

Improving Reliability in Plant Disease Semantic Segmentation by Quantifying Benchmark Contamination and Duplicate Family Exposure

Kotra Naik D¹

kotresh.cit@gmail.com

Malnad College of Engineering

Ananda Babu

Malnad College of Engineering

K C Suhas

Channabasaveshwara Institute of Technology

Research Article

Keywords: Plant disease segmentation, benchmark integrity, cross-split contamination, data leakage, semantic segmentation, duplicate-family exposure, group-disjoint splitting

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11109696/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Improving Reliability in Plant Disease Semantic Segmentation by Quantifying Benchmark Contamination and Duplicate Family Exposure

Kotra Naik D¹ · Dr. Ananda Babu J² · Dr. Suhas K C³

¹ Department of Information Science and Engineering, Malnad College of Engineering, Visvesvaraya Technological University, Belagavi, #21, Salagame Road, RangoliHalla, Hassan, Karnataka 573202, India

² Department of Computer Science and Engineering, Malnad College of Engineering, Visvesvaraya Technological University, Belagavi, #21, Salagame Road, RangoliHalla, Hassan, Karnataka 573202, India

³ Department of Computer Science and Engineering, Channabasaveshwara Institute of Technology, Visvesvaraya Technological University, Belagavi, Gubbi, Tumakuru, Karnataka, India

Kotra Naik D

Research Scholar, Department of Information Science and Engineering,
Malnad College of Engineering, Hassan, Karnataka, India
Email: kotresh.cit@gmail.com
ORCID: 0000-0002-2761-1915

Dr. Ananda Babu J

Professor and Head Department of Computer Science and Engineering,
Malnad College of Engineering, Hassan, Karnataka, India
Email: anand.tiptur@gmail.com
ORCID: 0000-0001-6158-7616

Dr. Suhas K C

Associate Professor, Department of Computer Science and Engineering,
Channabasaveshwara Institute of Technology, Gubbi, Tumakuru, Karnataka, India
Email: suhas2385@gmail.com
ORCID: 0000-0003-3222-5418

Abstract

Reliable plant-disease segmentation depends on evaluation images being independent of the data used for training. We audited PlantSeg, a large in-the-wild semantic-segmentation benchmark, to determine whether visually related image families cross its released partitions and whether such exposure changes the evidence produced by standard segmentation models. The audit combined multiple similarity signals with human adjudication, followed by contamination-aware re-evaluation, class-count-matched Monte Carlo controls, a two-seed training-exposure experiment, long-tail analysis, and group-disjoint benchmark reconstruction. Verified cross-split visual families involved about 10.7% of the released images. Across four canonical architectures, released-split evaluation was consistently more favorable than contamination-aware evaluation, and the main shifts persisted after controlling for disease-class composition. The training side exposure had very limited effect in the independent clean test cases but was seen to have an estimated 11–15% advantage in related contaminated test cases in both seeds. Deletion of family ties improved independence at the cost of rare class support; while the group-disjoint reconstruction kept all 115 disease classes intact while removing known family ties. This illustrates that benchmark contamination can be overestimated and family-aware split construction is a viable solution.

Keywords: Plant disease segmentation; benchmark integrity; cross-split contamination; data leakage; semantic segmentation; duplicate-family exposure; group-disjoint splitting

1. Introduction

Plant disease semantic segmentation supports lesion localization, severity estimation, phenotyping, and precision agriculture, but field imagery remains difficult because symptoms vary in size, shape, contrast, occlusion, illumination, and background context. Recent reviews describe these sources of variability as persistent challenges for automated plant-disease analysis [25]. Convolutional neural network (CNN) approaches such as modified U-Net remain widely used for lesion delineation [16], while DeepLab-based models have incorporated attention mechanisms for more difficult field scenes [6]. Transformer-based designs have also been introduced to capture broader spatial context [13]. These advances improve representation, but their reported performance still depends on the validity of the benchmark split on which they are evaluated.

PlantSeg provides 7,774 image-mask pairs covering 115 plant diseases and 34 host plants under in-the-wild imaging conditions [1]. It has already supported model development on the released benchmark. PlantFormer targets blurred lesion boundaries and anisotropic symptom spread [2], STAR-Net combines morphology-aware attention with a dynamic loss for imbalanced data [3], and ReLeaf-SAM adapts the Segment Anything Model (SAM) with reliability and local-detail compensation [4]. As PlantSeg becomes a comparative benchmark, independence between its training and evaluation partitions becomes as important as architectural performance. A score obtained on a contaminated split may remain useful for historical comparison while still overstating generalization to independent visual families.

Benchmark leakage is well documented outside plant pathology. SelfClean uses self-supervised representations to identify near-duplicates and label problems in visual datasets [38]. Ramos et al. showed that visual leakage can occur at different levels of similarity and can alter downstream evaluation [39]. In geospatial segmentation, de-duplication and split redesign have changed estimates of model performance when related imagery crossed partitions [40]. Similar findings were reported for geographically disjoint online-mapping benchmarks [41]. This suggests that repeated crops, frames, scenes, or even transformed images should be considered correlated records as opposed to the assumption that file-level uniqueness equates to independence in evaluation.

The presence of plant diseases introduces yet another challenge. Disease distributions tend to be highly imbalanced. Long-tailed structures have been observed in large agricultural data collections [23] and also in long-tailed semantic segmentation research has shown that class support impacts optimization and evaluation [45]. Pixel-wise long-tail objectives have also tackled this issue from another angle [46]. Removal of contaminated records may improve split independence, but at the cost of making estimates for rare diseases less stable. It is important to disentangle these two effects in benchmark auditing.

This work considers four interrelated questions: Does PlantSeg feature cross-split visual families, Is the evaluation different from contamination-aware evaluation in terms of disease distribution, Are the duplicate family exposures better for test cases compared to independently clean ones, and can disease coverage be maintained by group-disjoint reconstruction. The segmentation architectures are used as measurement tools for these questions rather than as objects of architecture ranking.

1.1 Key contributions

- A human-validated audit of PlantSeg that identifies cross-split visual families and separates label-consistent exposure from semantic conflicts.
- A multi-model released-versus-clean re-evaluation with exactly class-count-matched Monte Carlo controls to separate contamination effects from changes in test-set size and disease composition.
- A two-seed matched training-exposure experiment that isolates duplicate-family exposure using an equal-size, disease-class-matched control.
- A long-tail analysis and three group-disjoint benchmark reconstructions that preserve all 115 disease classes and eliminate verified family crossings.
- A reproducible audit-to-repair workflow linking visual-family validation, controlled effect estimation, exposure analysis, uncertainty assessment, and benchmark reconstruction.

Figure 1 summarizes the complete frozen workflow from audit through repaired-split validation.

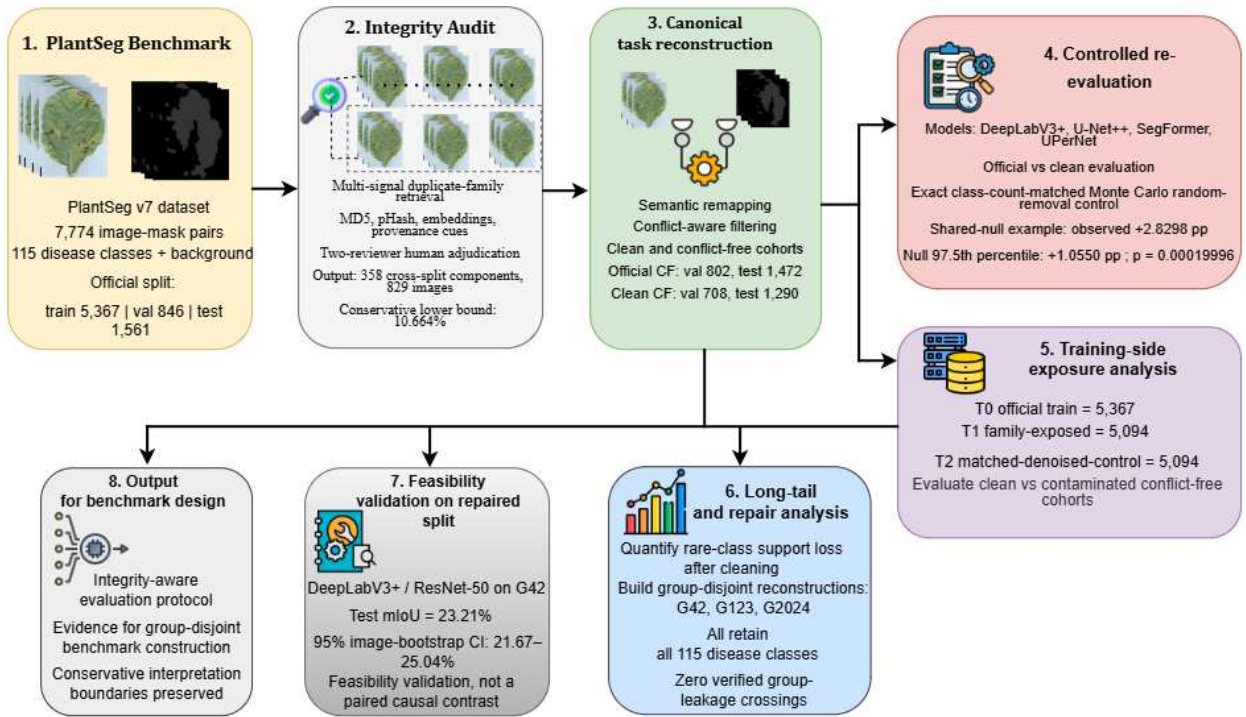


Figure 1. Overall PlantSeg benchmark-integrity workflow

2. Related Work

2.1 Plant-disease segmentation and in-the-wild benchmarks

Plant-disease segmentation has progressed through several modeling families. Convolutional approaches include modified U-Net for leaf-lesion segmentation [16] and an attention residual U-Net for multiclass leaf analysis [17]. DeepLab-based designs have been adapted to difficult field scenes [6] and to lightweight cotton-wilt segmentation [12]. Other work has used adversarial feature fusion for tomato disease segmentation [7] and lightweight transformers for grape disease segmentation [13]. Hybrid transfer-learning approaches [20] and attention-enhanced U-Net variants [22] extend the same line of work to combined segmentation and classification. These studies improve lesion representation and delineation, but they generally evaluate performance on the partitions supplied with the dataset.

Other crop-specific studies illustrate the diversity of downstream goals. Cross-modal fusion has been used to combine complementary signals for disease segmentation [5]. Lesion segmentation has been linked directly to severity estimation in natural field imagery [11], and mango pipelines combine segmentation with disease-severity analysis [21]. Few-shot segmentation has also been explored for annotation-efficient pest and disease analysis in field imagery [10]. Large-scale crop disease and pest datasets exhibit strong class imbalance [23]. These settings motivate evaluation protocols that account for both image dependence and long-tail support.

