

Foreground Sampling and Spatial Support in Pavement Crack Segmentation

Ehsan Ghaffari

ehsanghaffari@montana.edu

Montana State University

Kelvin C.P. Wang

Montana State University

Research Article

Keywords: Pavement crack segmentation, foreground sampling, patch sampling, class imbalance, spatial support, Crack500, deep learning

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11088886/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

Foreground-focused patch sampling is commonly used to address class imbalance in crack segmentation, but conditioning crops on foreground presence also changes which background regions can enter training. This study separates foreground prevalence from spatial support using a controlled experiment on the original full-resolution Crack500 scenes. Across 36 trained models, a fixed U-Net and training recipe were retained while sampling and normalization controls were varied with three paired initialization seeds. The prespecified screening stage identified a heterogeneous deficit under aggressive positive-only cropping, but the more direct mechanism evidence came from matched support controls. Exact enumeration of all valid 256 x 256 crop positions showed that positive-only sampling made 47.0% of training-image area, representing 39.1% of random-crop exposure, unreachable. At the primary two-pixel-tolerance average precision (AP2) endpoint, a natural-prevalence control retaining this restricted support remained 0.02659 below random cropping under batch normalization (BatchNorm), whereas a second control that matched its eligible negative-support cardinality but restored 99.98% spatial reachability returned to baseline (difference - 0.00024; 95% confidence interval [-0.00500, 0.00387]) and exceeded the restricted-support control by 0.02636. The same pattern did not replicate at a material magnitude under group normalization (GroupNorm), although aggressive positive-only sampling remained harmful. Foreground sampling should therefore be treated and reported as a spatial data-selection rule as well as a class-imbalance intervention, with crop size, natural positive-crop prevalence, spatial reachability, and eligible-support cardinality documented separately; its performance consequences can interact with normalization.

1. Introduction

Pavement cracking is among the most common and informative forms of surface deterioration used in roadway condition assessment, and reliable identification of crack location and extent underpins subsequent estimation of geometry, severity, and network condition. Conventional visual inspection is labor intensive, subjective, and difficult to apply consistently at network scale, which has driven extensive research on automated image-based assessment. Recent reviews trace a progression toward deep-learning pipelines while continuing to identify robustness, labeling, evaluation, and deployment as unresolved challenges (Mulyanto et al., 2025; Abdollahi and Okte, 2026).

Within this progression, pixel-level semantic segmentation has become especially important because it delineates irregular crack structures rather than merely classifying an image or localizing a bounding box. U-Net and its variants remain influential because the encoder-decoder structure combines contextual feature extraction with spatially precise decoding (Ronneberger et al., 2015), and pavement-specific work has expanded the design space through feature pyramids, Transformer encoders, and background-morphology modeling (Yang et al., 2020; Lau et al., 2020; Han et al., 2022; Liu et al., 2023; Yu et al., 2024; X. Guo et al., 2025; K. Wu et al., 2025). Architectural progress does not remove a more basic property: crack pixels are rare relative to intact pavement, and existing studies address this imbalance through loss design, augmentation, foreground-focused preprocessing, and preferential sampling (K.

Zhang et al., 2021; Tang et al., 2021; Y. Zhang et al., 2022; Al-Huda et al., 2023; Siriborvornratanakul, 2023).

Patch-based training makes the problem more subtle, because a foreground-enrichment rule changes more than the fraction of positive examples: the rule used to accept regions also determines which parts of the original scene can enter optimization. In the original Crack500 processing used by Yang et al. (2020), each image was divided into 16 non-overlapping regions and regions with fewer than 1,000 crack pixels were discarded, which increased crack concentration but also constructed a foreground-selected spatial distribution. More generally, if a crop is accepted only when it intersects a crack, regions sufficiently far from every crack become impossible to sample, so the intervention changes both class composition and spatial support.

This distinction matters because background information is not interchangeable. Non-crack pavement contains lane markings, joints, manholes, repaired boundaries, shadows, and texture variation, and field evaluation of a publicly trained model attributed its dominant error modes to exactly these crack-like non-crack structures (Lim et al., 2026). Chen et al. (2024) showed that marginal non-crack regions near crack boundaries carry informative structure, several recent models devote explicit capacity to background morphology or hard-negative suppression (X. Guo et al., 2025; K. Wu et al., 2025; Mao et al., 2026), and Nguyen and Thai (2023) warned that data-level imbalance remedies can remove precisely this contextual information.

The resulting methodological gap is that foreground-oriented sampling is usually described in terms of class balance or positive-patch frequency, whereas its spatial-selection consequences are rarely measured. Li et al. (2024), for example, proposed adaptive probabilistic sampling and reported improvements over random cropping on Crack500 and GAPs384. That establishes that sampling policy can materially affect performance, but it does not separate the effect of changing foreground prevalence from the effect of changing which background locations are reachable: two schemes can produce similar positive-crop prevalence while exposing the model to different backgrounds, and two schemes can retain equally many eligible negative positions while distributing them very differently.

