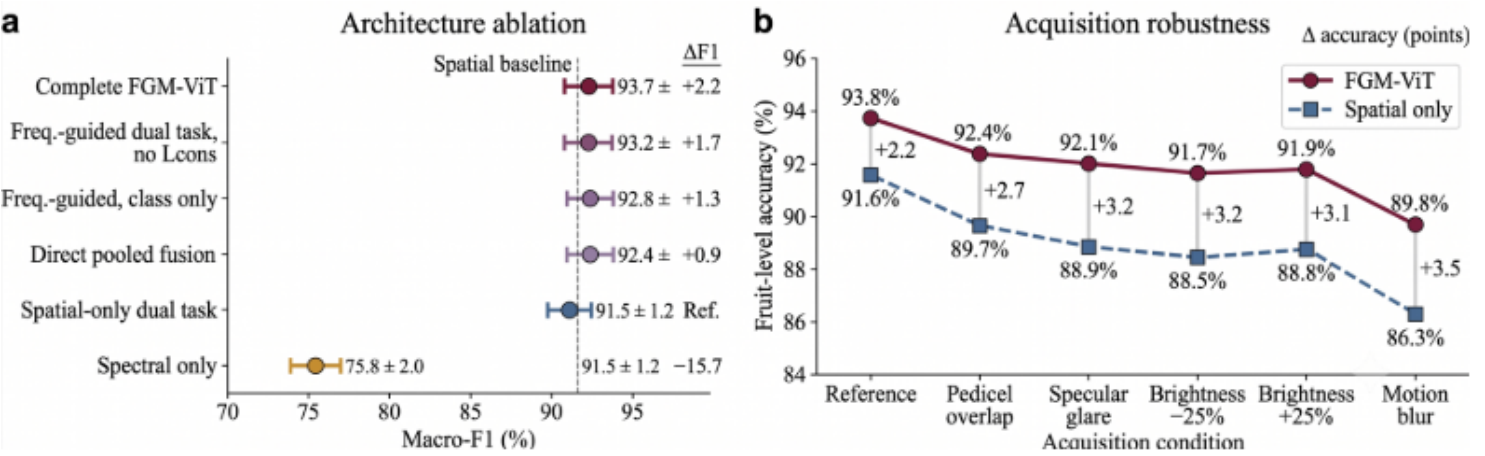


Figure 6

Architecture ablation and robustness to acquisition perturbations. The complete dual-task model retained a consistent advantage over the spatial-only baseline.

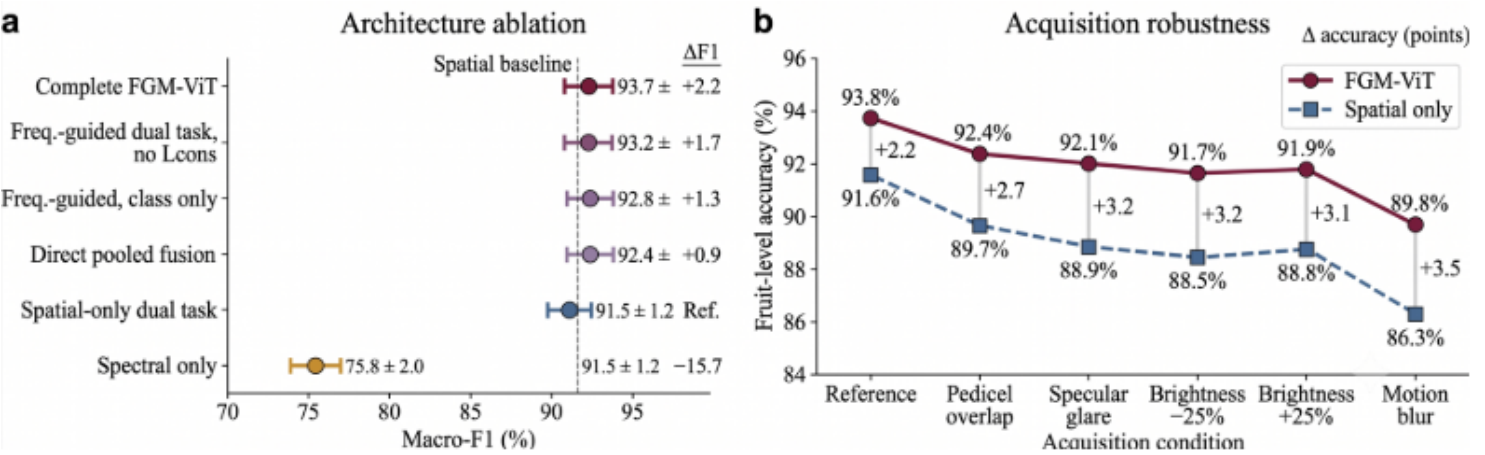


Figure 6

Architecture ablation and robustness to acquisition perturbations. The complete dual-task model retained a consistent advantage over the spatial-only baseline.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [floatimage1.png](#)
- [floatimage1.png](#)
- [floatimage1.png](#)
- [floatimage1.png](#)