2.2 Transformers, foundation models, and annotation-efficient segmentation

Transformer and foundation-model approaches have expanded plant-disease analysis through global context modeling, parameter-efficient adaptation, and promptable segmentation. PD-TR applies transformer-based end-to-end disease detection [18], while parameter-efficient transformer adapters have been studied for plant-leaf disease recognition [19]. A broader survey summarizes the role of Vision Transformers in semantic segmentation [34]. Leaf Only SAM demonstrated zero-shot use of the Segment Anything Model for leaf segmentation [14], and later work adapted SAM to multi-disease field imagery [15]. Enhanced multi-scale SAM variants have also been proposed for leaf-disease segmentation [8], as have unsupervised prompt-learning strategies [9]. Independent evaluations show that SAM performance varies substantially across real-world applications [36], and domain-specific adaptation has also been necessary in medical-image segmentation [37]. These findings reinforce the need to evaluate pretrained models on independent benchmark partitions.

Recent reviews consistently identify data quality and generalization as unresolved issues in plant-disease AI. Broad surveys cover the expanding range of deep-learning methods and their deployment challenges [26], while architectural trends are reviewed separately across automated disease-detection systems [27]. Task-specific

reviews narrow this discussion to semantic segmentation in apple disease [28] and integrated detection, segmentation, and classification in maize [29]. Other reviews emphasize persistent practical limitations, data quality, and generalization in plant-disease AI [30], with similar concerns reported in broader crop-disease and computer-vision surveys [32]. Challenges that remain important for deployment are discussed explicitly by Khan et al. [31], and Qadri et al. place them within the wider development of computer-vision methods for plant disease analysis [33]. Domain shift remains a parallel concern in semantic segmentation [35]. Benchmark leakage and training exposure to duplicate visual families, however, remain much less studied experimentally in agricultural segmentation.

2.3 Benchmark integrity, data leakage, and duplicate auditing

Benchmark-integrity studies show why file-level de-duplication is insufficient. SelfClean treats data auditing as a retrieval and scoring problem for near-duplicates and label errors [38]. Ramos et al. distinguish several forms of visual leakage and show that both obvious and subtle overlap can alter downstream reliability [39]. Earlier work on CIFAR likewise showed that near-duplicate purging can change evaluation outcomes [42]. In PlantSeg’s case, these works imply that similarity-derived signals are helpful for candidate identification; however, benchmark claims should be backed by image relationships rather than mere similarity scores.

In addition to the presence of identical pixels, structured dependence is possible without it. De-duplication research on geospatial imagery implies that related images yield overly positive segmentation estimates [40]. Similarly, geographically disjoint evaluation has affected results in online mapping benchmarks [41]. As such, cross-validation guidelines suggest clustering of correlated samples before splitting [43]. Bias-aware data management guidelines make the same general statement: correlations induced during data collection can be reflected in the evaluation process if these correlations are not considered [44]. Correlation family, in plant imagery case, would include samples from the same image, leaf, lesion, scene, acquisition session, or crop. This work formalizes that idea by constructing a family graph and disjoint group reconstruction based on it.

2.4 Long-tailed evaluation and cost of benchmark cleaning

Plants and diseases datasets suffer from imbalance, and benchmark cleaning may decrease support for less frequent classes. DLCPD-25 illustrates the long-tailed character of large crop disease and pest collections [23]. Diffusion-based augmentation has been proposed specifically to improve long-tailed pest classification [24]. Frequency-Based Matcher deals with class imbalance using frequency-aware matching for semantic segmentation [45], while AUCSeg considers pixel-level long tail problem [46]. These results make class support part of benchmark-integrity analysis rather than a separate issue.

A contamination-aware benchmark therefore cannot be repaired by deleting related images without checking the resulting class distribution. Macro-averaged segmentation metrics can become unstable when support for tail classes is reduced, so repair should preserve category coverage wherever possible. The same concern appears in other long-tailed diagnostic settings. Large-scale radiology benchmarks show that rare-category support affects evaluation reliability [47], and challenge analyses of long-tailed chest X-ray classification reach similar conclusions [48]. The relevant lesson for PlantSeg is that decontamination should improve independence without needlessly sacrificing semantic coverage.

2.5 Research gap and study rationale

The literature leaves three practical gaps for a benchmark such as PlantSeg. First, no published evaluation has established a human-validated cross-split visual-family audit of the released partitions. Second, the effect of verified contamination on semantic-segmentation scores has not been separated from changes in class composition. Third, training-side exposure and group-disjoint benchmark repair have not been evaluated together in the same long-tailed plant-disease setting.

This study addresses those gaps through an audit-to-repair design: verified visual-family detection, contamination-aware re-evaluation, disease-class-matched composition controls, a matched T1orT2 exposure experiment, and group-disjoint reconstruction. Table 1 positions these elements relative to representative plant-disease and benchmark-integrity studies.

The primary object of study is the evaluation protocol itself. The segmentation models provide independent measurement channels for assessing how benchmark structure changes the evidence of generalization.

Table 1. Comparative positioning of representative studies relevant to benchmark integrity in plant-disease semantic segmentation.

Study	Focus	Integrity audit	Human validation	Controlled effect	Exposure or repair

Wei et al. [1]	PlantSeg benchmark	No dedicated family audit	Not reported	Not reported	Benchmark creation
Feng et al. [2]	PlantFormer on PlantSeg	Not reported	Not reported	Not reported	Model-focused evaluation
Fan et al. [3]	STAR-Net on PlantSeg	Not reported	Not reported	Not reported	Imbalance-aware training
Liu et al. [4]	ReLeaf-SAM	Not reported	Not reported	Not reported	Released-split model evaluation
Ge et al. [10]	Few-shot field segmentation	Not reported	Not reported	Not reported	Annotation-efficiency focus
Keshavarzi et al. [28]	Plant segmentation review	Data limitations discussed	Not reported	Not reported	No controlled leakage study
Groger et al. [38]	General visual-data auditing	Yes	Issue inspection	Cleaning reliability	No task-specific exposure test
Ramos et al. [39]	Visual-leakage analysis	Yes	Not plant-specific	Reliability impact	No matched exposure design
Adimoolam et al. [40]	Geospatial de-duplication	Yes	Not plant-specific	Seenorunseen comparison	No class-matched exposure control
Lilja et al. [41]	Geographically disjoint mapping	Structured overlap audit	Not reported	Yes	Disjoint-split remedy
Barz & Denzler [42]	CIFAR near-duplicate purging	Yes	Manual or algorithmic	Yes	Duplicate-free replacement
Li et al. [45]	Long-tailed segmentation	No	Not reported	Not reported	Tail-focused evaluation
Han et al. [46]	Pixel-level long-tail segmentation	No	Not reported	Not reported	Tail-focused evaluation
Zheng et al. [47]	Long-tailed diagnosis	Not family-focused	Not reported	Not reported	Rare-category evaluation
Present study	PlantSeg integrity evaluation	Yes; visual-family graph	Two-reviewer adjudication	Multi-model composition control	T0 or T1 or T2 exposure + group-disjoint repair

3. Materials and Methods

3.1 Study design and analytical scope

The benchmark-reliability design was organized around four questions: whether verified cross-split visual families are present; whether released and contamination-aware evaluations differ; whether training-side family exposure benefits related evaluation cases; and whether group-disjoint reconstruction can preserve semantic coverage.

3.1.1 Problem definition and evaluation framework

Let each record contain image x_i , mask y_i , semantic disease label c_i , and released split s_i . Let $M(f, E)$ denote class-macro mean Intersection over Union (mIoU) for model f on cohort E . Equations (1)-(5) define the visual-relation graph, cross-split families, contamination fraction, and semantic conflict used throughout the audit.

$$\mathcal{D} = \{(x_i, y_i, c_i, s_i)\}_{i=1}^N, \quad s_i \in \{\text{tr}, \text{val}, \text{te}\} \quad (1)$$

$$A_{ij} = 1 \Leftrightarrow (i, j) \text{ is accepted as a verified visual relation}; \quad G = (V, E), \quad E = \{(i, j): A_{ij} = 1\} \quad (2)$$

$$\mathcal{C}_{\text{cross}} = \{C_k: |\{s_i: i \in C_k\}| > 1\} \quad (3)$$

$$\rho_{\text{cont}} = \frac{\left| \bigcup_{C_k \in \mathcal{C}_{\text{cross}}} C_k \right|}{N} \quad (4)$$

$$q_k(c) = \mathbf{1}(|\{c_i: i \in C_k\}| > 1) \quad (5)$$

A connected component C_k of the verified visual-relation graph defines one visual family. Cross-split components contribute to the conservative contamination estimate, and disease, class, and host-plant conflicts are tracked separately. Algorithm 1 links these definitions to the evaluation, exposure, and repair analyses.

Algorithm 1. Integrity-aware benchmark evaluation and repair framework.

1. Input & graph construction. Start from released records $\mathcal{D}$. Retrieve candidate cross-split relations using MD5, pHash, structural/deep similarity and provenance cues; accept a relation only after the frozen automated/human validation rule. Build $G=(V, E)$, with connected components C_k defining verified visual families.

$$A_{ij} = 1, \quad |\{s_i: i \in C_k\}| > 1 \text{ [Eqs. (1)-(3)]}.$$

2. Integrity quantification. Compute the conservative lower-bound contamination fraction and flag components with semantic, disease, or host-plant conflict. Contradictory components are quarantined from the primary exposure-mechanism analysis.

$$\rho_{\text{cont}} = \frac{\left| \bigcup_{C_k \in \mathcal{C}_{\text{cross}}} C_k \right|}{N}, \quad q_k(c) = \mathbf{1}(|\{c_i: i \in C_k\}| > 1) \text{ [Eqs. (4)-(5)]}.$$

3. Canonical task & clean cohorts. Apply the frozen 116-output semantic mapping, retain the earliest partition member of each verified family for contamination-aware evaluation, and construct official/clean conflict-free cohorts.

$$M(f, E) = \text{mIoU}(E) \text{ over supported classes [Eqs. (6)-(9)]}.$$

4. Evaluation-side contrast. For each frozen model/seed, quantify the released-to-clean shift and compare it with exactly class-count-matched random removals so that test-size and disease-composition changes are controlled.

$$\Delta_{\text{eval}} = M(f, E_{\text{official}}) - M(f, E_{\text{clean}}); p_{\text{MC}} = \frac{1 + \sum_{b=1}^B \mathbf{1}[\Delta_{\text{null}}^{(b)} \geq \Delta_{\text{eval}}]}{B+1}, \quad B = 5,000 \text{ [Eqs. (10)-(12)]}.$$

5. Training-side exposure contrast. Construct T_1 by removing the 273 label-consistent duplicate-family-exposed training images and T_2 by removing 273 ordinary images with exactly matched semantic-class counts. Train under identical protocols and replicate across seeds 42 and 123.

$$T_1 = T_0 \setminus L, \quad T_2 = T_0 \setminus R, \quad |L| = |R|, \quad n_c(R) = n_c(L). \text{ Estimate } \Delta_{\text{clean}} \text{ and } \Delta_{\text{fam}} \text{ [Eqs. (13)-(15)]}.$$

6. Exposure-by-cohort interaction. Test whether retaining duplicate-family exposure benefits related contaminated cases more strongly than independent clean cases; preserve controlled/quasi-experimental wording.

$$\Delta_{\text{int}} = \Delta_{\text{fam}} + \Delta_{\text{clean}} \text{ [Eq. (16)]}.$$

7. Uncertainty. Use paired image bootstrap for clean cases and leakage-component clustered bootstrap for contaminated cases; carry all members of a verified family together in each clustered resample.

$$\text{CI}_{95\%}(\Delta) = [Q_{0.025}(\Delta_{\text{boot}}), Q_{0.975}(\Delta_{\text{boot}})], B=5,000 \text{ [Eq. (17)]}.$$

8. Long-tail & repair. Quantify support lost after cleaning, then create $G_{42}/G_{123}/G_{2024}$ group-disjoint reconstructions that preserve all disease classes while constraining each verified family to one partition.

$$\forall C_k, \quad |\{s_i: i \in C_k\}| = 1; \text{ target proportions } \approx 0.70/0.10/0.20 \text{ [Eq. (18)]}.$$

9. Feasibility validation. Run the prespecified DeepLabV3+–ResNet50 seed-42 validation on G42 using the same canonical training recipe. Interpret the resulting score only as repaired-protocol feasibility.