The present study fixes a U-Net and training recipe and treats crop sampling as the experimental intervention on audited full-resolution Crack500 scenes. Exhaustive enumeration of all valid 256 x 256 crops quantifies natural positive-crop prevalence, spatial reachability, and eligible-negative cardinality. Four screening conditions span low, random, moderate, and positive-only sampling; matched E/F controls then separate restricted support from reduced cardinality, while batch normalization (BatchNorm) and group normalization (GroupNorm) controls assess normalization dependence, with average precision (AP2) as the prespecified primary endpoint. The study therefore contributes an experimental separation of these three properties and a reporting framework that treats patch sampling as a spatial data-selection rule.

The broader implication is that a statement such as “training used 58% positive patches” does not fully describe the resulting training distribution unless the associated crop geometry and background

coverage are also known. This paper therefore asks not only how much foreground a model should see, but also what source-image information is removed when that foreground exposure is enforced.

2. Related Work

2.1 Segmentation Architectures

Pavement crack segmentation has progressed from U-Net-based convolutional networks toward feature-pyramid, attention, Transformer, and morphology-aware architectures (Lau et al., 2020; Yang et al., 2020; Han et al., 2022; Liu et al., 2023; Yu et al., 2024; X. Guo et al., 2025; K. Wu et al., 2025), while comparative studies continue to identify generalization, dataset construction, and evaluation as major challenges (Y. Zhang and L. Zhang, 2024; Mulyanto et al., 2025; Abdollahi and Okte, 2026). Because the present question concerns sampling rather than architecture, all experiments use one fixed U-Net.

2.2 Imbalance and Sampling

Class imbalance has been addressed at both the algorithm and data levels. Algorithm-level approaches alter losses, weighting, or network behavior so that minority crack pixels contribute more effectively: encoder-decoder designs targeting crack/non-crack quantity imbalance (Tang et al., 2021), recurrent adaptive balanced learning (Y. Zhang et al., 2022), an asymmetric dual-decoder U-Net (Al-Huda et al., 2023), and methods for very thin cracks under severe imbalance (Siriborvornratanakul, 2023).

Data-level interventions instead alter which training samples are presented. CrackGAN used crack-patch-only supervision to prevent an all-background failure mode under partially accurate ground truth (K. Zhang et al., 2021). The widely reused Crack500 preprocessing from Yang et al. (2020) retained only non-overlapping regions containing more than 1,000 crack pixels, deriving the commonly used 1,896, 348, and 1,124 crop sets from the original release. Reményi et al. (2026) likewise used crack-affected cropped regions to reduce extreme pixel imbalance, and Tao (2024) used cropping and pasting within a local-to-global scheme. Each is reasonable within its pipeline, yet each illustrates that an imbalance intervention can simultaneously become a spatial-selection rule.

Evidence that data composition can dominate other design choices is accumulating independently of architecture. Fan et al. (2025) compared preprocessing variants and dataset balance across four crack datasets and concluded that balance influenced accuracy more strongly than the preprocessing method, identifying resampling as the direction for further work, and Yin et al. (2025) treated sample selection as a design variable in road-damage detection. Li et al. (2024) makes the issue most explicit, reporting advantages for adaptive probabilistic sampling over random cropping. This raises a specific identification question: when sampling changes performance, is the relevant intervention the increased frequency of crack-positive crops, the change in the distribution of background locations, or both?

2.3 Background Context

The assumption that most non-crack pixels are interchangeable is difficult to reconcile with recent research. Nguyen and Thai (2023) warned that deleting non-crack patches or undersampling the majority class can reduce global contextual information in complex scenes. Patch filtering does not merely remove redundant easy negatives; depending on geometry, it can remove entire source-image regions and the structures they contain.

Independent field evidence supports the concern. Lim et al. (2026) evaluated a model trained on integrated public crack datasets against unmanned aerial vehicle (UAV)-acquired imagery and attributed the major error modes to crack-like non-crack factors including lane markings, manholes, repaired boundaries, and shadows rather than to the cracks themselves. Chen et al. (2024) developed a two-stage representation-learning framework because ambiguous marginal non-crack regions improve discrimination, MorFormer reconstructs background morphology to suppress crack-like noise (X. Guo et al., 2025), and Zhong et al. (2024) synthesized crack images with varied backgrounds. The background therefore carries structured information influencing false positives and generalization.

These observations motivate a distinction between negative-support cardinality and negative-support identity. A sampler can retain many negative coordinates drawn from a narrow neighborhood, or the same number distributed broadly across the scene. The two distributions have equal eligible-negative counts but different environmental support, and a positive-patch percentage cannot reveal that difference.

2.4 Evaluation Endpoints

The metric used to score a sampling intervention is not neutral with respect to it. Changing the positive-crop rate changes the effective class prior seen during training and can shift the predicted probability scale, so a threshold-dependent metric can register part of that shift as a change in segmentation quality. Average precision evaluates ranking across operating thresholds rather than at one fixed threshold, so the study uses tolerance-aware pixel average precision as the primary endpoint and reserves fixed-threshold metrics for calibration diagnostics.