Output: zero verified group crossings, all 115 disease classes retained, G42 test mIoU=23.21% (95% image-bootstrap CI 21.67–25.04%).

3.2 Dataset and semantic segmentation task

The audited dataset was PlantSeg v7 (Zenodo 10.5281/zenodo.17719108; accessed 15 September 2026), containing 7,774 image-mask pairs, 115 disease categories, and 34 host plants [1]. The released split contained 5,367 training, 846 validation, and 1,561 test images. Background was class 0 and disease classes were mapped to IDs 1–115, giving 116 output channels. The archive version and checksum were fixed before analysis so every audit relation and split reconstruction remained release specific.

The mask audit found 7,762 samples consistent with the verified semantic identity and 12 that required in-memory disease-ID correction. Source files were not modified. These corrections prevented target-ID inconsistencies from being mistaken for model or contamination effects; the complete cohort and loader definitions are given in Supplementary Tables S2–S3.

3.3 Cross-split integrity audit

The integrity audit combined exact matching with perceptual and structural retrieval. Message Digest 5 (MD5) hashes identified exact matches; perceptual hash (pHash) distance, Structural Similarity Index Measure (SSIM), color-histogram correlation, keypoint evidence, and Random Sample Consensus (RANSAC) geometry were used to retrieve transformed duplicates and same-scene or cropped candidates. These signals generated candidates for review rather than serving as ground-truth labels.

A relation entered the verified graph only when the evidence supported a common image or acquisition identity, including modified versions or another frame or crop of the same recognizable plant, leaf, lesion, or scene. Visually similar but independent images were excluded, and provenance matches were treated as supporting evidence only. The test of expansion $pHash \leq 12$ on novel cross-split pairs used the stored descriptors for each pair. Visual families were generated through connected components; since the recall of retrieval is not exhaustive, the contamination rate is stated as a lower bound.

3.4 Human validation and candidate triage

Two reviewers independently assessed 100 substantive image-pair cases plus six attention checks, assigning both a binary leakage decision and a relation taxonomy. Disagreements were adjudicated. A Random Forest (RF) triage model was then fitted to 94 human-consensus cases using pHash, SSIM, keypoint, and RANSAC features. Model discrimination was assessed with out-of-fold (OOF) predictions. Operational thresholds were fixed at $p \geq 0.51$ for leakage and $p \leq 0.49$ for non-leakage, with manual review between them; details are provided in Supplementary Methods Note S1 and Tables S6–S7.

The RF used 500 trees, equal class weights, minimum leaf size 2, median imputation, and a fixed random state. Thresholds were locked before the targeted 90-pair expansion. More stringent thresholds were examined only as sensitivity analyses and did not alter the accepted graph decisions materially.

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \quad P_e = p_{1,yes}p_{2,yes} + p_{1,no}p_{2,no}$$

$$AC1 = \frac{P_o - P_{e,AC1}}{1 - P_{e,AC1}}, \quad P_{e,AC1} = 2\bar{p}(1 - \bar{p}), \quad \bar{p} = \frac{p_{1,yes} + p_{2,yes}}{2}$$

The frozen Random-Forest decision rule was:

$$d(p) = \begin{cases} \text{leakage,} & p \geq 0.51, \\ \text{non-leakage,} & p \leq 0.49, \\ \text{manual review,} & 0.49 < p < 0.51, \end{cases}$$

3.5 Contamination-aware split construction

For contamination-aware evaluation, later-partition members of verified cross-split families were removed with priority given to retaining training members, yielding clean validation and test cohorts of 718 and 1,290 images. Components with contradictory semantic labels were quarantined separately, producing conflict-free official and clean cohorts. Label-consistent exposed training images were retained for the T1 or T2 experiment. This separation allowed visual-family exposure and contradictory supervision to be analyzed as distinct integrity problems.

3.6 Semantic segmentation benchmarking

Four established architectures were evaluated with segmentation_models_pytorch 0.5.0: DeepLabV3+-ResNet50, U-Net++-EfficientNet-B0, SegFormer-MiT-B0, and UPerNet-ResNet50. DeepLabV3+ was trained with seeds 42 and 123; the other models used seed 42. The experiment tested whether the released-versus-clean direction was consistent across architectures, not whether one architecture was more susceptible to contamination. Images and masks were resized with aspect-ratio preservation and zero padding to 512x512; the full configuration is given in Supplementary Table S3.

All canonical runs used the same preprocessing and augmentation policy. Horizontal flips, scale changes, and modest intensity perturbations were applied during training. Background was class 0, disease classes were 1-115, and label 255 was reserved as the ignore index. Padding used background 0, so padded pixels contributed to background loss and metric counts.

Models were optimized with multiclass cross-entropy and stochastic gradient descent (SGD), using the same polynomial learning-rate schedule, effective batch size, epoch budget, and early-stopping rule. Mixed-precision training was used on the graphics processing unit (GPU). Checkpoints were selected only by cleanorconflict-free validation mIoU, ensuring that model selection did not use the contaminated official validation score. Exact hyperparameters are reported in Supplementary Table S3.

As a plausibility check, DeepLabV3+-ResNet50 seed 42 reached 24.30% mIoU on the released test set, close to the 25.08% value reported with the PlantSeg reference implementation [1]. The comparison was used only to confirm task consistency because the present implementation used a different software stack.

The optimization objective and polynomial learning-rate schedule are given below for reproducibility. Pixels assigned the reserved ignore index are excluded; zero-valued padding remains part of the background class. $\Omega_{\text{valid}} = \{p \in \Omega: y(p) \neq 255\} p_{\theta}(c | x, p)$

$$\mathcal{L}_{\text{CE}}(\theta) = -\frac{1}{|\Omega_{\text{valid}}|} \sum_{p \in \Omega_{\text{valid}}} \sum_{c=0}^{115} \mathbf{1}[y(p) = c] \log p_{\theta}(c | x, p) \quad (6)$$

$$\eta_t = \eta_{\min} + (\eta_0 - \eta_{\min}) \left(1 - \frac{t}{T}\right)^{\gamma}, \quad \eta_0 = 0.01, \quad \eta_{\min} = 10^{-4}, \quad \gamma = 0.9 \quad (7)$$

3.7 Evaluation metrics

Intersection over Union (IoU) was computed from pixel-level confusion counts. Mean IoU (mIoU) was the class-macro average over classes present in either predictions or references; disease-only mIoU excluded background. Mean class accuracy (mAcc) and aggregate pixel accuracy (aAcc) were secondary measures. Pixels with label 255 were excluded from metric accumulation, while zero-valued padding contributed to the background class.

$$\text{IoU}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c + \text{FN}_c} \quad (8)$$

$$\text{mIoU}(E) = \frac{1}{|\mathcal{K}_E|} \sum_{c \in \mathcal{K}_E} \text{IoU}_c \quad (9)$$

$$\Delta_{\text{eval}}(m) = M(f_m, E_{\text{official}}) - M(f_m, E_{\text{clean}}) \quad (10)$$

The primary evaluation-side estimand was the difference between released and contamination-aware mIoU for the same frozen checkpoint. Using the identical checkpoint under both evaluations isolates the effect of evaluation-set composition from training stochasticity.

3.8 Class-matched Monte Carlo control

An exactly class-count-matched Monte Carlo test was used to determine whether the released-to-clean shift could be explained by the smaller test set or altered disease composition. Cleaning removed 271 of 1,561 test images. In each of 5,000 draws, the same number of images was removed within every affected disease class, so the observed disease-specific removal profile was preserved while image identity varied.

The resulting null distribution answers whether the observed shift is unusual relative to other removals with the same disease composition. The plus-one correction sets the minimum attainable Monte Carlo probability at $1/(B+1)$.

For each draw, mIoU was recomputed from stored per-image intersectionorunion arrays and compared with the observed shift using the plus-one Monte Carlo probability:

$$\Delta_{\text{null}}^{(b)} = M(f, E_{\text{official}}) - M(f, E_{\text{official}} \setminus R^{(b)}), \quad b = 1, \dots, B \quad (11)$$

$$p_{\text{MC}} = \frac{1 + \sum_{b=1}^B \mathbf{1}[\Delta_{\text{null}}^{(b)} \geq \Delta_{\text{eval}}]}{B+1}, \quad B = 5,000 \quad (12)$$

Each Monte Carlo sample preserved the observed disease-specific removal counts exactly; only the identity of the removed images changed.

The 97.5th null percentile was used as a reference threshold. Common-class and disease-only decompositions were also computed to check whether the observed shift depended on classes disappearing after cleaning.

The five modelorseed composition-control tests were treated as one multiplicity family and adjusted with the Holm procedure. The T1orT2 exposure analyses addressed a separate mechanism question and were reported with bootstrap intervals rather than included in this family.

Formally, for ordered probabilities $p_{(1)} \leq \dots \leq p_{(m)}$, the Holm step-down rule compares $p_{(i)} \leq \frac{\alpha}{m-i+1}$ sequentially and stops at the first non-rejection. Here $m=5$ because the five cross-modelorseed composition-control tests formed one prespecified family.

3.9 Training-side duplicate-family exposure experiment

Training-side exposure was evaluated with DeepLabV3+-ResNet50 under three training sets: T0, the released training split (5,367 images); T1, after removing 273 label-consistent duplicate-family-exposed images; and T2, after removing 273 ordinary images with exactly the same disease-class removal counts. T1 and T2 therefore contained the same number of images and the same class-removal profile.

The principal comparison was T1 versus T2. Seed 42 was replicated with seed 123 using the same T1orT2 image sets and training protocol, changing only training stochasticity.

Three candidate T2 controls were generated before freezing one control set. Disease-class removal counts were identical by design, while foreground fraction, component count, and mask-morphology distributions were checked as secondary balance diagnostics. Because these covariates were only approximately balanced, the experiment is described as controlled or quasi-experimental rather than fully randomized.

Secondary balance was evaluated with foreground fraction, component count, and a frozen nine-feature mask-morphology descriptor; definitions and Gaussian-mixture details are reported in Supplementary Table S10.

Across the candidate controls, standardized mean difference (SMD) and total variation (TV) diagnostics indicated modest residual imbalance. These diagnostics were used to characterize the matched control rather than to claim randomization; full values are reported in Supplementary Table S12. $\text{SMD} = \frac{\bar{x}_L - \bar{x}_R}{\sqrt{(s_L^2 + s_R^2)/2}}$ $\text{TV}(P, Q) = \frac{1}{2} \sum_k |P_k - Q_k|$

Two prespecified evaluation channels separated independent generalization on the clean test set from performance on related contaminated, conflict-free cases. A selective advantage on the latter would indicate family-specific exposure rather than a stable general training benefit.

Let L denote the 273 label-consistent duplicate-family-exposed training images and R the frozen matched-random control set:

$$\mathcal{T}_0 = \mathcal{T}_{\text{official}}, \quad \mathcal{T}_1 = \mathcal{T}_0 \setminus L, \quad \mathcal{T}_2 = \mathcal{T}_0 \setminus R \quad (13)$$

$$|L| = |R| = 273, \quad n_c(R) = n_c(L), \quad \forall c \quad (14)$$

$$\Delta_{\text{clean}} = M(f_{T_1}, E_{\text{clean}}) - M(f_{T_2}, E_{\text{clean}}), \quad \Delta_{\text{fam}} = M(f_{T_2}, E_{\text{contam}}) - M(f_{T_1}, E_{\text{contam}}) \quad (15)$$

$$\Delta_{\text{int}} = \Delta_{\text{fam}} + \Delta_{\text{clean}} = [M(f_{T_2}, E_{\text{contam}}) - M(f_{T_1}, E_{\text{contam}})] - [M(f_{T_2}, E_{\text{clean}}) - M(f_{T_1}, E_{\text{clean}})] \quad (16)$$

Equation (16) is positive when retaining duplicate-family exposure benefits related contaminated cases more than independent clean cases. The interaction is interpreted as a controlled mechanism contrast because T1 and T2 match training size and disease-class removal profile.

3.10 Bootstrap analysis of training-side exposure

Bootstrap inference used stored per-image intersection and union counts without rerunning the models. Clean T1-T2 differences used 5,000 paired-image resamples. Contaminated-cohort differences used 5,000 visual-family clustered resamples so all members of a verified family remained together.

Pairing preserved the within-image T1orT2 comparison on clean cases, while clustering accounted for dependence among contaminated cases from the same verified family.

Percentile confidence intervals (CIs) were calculated separately for each seed. Empirical bootstrap tail probabilities were one-sided for the prespecified contaminated-cohort direction and two-sided for clean effects. Two-seed summaries are descriptive because two fitted seeds do not estimate population-level training variance.

$$CI_{95\%}(\Delta) = \left[Q_{0.025} \left(\{\Delta_{\text{boot}}^{(b)}\}_{b=1}^B \right), Q_{0.975} \left(\{\Delta_{\text{boot}}^{(b)}\}_{b=1}^B \right) \right] \quad (17)$$

The same cohort-specific resampling was applied to the exposure-by-cohort interaction in Equation (16); Supplementary Table S1 summarizes the estimands and inference scheme.

3.11 Long-tail support analysis

Long-tail effects were quantified from test-set class support before and after cleaning using median class support and counts of classes with fewer than 10 and fewer than 5 images. Spearman correlation assessed the relationship between class support and class IoU.

3.12 Group-disjoint benchmark reconstruction

After contradiction quarantine, each label-consistent visual family was assigned wholly to one partition under approximate 70or10or20 trainorvalidationortest proportions. Reconstructions were generated with seeds 42, 123, and 2024 and checked for zero verified family crossings and retention of all disease classes.

$$\forall C_k, \quad |\{s_i: i \in C_k\}| = 1, \quad \text{target split proportions} \approx (0.70, 0.10, 0.20) \quad (18)$$

Equation (18) defines the group-disjoint constraint. Feasibility was evaluated with one prespecified DeepLabV3+-ResNet50 seed-42 run on G42 using the canonical training recipe. G42 contained 5,262or756or1,503 trainorvalidationortest images, retained all 115 disease classes in every partition, and had zero verified family crossings.

Checkpoint selection used G42 validation mIoU only. Because validation performance was still improving at epoch 30, the resulting test score is treated as evidence that the repaired protocol is trainable rather than as an optimized performance ceiling.

4. Results

Results revolve around the verified audit, evaluation side shift, training exposure, long-tail effects, and group-disjoint fix.

4.1. Cross-split contamination and human validation

The verified graph consisted of 358 cross-split elements that included 829 out of 7,774 images, resulting in the conservative contamination lower bound of 10.664%. Semantic conflicts were found in 105 elements; conflicting images were isolated from the primary exposure analysis. The more extended provenance-sensitive screening identified additional images, although they were not included into the verified graph without any visual evidence. Human validation confirmed the audit choices. In 100 representative examples, binary agreement between reviewers was 94.0% (Cohen's kappa=0.864; Gwet's AC1=0.892), and resolved positive validation stratum consisted of 48 confirmed contamination cases. Detailed matrices and confidence intervals are reported in Supplementary Table S8.

The Random Forest triage model showed high out-of-fold discrimination on the 94 human-consensus cases, with receiver operating characteristic area under the curve (ROC-AUC) 0.9979 and precision-recall area under the curve (PR-AUC) 0.9991. These values describe calibration-stage diagnostics rather than external validation. Threshold-sensitivity results are given in Supplementary Table S7.

More stringent triage thresholds primarily transferred cases to manual review instead of reversing accepted decisions, indicating that the verified graph was not driven by marginal classifier probabilities.

Conflict structure is summarized in Table 2, and Figure 2 shows representative exact, transformed, same-scene, and excluded visually independent relationships.

Table 2. Final contamination audit and human-validation outcomes.

Outcome	Result	Interpretation
Dataset size	7,774 images	Canonical audited image-mask pairs
Cross-split components	358	Verified cross-split connected components spanning released splits
Implicated images	829 (10.664%)	Conservative lower-bound contamination
Class-conflicting components	105	Component contains conflicting semantic class labels
Disease-conflicting components	101	Component contains conflicting disease labels
Plant-conflicting components	59	Component contains conflicting host-plant labels
Binary human agreement	94.0%; $\kappa=0.864$; AC1=0.892	Independent reviewer validation
Resolved positive precision	48or48 = 100%	95% Wilson CI: 92.59–100%
RF triage classifier	Accuracy 98.94%; precision 100%; recall 98.44%	Out-of-fold discrimination or probability diagnostics; n=94 substantive consensus cases

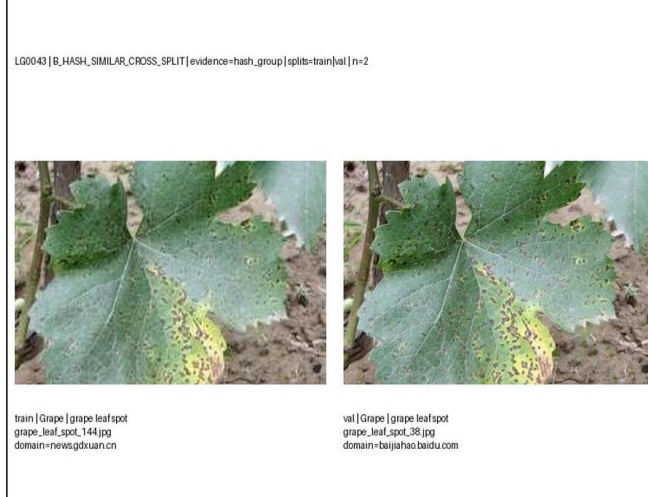
A. Exact duplicate + semantic conflict



B. Modified/near-identical copy



C. Same scene / same subject



D. Same-source but independent images (excluded)



Figure 2. Representative PlantSeg audit relationships: exact duplicate, modified duplicate, same-scene or same-subject relation, and a provenance-matched but visually independent negative example. The examples distinguish verified leakage identity from visual or provenance similarity alone.

4.2. Contamination effects across segmentation models

All four seed-42 models scored higher on the released evaluation than on the contamination-aware clean evaluation. DeepLabV3+, SegFormer, and UPerNet exceeded the 97.5th percentile of the disease-class-matched Monte Carlo null; U-Net++ showed the same direction but did not reach the prespecified significance threshold. Holm adjustment retained the four model or seed results with raw $p=0.00020$.

The released-to-clean shift ranged from +0.20 to +2.83 percentage points across the seed-42 architectures. Model ordering was unchanged between the two evaluation protocols, indicating that contamination affected absolute evidence more clearly than relative ranking in this experiment. Complete model-wise values are reported in Table 3.

DeepLabV3+ seed 42 decreased from 24.30% released to 21.47% clean mIoU, a +2.83-point released-minus-clean shift. Seed 123 reproduced the pattern with a +2.05-point shift. Both exceeded their class-count-matched null thresholds (Monte Carlo $p=0.00020$), reducing reliance on a single training realization.

Table 3 and Figure 3 summarize the cross-model and replicated DeepLabV3+ results. Between-architecture differences in shift magnitude are not interpreted as susceptibility rankings because the non-DeepLab models were evaluated with one training seed.

Supplementary Figure S1 reports the frozen training histories and documents that some models were still improving near the 30-epoch limit.

Table 3. Cross-model official-versus-clean mIoU and exactly class-count-matched Monte Carlo composition-control results. Here, pp denotes percentage points; null percentiles are reference thresholds, not confidence intervals (CIs) for observed shifts.

Model	Seed	Official mIoU (%)	Clean mIoU (%)	Shift (pp)	Shared null 97.5% (pp)	Monte Carlo p
DeepLabV3Plus (resnet50)	42	24.30	21.47	+2.83	+1.0550	0.00020
UnetPlusPlus (efficientnet-b0)	42	6.16	5.96	+0.20	+0.27	0.08098
Segformer (mit b0)	42	12.04	10.81	+1.23	+0.65	0.00020
UPerNet (resnet50)	42	29.78	27.13	+2.65	+1.05	0.00020
DeepLabV3Plus (resnet50)	123	24.94	22.89	+2.05	+0.92	0.00020

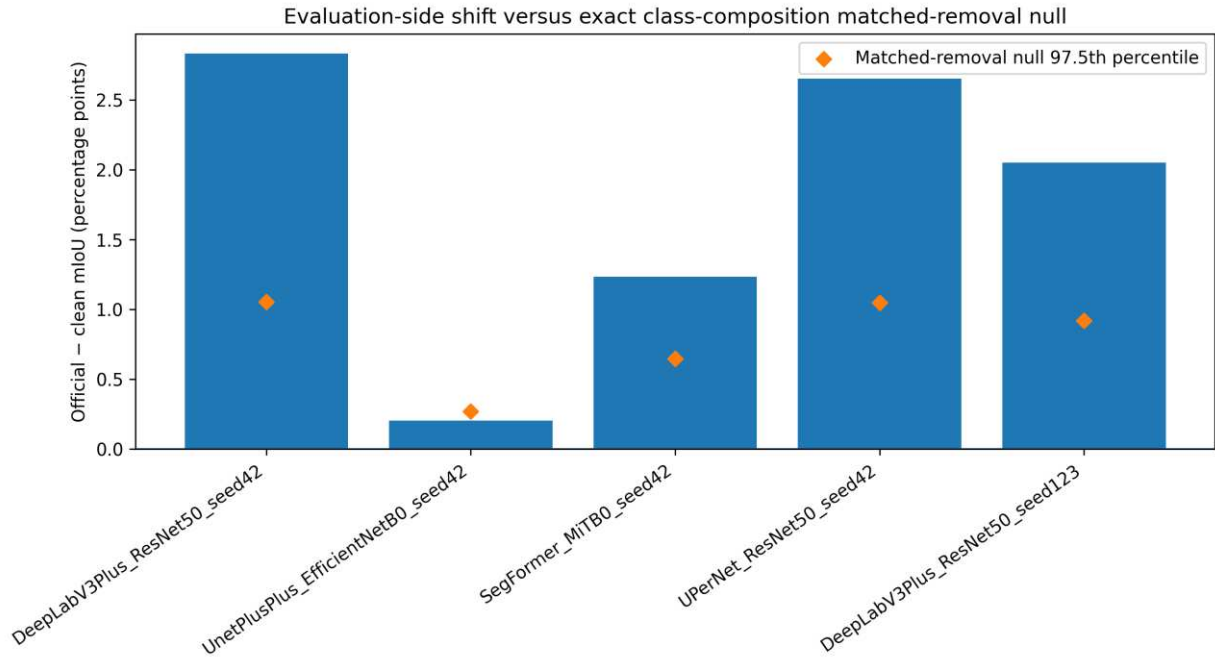


Figure 3. Released-minus-clean mIoU shift for the five model/seed evaluations relative to the shared class-count-matched Monte Carlo null.

4.3. Training-side exposure on clean and contaminated cases

T1 and T2 each contained 5,094 training images. On the independent clean test set, T1-T2 effects were small and changed direction across seeds: +0.40 percentage points for seed 42 and -1.32 for seed 123, with confidence intervals spanning or approaching zero. The experiment therefore did not show a reproducible clean-generalization benefit from duplicate-family exposure.

The pattern was different on related contaminated, conflict-free cases. Retaining family exposure in T2 produced advantages of +11.04 percentage points for seed 42 and +15.20 for seed 123; the visual-family clustered confidence intervals remained above zero for both seeds. Table 4 reports the corresponding cohort scores and uncertainty.

The exposure-by-cohort interaction was also positive in both seeds (+11.44 and +13.88 percentage points), showing that the exposure advantage was concentrated in related contaminated families rather than independent clean cases. Figure 5 summarizes these contrasts.

Table 4. Controlled T1 versus T2 training-side results and bootstrap uncertainty. Contam.-CF denotes the contaminated, conflict-free cohort; pp denotes percentage points.

Seed	Clean T1 (%)	Clean T2 (%)	Clean T1-T2 (pp)	Clean 95% CI (pp)	Contam.-CF T1 (%)	Contam.-CF T2 (%)	Contam. T2-T1 (pp; 95% CI)
42	22.44	22.04	+0.40	[-0.78, +1.45]	26.38	37.43	+11.04 [+6.34, +15.76]
123	21.81	23.13	-1.32	[-2.36, +0.02]	25.37	40.57	+15.20 [+9.65, +19.60]

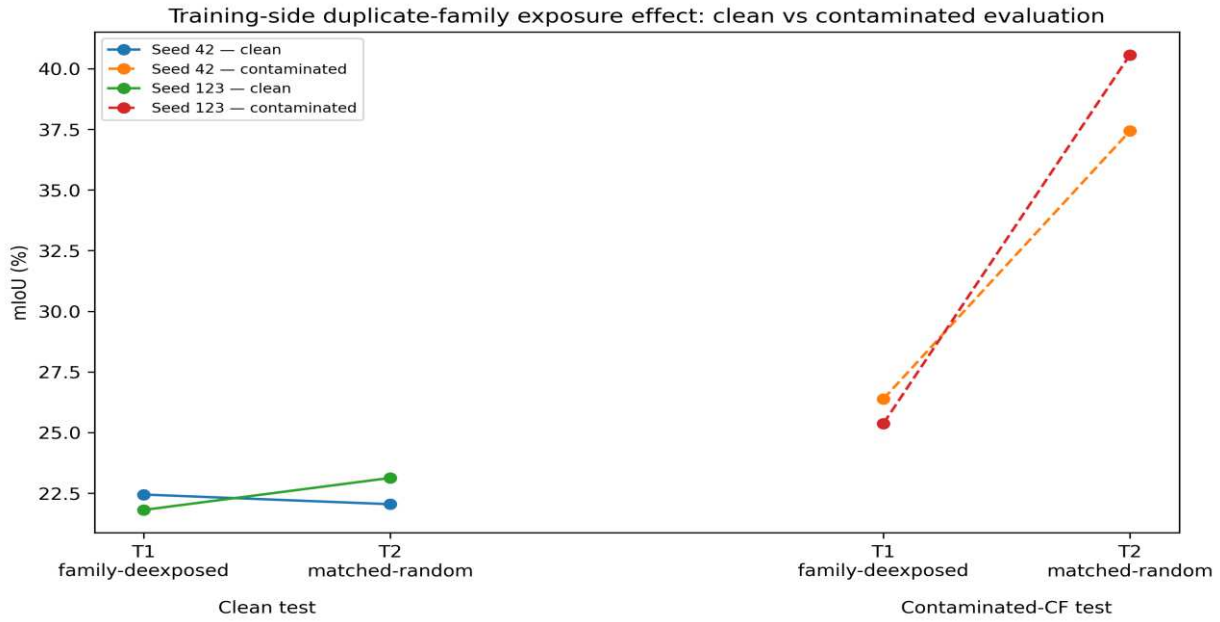


Figure 4. Seed-wise DeepLabV3+ T1orT2 performance on clean and contaminated, conflict-free cohorts.

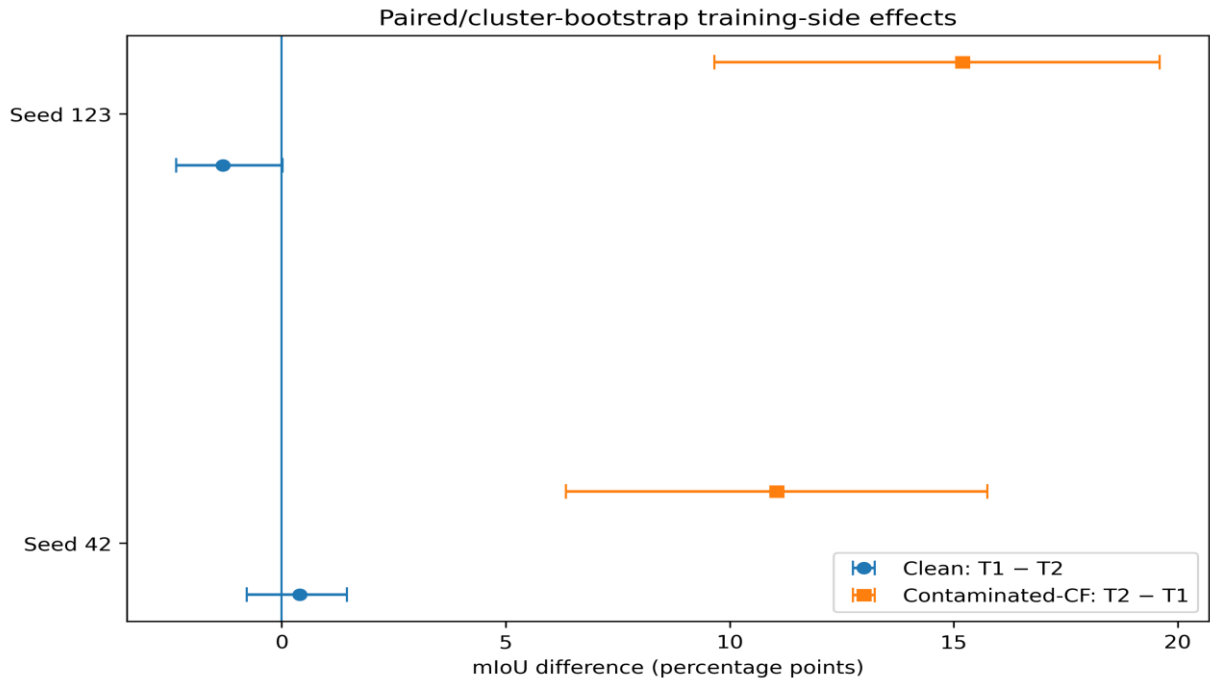


Figure 5. Seed-specific clean and contaminated-cohort exposure effects with bootstrap intervals.

4.4. Long-tail effects of contamination-aware cleaning

Cleaning reduced the test set from 1,561 to 1,290 images while retaining the same 114 disease classes represented in the released test cohort. Median support fell from 11 to 9 images, and the number of classes with fewer than 10 test images increased from 48 to 61. Table 5 gives the full support distribution.

Class support was positively associated with class IoU in both released and clean evaluations (Figure 6). Cleaning therefore improved split independence while increasing sparsity for rare classes, which motivates preserving the complete class-level decomposition in Supplementary Table S11.

Table 5. Effect of contamination-aware cleaning on test-set class support.

Metric	Official test	Clean test
Number of images	1,561	1,290
Disease classes represented	114	114
Median images per disease class	11	9
Classes with <10 images	48	61
Classes with <5 images	20	26
Support-IoU Spearman ρ	0.605	0.693
Support-IoU p-value	9.71e-13	1.26e-17

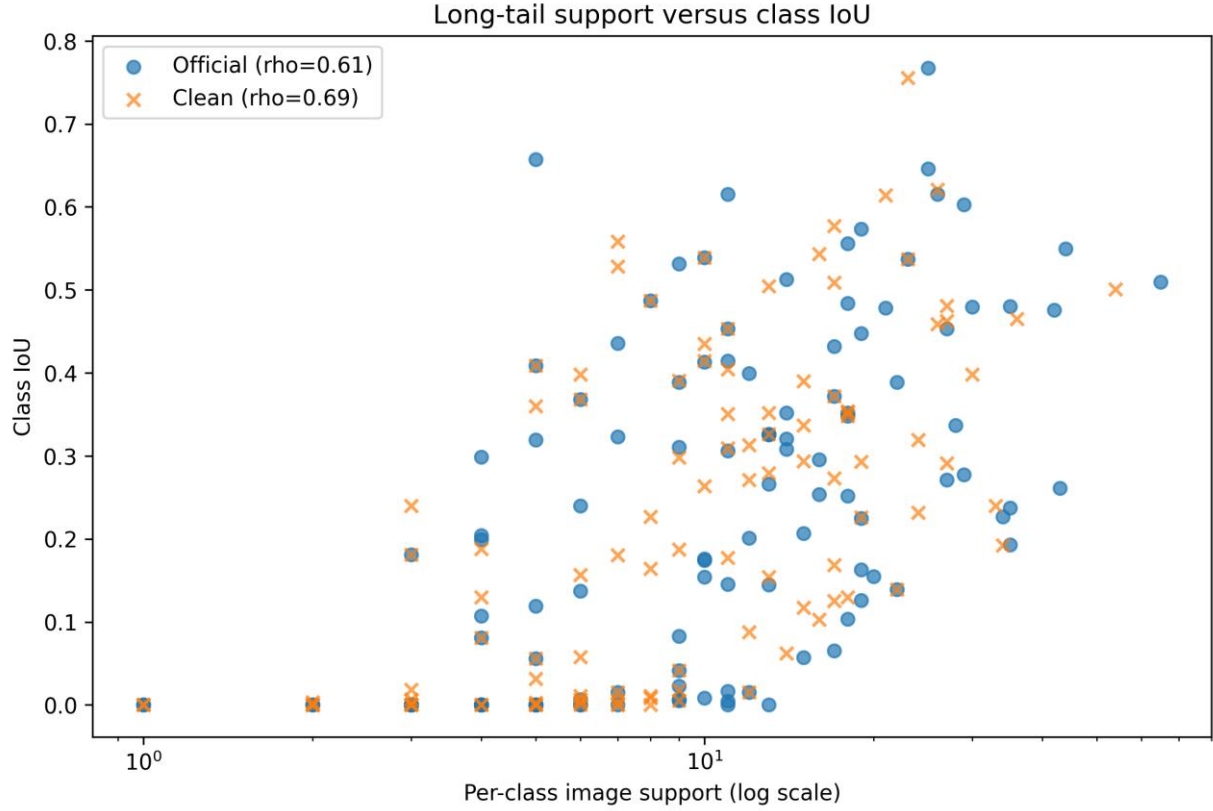


Figure 6. Relationship between per-class image support and class IoU under released and clean evaluation.

4.5. Group-disjoint benchmark reconstruction

The three group-disjoint reconstructions retained 7,521 images, all 115 disease classes in every partition, and zero verified family crossings. Their test sets contained about 1,500 images, substantially closer to the released test size than the deletion-only clean cohort.

The prespecified DeepLabV3+-ResNet50 run on G42 reached 23.21% test mIoU, with a 95% image-bootstrap confidence interval of 21.67-25.04%. Because the model was trained on a different partition, this result is used only to show that the repaired protocol is operational and trainable. Tables 6 and 7 report the split composition and secondary metrics.

Table 6. Group-disjoint benchmark reconstruction outcomes.

Seed	Total	Train	Validation	Test	Classes in each split	Group-leakage violations
42	7521	5262	756	1503	115or115or115	0
123	7521	5263	755	1503	115or115or115	0
2024	7521	5263	753	1505	115or115or115	0

Table 7. Canonical DeepLabV3+ validation of the primary G42 group-disjoint reconstruction.

Item	G42 validation result
Architecture or encoder	DeepLabV3+ or ResNet-50
Training seed	42
Train or validation or test images	5,262 or 756 or 1,503
Disease classes in each split	115 or 115 or 115

Item	G42 validation result
Verified duplicate-family crossings	0
Maximum epochs or selected epoch	30 or 30
Best validation mIoU	23.25%
Test semantic mIoU	23.21%
95% image-bootstrap CI for test mIoU	21.67–25.04% (B=5,000)
Disease-only mIoU	22.62%
Mean class accuracy	33.82%
Aggregate pixel accuracy	88.63%
Interpretation	Feasibility validation of the repaired benchmark; not a paired causal contrast with the released split.

4.6. Robustness and diagnostic analyses

4.6.1. Retrieval validation and targeted expansion

Validation checks confirmed the reliability of the audit decisions across reviewed positive, negative, and difficult candidate strata. Detailed Wilson intervals, classifier diagnostics, and threshold sensitivity results are reported in Supplementary Tables S4 and S7.

A targeted expansion of 90 previously unsurfaced pairs with pHash distance ≤ 12 identified additional accepted relations across several perceptual-distance strata (Figure 7). The decreasing yield at larger distances supports using multiple similarity signals and human review rather than a single perceptual-hash cutoff.

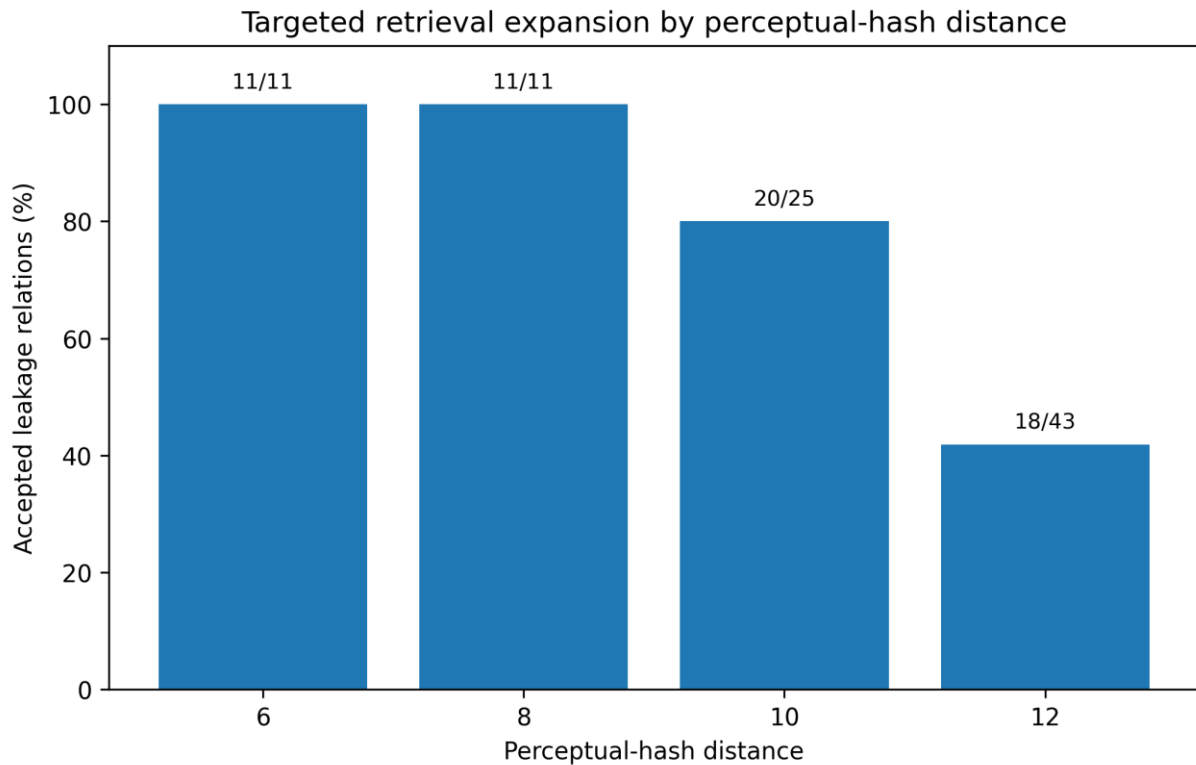


Figure 7. Accepted leakage relations across the targeted pHash-distance expansion.

4.6.2. Composition-control decomposition and matched-control balance

Composition-control decomposition showed that the released-minus-clean shift persisted when the comparison was restricted to common classes and when background was excluded. The class-level shift was larger in disease classes from which a greater fraction of test images had been removed; complete decompositions and T1orT2 balance diagnostics are provided in Supplementary Table S12.

For DeepLabV3+ seed 42, the observed released-minus-clean shift lay beyond the 97.5th percentile of the shared 5,000-draw class-count-matched null (Figure 8). The same draw set was used across all modelorseed evaluations.

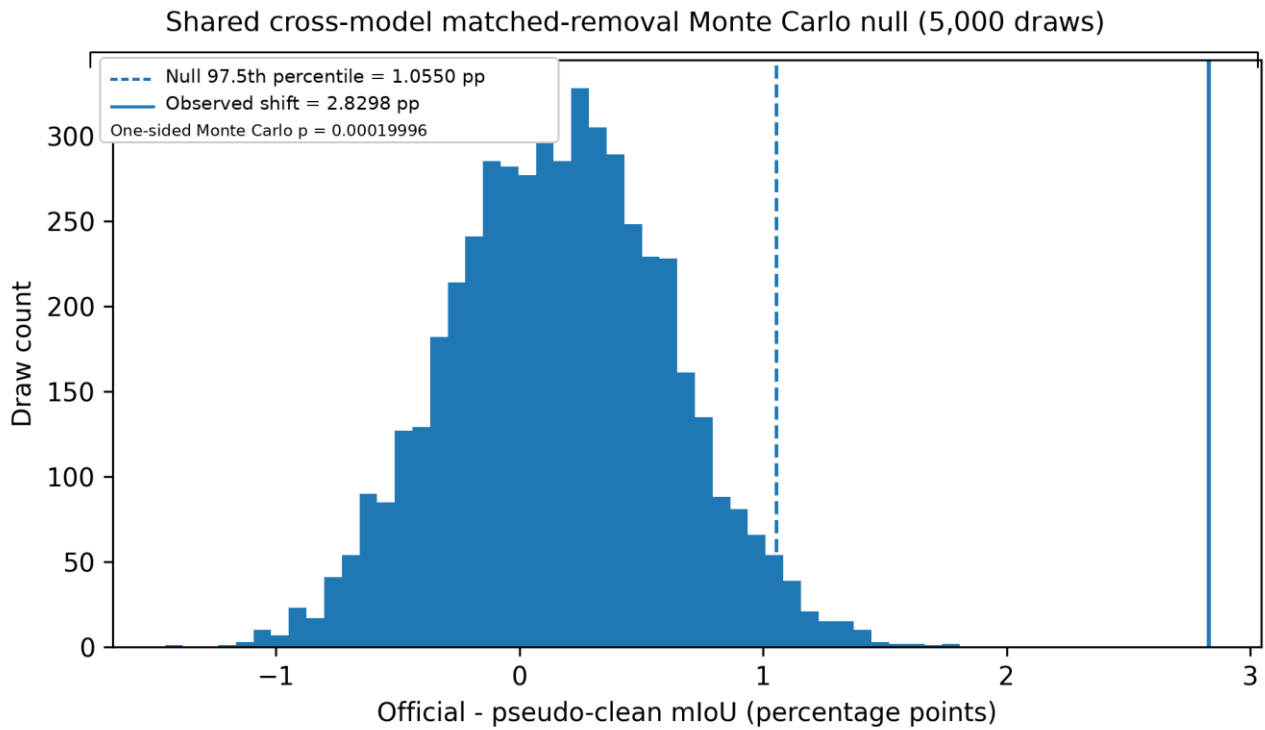


Figure 8. Shared class-count-matched Monte Carlo null for the DeepLabV3+ seed-42 released-minus-clean shift.

At class level, disease classes losing a larger fraction of test support tended to show larger released-minus-clean IoU differences (Figure 9). The relationship was heterogeneous and is treated as descriptive rather than causal; the full class-level analysis is provided in Supplementary Table S11.

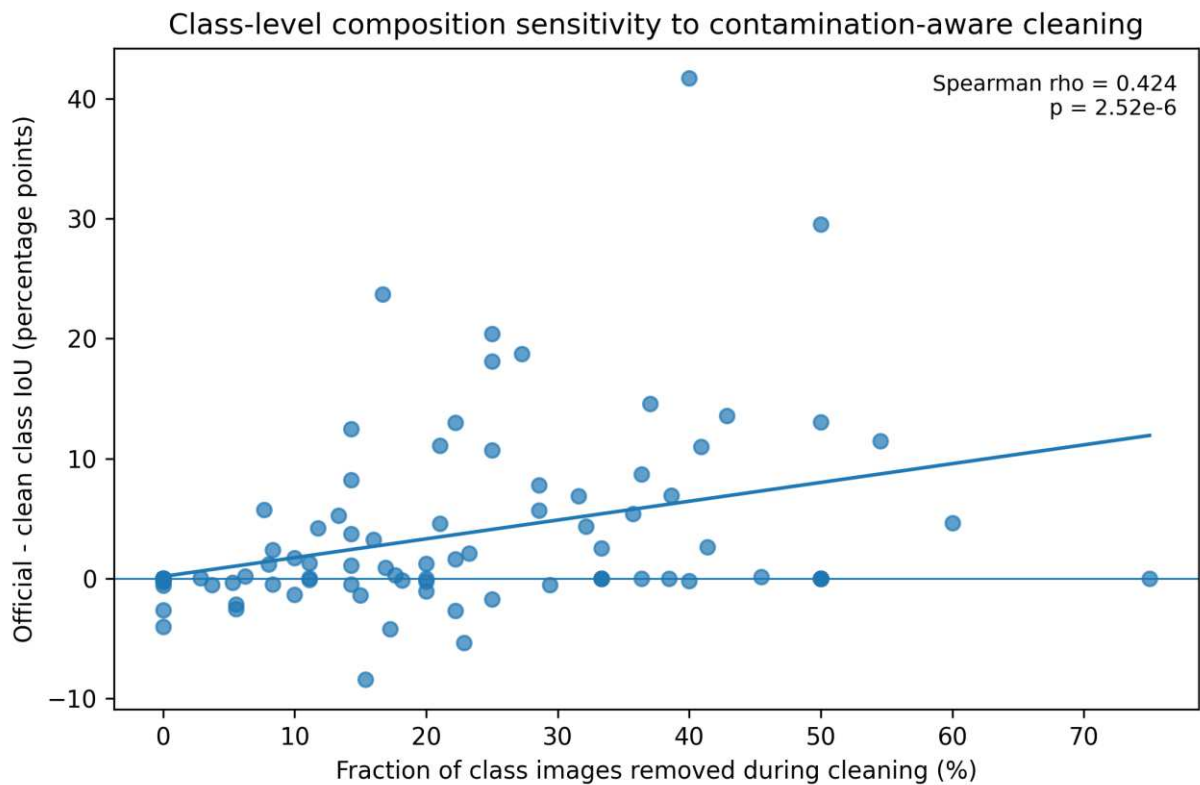


Figure 9. Class-level removed fraction versus released-minus-clean IoU difference across supported disease classes.

4.6.3. Exposure-by-cohort interaction

The exposure-by-cohort interaction was positive for both DeepLabV3+ seeds, providing a compact summary of the mechanism experiment: retained duplicate-family exposure benefited related contaminated cases much more than independent clean cases (Table 8).

Table 8. Exposure-by-cohort interaction derived from the frozen T1orT2 predictions (Equation 16). pp denotes percentage points; CI denotes confidence interval.

Seed	Clean T2-T1 (pp)	Contaminated-CF T2-T1 (pp)	Interaction Delta_int (pp)	95% bootstrap CI (pp)	Empirical one-sided tail probability
42	-0.40	+11.04	+11.44	[+6.47, +16.44]	0.00020
123	+1.32	+15.20	+13.88	[+8.18, +18.52]	0.00020

5. Discussion

The results establish a coherent benchmark-integrity pattern. Human-validated cross-split families were present in PlantSeg; released evaluation was more favorable than contamination-aware evaluation across several architectures; and the controlled exposure experiment localized the largest advantage to related contaminated cases. Group-disjoint reconstruction then showed that split independence can be improved without sacrificing disease-class coverage.

5.1. Benchmark contamination as an evaluation problem rather than a simple duplication count

The integrity problem extends beyond byte-identical duplication. Modified versions or same scene/crop images retain lesion appearance, spatial arrangement, and contextual background through partitioning despite having different pixel arrays. This therefore justifies the use of visual family-based grouping rather than file-level de-duplication, and is supported by general research into near-duplicates and visual leakage auditing [39]. In semantic segmentation, this reliance is particularly crucial since dense predictions can make use of spatial information alongside object appearances. Repeated scenes or crop images could retain the appearance, scaling, vein pattern, and structural background of the lesions such that the evaluation image is somewhat familiar but not a duplicate.

Semantic conflict is an issue distinct from repeated exposure. Families with contradictory class, disease, or host-plant labels were quarantined before the exposure analysis so the training-side mechanism was not confounded by inconsistent supervision. Label-consistent families, by contrast, can be retained safely if all members are assigned to the same partition. The reported contamination fraction should be read as a verified lower bound because the retrieval stage was intentionally conservative; exhaustive recall cannot be established. Provenance-only matches were not counted as leakage without supporting visual evidence.

5.2. Released versus contamination-aware evaluation

The cross-model results show that benchmark contamination affected the absolute evidence of generalization. Every seed-42 architecture scored higher on the released evaluation, and the DeepLabV3+ result reproduced under a second seed. Disease-class-matched Monte Carlo controls showed that the main shifts were larger than expected from test-set size and composition changes alone. Relative model ordering remained stable, so the study does not support an architecture-susceptibility ranking. The more important point is that a model ranking can remain unchanged even when the absolute estimate of independent generalization is optimistic. The difference is significant when improvements on architecture are only few percentages. It is very useful for class-dependent metrics since pixel accuracy for the background may be unchanged while disease metrics may change significantly. The common class and disease only breakdowns have the same benchmark level interpretation.

5.3. Training side related cases exposure

In the case of the T1orT2 experiment, we get additional information regarding the mechanism that goes along with the previous experiment. The experiment compares the effect on independent clean cases versus related contaminated cases when training sample size and disease-class exclusion are held constant. There was no significant clean effect in both seeds; however, there were significant replicated effects on related contaminated cases when duplicate family exposure was maintained. It indicates a specific family exposure effect and not a general improvement in representation of the images. The experiment reduces the interpretation of released-versus-clean effect but it cannot prove the memorization or causation due to randomization. The residual morphological or acquisition difference might persist in the case of matched T1 and T2 controls.

5.4. Long-tail costs of naïve cleaning

Although cleaning increased independence, it decreased the level of support of rare classes. Median support levels decreased and the number of classes that had a low number of test images increased. Since support is positively correlated with IoU, it follows that the removal process results in reduced leakage bias and increased uncertainty about tail-class support. Group disjoint reconstruction provides a better solution to this problem. Both group disjoint reconstructions have all the classes present, preserve similar test sets in terms of size to the one that is publicly released and remove verified family crossings. Therefore, it makes sense to define families first and then allocate partitions as opposed to performing deletion post hoc. It should be noted that the released partitioning scheme can still be used for comparison purposes.

5.5. Implications for benchmark design and reporting

Three approaches flow from these findings. First, the process of constructing a benchmark should recognize the existence of exact and transformed duplicates, same-scene relations, and semantic contradictions prior to splitting. Second, visual family determinations that are ambiguous should use both automated searching and manual verification and not rely on hash comparisons alone. Third, studies that have been published should note the precise version of the dataset used as well as the definition of the splits and integrity-related filtering that is applied such that it can be revisited if new relations are discovered. In the case of existing benchmarks, contamination-insensitive and contamination-aware results can be reported separately. The former ensures continuity while the latter is a more defensible estimation of generalization to unrelated visual families.

5.6. Scope and limitations

There are five primary limitations to the study. Firstly, the results pertain to one benchmark, and generalization of the extent of contamination is not possible without audits on other datasets. Secondly, the exposure experiment was performed on two DeepLabV3+ training seeds, while other architectures were tested using only one seed. Thirdly, the retrieval audit was deliberately conservative and therefore did not give a global recall estimate; as a result, other related families might have been left undetected. Fourthly, the primary G42 repaired-split validation was performed using one model and one seed, hence serving to prove the feasibility of the protocol and not its superiority. Fifthly, the T1orT2 experiment involved one frozen matched T2 removal set; disease class numbers were exact, but secondary morphology and acquisition covariates were roughly balanced.

6. Conclusion

The PlantSeg dataset includes confirmed cross-split visual families which significantly influence the interpretation of the semantic-segmentation performance evaluation. In this audit, the contamination effect has been identified through setting up a contamination lower bound and it is shown that the released-split evaluation consistently favors the contamination-aware evaluation in multiple models. Also, a systematic experiment in the controlled training environment has pinpointed the source of the greatest benefit to the related contaminated images rather than the independent clean images. While direct deletion helps improve independence, it compromises the rare disease classification. Group-disjoint reconstruction helps achieve both with keeping 115 disease classes intact.

Acknowledgments: Not applicable

Author contributions: **Author 1:** Conceptualization, methodology, data curation, formal analysis, investigation, software implementation, experimentation, visualization, and original draft preparation. **Author 2:** Conceptualization, methodology, supervision, validation, interpretation of results, critical review and editing of the manuscript, and project administration. **Author 3:** Data validation, literature review, dataset preparation, analysis of results, visualization support, manuscript review and editing, and verification of the final manuscript.

Clinical trial number: Not applicable

Consent to participate: Not applicable

Consent to publish: Not applicable

Ethical approval: Not applicable

Data availability statements: The PlantSeg dataset analyzed in this study is publicly available through Zenodo at <https://doi.org/10.5281/zenodo.17719108>. The analysis was conducted on the PlantSeg v7 release comprising 7,774 image–mask pairs across 115 plant-disease categories.

Declaration of Competing Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AI Usage Declaration: During manuscript preparation, the authors used Claude for manuscript organization, language editing, consistency checks, and limited assistance with code troubleshooting and documentation. Claude was not used to generate experimental data or determine statistical results. All analyses, interpretations, and conclusions were independently reviewed by the authors, who take full responsibility for the work reported here.

References

- [1] Wei T, Chen Z, Yu X, Chapman S, Melloy P, Huang Z. A Large-Scale In-the-wild Dataset for Plant Disease Segmentation. *Scientific Data*. 2026;13:205. doi:10.1038/s41597-025-06513-4.
- [2] Feng C, Fan Y, Chen X, Wang H. PlantFormer: a precise plant disease segmentation network with interactive backbone and global-anisotropic context aggregation. *Frontiers in Plant Science*. 2026;17:1853571. doi:10.3389/fpls.2026.1853571.
- [3] Fan Y, Yu M, Shen L, Ma J, Zeng Z, Wang H. Robust plant disease segmentation in complex field environments: an in-depth analysis and validation with STAR-Net. *Frontiers in Plant Science*. 2026;16:1706072. doi:10.3389/fpls.2025.1706072.
- [4] Liu F, Hu Y, Zhou Q, Zou H. ReLeaf-SAM: reliability-guided detail compensation for SAM-based plant disease segmentation. *Frontiers in Plant Science*. 2026;17:1865649. doi:10.3389/fpls.2026.1865649.
- [5] Yang Z, Sun L, Deng J, Cai A, Liu Z, Liu G, et al. A multi-scale linear cross-modal fusion architecture for tomato leaf disease segmentation. *Computers and Electronics in Agriculture*. 2026;240:111155. doi:10.1016/j.compag.2025.111155.
- [6] Sharma N, Gupta S, Al-Yarimi FAM, Ghadi YY, Bharany S, Rehman AU, et al. DBA-DeepLab: Dual-Backbone Attention-Enhanced DeepLab V3+ Model for Plant Disease Segmentation. *Food Science & Nutrition*. 2025;13(7):e70668. doi:10.1002/orsn3.70668.
- [7] Yang Z, Sun L, Liu Z, Deng J, Zhang L, Huang H, et al. FATDNet: A fusion adversarial network for tomato leaf disease segmentation under complex backgrounds. *Computers and Electronics in Agriculture*. 2025;234:110270. doi:10.1016/j.compag.2025.110270.
- [8] Li J, Feng Q, Zhang J, Yang S. EMSAM: enhanced multi-scale segment anything model for leaf disease segmentation. *Frontiers in Plant Science*. 2025;16:1564079. doi:10.3389/fpls.2025.1564079.
- [9] Tian L, Yuan Y, En Q, Ma W, Zhang G, Liang F. Attention-based unsupervised prompt learning for SAM in leaf disease segmentation. *Knowledge-Based Systems*. 2025;320:113652. doi:10.1016/j.knsys.2025.113652.
- [10] Ge Z, Fan X, Zhang J, Jin S. SegPPD-FS: Segmenting plant pests and diseases in the wild using few-shot learning. *Plant Phenomics*. 2025;7(4):100121. doi:10.1016/j.plaph.2025.100121.
- [11] Zhao C, Li C, Wang X, Wu X, Du Y, Chai H, et al. Plant Disease Segmentation Networks for Fast Automatic Severity Estimation Under Natural Field Scenarios. *Agriculture*. 2025;15(6):583. doi:10.3390/agriculture15060583.
- [12] Xu Y, Ma B, Yu G, Zhang R, Tan H, Dong F, Bian H. Accurate cotton verticillium wilt segmentation in field background based on the two-stage lightweight DeepLabV3+ model. *Computers and Electronics in Agriculture*. 2025;229:109814. doi:10.1016/j.compag.2024.109814.
- [13] Zhang X, Feng Z, Mu W. Reformer: Re-parameterized kernel lightweight transformer for grape disease segmentation. *Expert Systems with Applications*. 2025;263:125757. doi:10.1016/j.eswa.2024.125757.
- [14] Williams D, Macfarlane F, Britten A. Leaf only SAM: A segment anything pipeline for zero-shot automated leaf segmentation. *Smart Agricultural Technology*. 2024;8:100515. doi:10.1016/j.atech.2024.100515.
- [15] Mouppou E, Retraint F, Tapamo H, Nkenliffack M, Kacfab C, Tagne A. Segment Anything Model and Fully Convolutional Data Description for Plant Multi-Disease Detection on Field Images. *IEEE Access*. 2024;12:102592-102605. doi:10.1109/ACCESS.2024.3433495.
- [16] Zhang S, Zhang C. Modified U-Net for plant diseased leaf image segmentation. *Computers and Electronics in Agriculture*. 2023;204:107511. doi:10.1016/j.compag.2022.107511.
- [17] Abinaya S, Kandagatla UK, Alphonse AS. Cascading Autoencoder With Attention Residual U-Net for Multi-Class Plant Leaf Disease Segmentation and Classification. *IEEE Access*. 2023;11:98153-98170. doi:10.1109/ACCESS.2023.3312718.
- [18] Wang H, Nguyen T-H, Nguyen TN, Dang M. PD-TR: End-to-end plant diseases detection using a transformer. *Computers and Electronics in Agriculture*. 2024;224:109123. doi:10.1016/j.compag.2024.109123.
- [19] Xu X, Wang Y, Shang Y, Hua Z, Wang Z, Song H. Plant leaf disease identification by parameter-efficient transformer with adapter. *Engineering Applications of Artificial Intelligence*. 2024;138(B):109466. doi:10.1016/j.engappai.2024.109466.
- [20] Naveenkumar M, Nandagopal S. Adaptive hybrid segmentation combined with meta heuristic optimization in transfer learning for plant leaf disease classification. *Scientific Reports*. 2025;15:9838. doi:10.1038/s41598-025-93225-9.
- [21] Faye D, Diop I, Mbaye N, Dione D, Diedhiou MM, et al. Mango fruit diseases severity estimation based on image segmentation and deep learning. *Discover Applied Sciences*. 2025;7:143. doi:10.1007/s42452-025-06550-z.
- [22] Mythili S, Kowsalya S, Kousalya A, Naveenkumar M. Plant Leaf Disease Segmentation and Classification Using Channel and Spatial Attention Block With UNet Model. *Journal of Phytopathology*. 2026;174(2):e70299. doi:10.1111/orjph.70299.

- [23] Zhang H-W, Wang R-F, Wang Z, Su W-H. DLCPD-25: A Large-Scale and Diverse Dataset for Crop Disease and Pest Recognition. *Sensors*. 2025;25(22):7098. doi:10.3390/s25227098.
- [24] Du M, Wang F, Wang Y, Li K, Hou W, Liu L, et al. Improving long-tailed pest classification using diffusion model-based data augmentation. *Computers and Electronics in Agriculture*. 2025;234:110244. doi:10.1016/j.compag.2025.110244.
- [25] Upadhyay A, Chandel NS, Singh KP, Chakraborty SK, et al. Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture. *Artificial Intelligence Review*. 2025;58:92. doi:10.1007/s10462-024-11100-x.
- [26] Shoaib M, Sadeghi-Niaraki A, Ali F, Hussain I, Khalid S. Leveraging deep learning for plant disease and pest detection: a comprehensive review and future directions. *Frontiers in Plant Science*. 2025;16:1538163. doi:10.3389/fpls.2025.1538163.
- [27] Kaya Y, Gursoy E. A review of deep learning architectures for plant disease detection. *Turkish Journal of Biology*. 2025;49(5):459-497. doi:10.55730/or1300-0152.2761.
- [28] Keshavarzi M, Mesarich C, Bailey D, Johnson M, Sen Gupta G. A review of semantic segmentation methods and their application in apple disease detection. *Computers and Electronics in Agriculture*. 2025;237(A):110531. doi:10.1016/j.compag.2025.110531.
- [29] Timilsina S, Sharma S, Kondo S. Advancements in maize leaf disease detection, segmentation and classification: A review. *Biosystems Engineering*. 2025;255:104162. doi:10.1016/j.biosystemseng.2025.104162.
- [30] Shafay M, Hassan T, Owais M, Hussain I, Khawaja SG, Seneviratne L, et al. Recent advances in plant disease detection: challenges and opportunities. *Plant Methods*. 2025;21:140. doi:10.1186/s13007-025-01450-0.
- [31] Khan SU, Alsuhaibani A, Alabduljabbar A, Almarshad F, Altherwy YN, et al. A review on automated plant disease detection: motivation, limitations, challenges, and recent advancements for future research. *Journal of King Saud University - Computer and Information Sciences*. 2025;37:34. doi:10.1007/s4443-025-00040-3.
- [32] Ngugi HN, Ezugwu AE, Akinyelu AA, Abualigah L. Revolutionizing crop disease detection with computational deep learning: a comprehensive review. *Environmental Monitoring and Assessment*. 2024;196:302. doi:10.1007/s10661-024-12454-z.
- [33] Qadri SAA, Huang N-F, Wani TM, Bhat SA. Advances and Challenges in Computer Vision for Image-Based Plant Disease Detection: A Comprehensive Survey of Machine and Deep Learning Approaches. *IEEE Transactions on Automation Science and Engineering*. 2025;22:2639-2670. doi:10.1109/TAASE.2024.3382731.
- [34] Thisanke H, Deshan C, Chamith K, Seneviratne S, Vidanaarachchi R, Herath D. Semantic segmentation using Vision Transformers: A survey. *Engineering Applications of Artificial Intelligence*. 2023;126(A):106669. doi:10.1016/j.engappai.2023.106669.
- [35] Rafi TH, Mahjabin R, Ghosh E, Ko Y-W, Lee J-G, et al. Domain generalization for semantic segmentation: a survey. *Artificial Intelligence Review*. 2024;57:247. doi:10.1007/s10462-024-10817-z.
- [36] Ji W, Li J, Bi Q, Liu T, Li W, Cheng L. Segment Anything Is Not Always Perfect: An Investigation of SAM on Different Real-world Applications. *Machine Intelligence Research*. 2024;21:617-630. doi:10.1007/s11633-023-1385-0.
- [37] Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. *Nature Communications*. 2024;15:654. doi:10.1038/s41467-024-44824-z.
- [38] Groger F, Lionetti S, Gottfrois P, Gonzalez-Jimenez A, Amruthalingam L, Labelling Consortium, et al. Intrinsic Self-Supervision for Data Quality Audits. *Advances in Neural Information Processing Systems* 37. 2024:92273-92316. doi:10.52202/or079017-2930.
- [39] Ramos P, Ramos R, Garcia N. Data Leakage in Visual Datasets. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2025:6368-6378. doi:10.1109/ICCVW69036.2025.00661.
- [40] Adimoolam YK, Poullis C, Averkiou M. Data Leakage Detection and De-duplication in Large Scale Geospatial Image Datasets. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2026. arXiv:2304.02296.
- [41] Lilja A, Fu J, Stenborg E, Hammarstrand L. Localization Is All You Evaluate: Data Leakage in Online Mapping Datasets and How to Fix It. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024:22150-22159. doi:10.1109/CVPR52733.2024.02091.
- [42] Barz B, Denzler J. Do We Train on Test Data? Purging CIFAR of Near-Duplicates. *Journal of Imaging*. 2020;6(6):41. doi:10.3390/jimaging6060041.
- [43] Bradshaw TJ, Huemann Z, Hu J, Rahmim A. A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging. *Radiology: Artificial Intelligence*. 2023;5(4):e220232. doi:10.1148/oryai.220232.
- [44] Rouzrokh P, Khosravi B, Faghani S, Moassefi M, Vera Garcia DV, Singh Y, et al. Mitigating Bias in Radiology Machine Learning: 1. Data Handling. *Radiology: Artificial Intelligence*. 2022;4(5):e210290. doi:10.1148/oryai.210290.
- [45] Li S, Yang L, Cao P, Li L, Ma H. Frequency-Based Matcher for Long-Tailed Semantic Segmentation. *IEEE Transactions on Multimedia*. 2024;26:10395-10405. doi:10.1109/TMM.2024.3407679.
- [46] Han B, Xu Q, Yang Z, Bao S, Wen P, Jiang Y, Huang Q. AUCSeg: AUC-Oriented Pixel-Level Long-Tail Semantic Segmentation. *Advances in Neural Information Processing Systems* 37. 2024:126863-126907. doi:10.52202/or079017-4029.
- [47] Zheng Q, Zhao W, Wu C, Zhang X, Dai L, Guan H, et al. Large-scale long-tailed disease diagnosis on radiology images. *Nature Communications*. 2024;15:10147. doi:10.1038/s41467-024-54424-6.
- [48] Holste G, Zhou Y, Wang S, Jaiswal A, Lin M, Zhuge S, et al. Towards long-tailed, multi-label disease classification from chest X-ray: Overview of the CXR-LT challenge. *Medical Image Analysis*. 2024;97:103224. doi:10.1016/j.media.2024.103224.

Supplementary Files

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

- [PlantSegSupplementary.docx](#)