Matching tolerance is the second relevant axis. Rill-García et al. (2022) observed that crack results are conventionally reported with a tolerance margin of a few pixels, acceptable for locating cracks but not for measuring width, and showed that inaccurate annotations can depress tolerance-free scores by as much as 20%. Silva et al. (2026) reported materially different conclusions under exact-match versus dilated-tolerance criteria, and Ma et al. (2027) note that comprehensive comparative evaluation remains limited. Reporting across a graded range of tolerances therefore carries mechanistic information: an effect that persists as tolerance relaxes is unlikely to be a pure localization artifact, whereas one that attenuates is.

2.5 Sampling Beyond Pavement

The distinction developed here is not unique to pavement imagery. In object detection, online hard example mining selects training examples by difficulty rather than uniformly (Shrivastava et al., 2016)

and focal loss reweights easy negatives rather than discarding them (Lin et al., 2017). In patch-based medical image segmentation, nnU-Net implements foreground oversampling so that rare structures appear in sampled patches (Isensee et al., 2021), and systematic studies have examined oversampling and undersampling as first-class experimental variables (Buda et al., 2018).

Two things follow. First, the practice interrogated here is widespread and in some fields codified in training pipelines, which raises rather than lowers the value of measuring its side effects. Second, what distinguishes the present work is the geometric consequence specific to cropping from large scenes: when patches are accepted by a foreground criterion, a measurable portion of the source imagery becomes unreachable. Hard example mining and foreground oversampling change which examples are emphasized among those available; foreground-conditioned cropping can additionally change which source locations can enter training at all.

2.6 Research Gap

Across these lines of work, the unresolved issue is not whether sampling composition matters, but what changes when foreground-oriented cropping is imposed. Prior studies report crack-focused filtering, crack-patch-only supervision, adaptive sampling, and explicit background modeling, but do not jointly separate foreground prevalence from the spatial support and eligible-negative cardinality induced by the crop rule. Table 1 compares six especially relevant studies, which are revisited in Section 5.2.

Table 1

Representative prior work on crack sampling, imbalance, and background information, and its relation to the present study.

Prior study	Relevant prior practice or conclusion	Relation to the present study
Yang et al. (2020), Feature Pyramid and Hierarchical Boosting Network	For Crack500, divided each full image into 16 non-overlapping regions and retained only regions containing more than 1,000 crack pixels.	The reported benchmark and model results remain valid for the derived crop dataset. The present study treats this preprocessing additionally as a spatial-selection rule and measures the training-image support it removes.
K. Zhang et al. (2021), CrackGAN	Used crack-patch-only supervision to prevent the all-background failure mode under partially accurate labels and severe imbalance.	The benefit is not contradicted. The present study examines the complementary cost that can arise in fully supervised patch segmentation when conditioning on crack-containing crops removes background support.
Nguyen and Thai (2023), Crack segmentation of imbalanced data	Warned that deleting non-crack patches or undersampling the majority class can remove global contextual information; emphasized loss-based alternatives across varying imbalance severity.	The present study operationalizes that warning by quantifying reachable support and testing whether restricted support can reproduce a sampling deficit independently of positive-crop prevalence.
Li et al. (2024), Multi-level attention DeepLabV3+	Proposed adaptive probabilistic sampling and reported advantages over random cropping in tested Crack500 and GAPs384 configurations.	Their empirical gains are not a direct target of replication. The present study tests whether foreground enrichment is uniformly beneficial and separates enrichment from the spatial-support change induced by the sampler.
Chen et al. (2024), Mind Marginal Non-Crack Regions	Showed that ambiguous non-crack regions near crack boundaries contain useful information and designed representation learning specifically for those regions.	The findings are complementary: Chen et al. establish the value of difficult local background, whereas the present study examines whether a sampler can make broader training-image background unreachable before learning begins.
Reményi et al. (2026), HoughNet-extended U-Net	Used crack-affected cropped Crack500 regions to reduce extreme pixel-level imbalance before training and optimized a recall-oriented segmentation pipeline.	The within-pipeline comparisons remain valid. The present study highlights that crack-focused cropping also changes spatial support and that support cardinality and spatial coverage should be reported separately.

3. Materials and Methods

3.1 Dataset Integrity

The experiments used the 500 original full-resolution Crack500 scenes rather than the redistributed pre-cropped release, because that release had already applied crack-content filtering and could not support unbiased enumeration of the crop domain. Orientation was audited before training because JPEG images can carry Exchangeable Image File Format (EXIF) orientation metadata that rotates the image while leaving the paired PNG mask in its stored orientation. A decoder honoring EXIF made 247 of the 500 pairs appear dimensionally mismatched, so a naive filter would have rejected all 247; decoding in raw stored orientation prevented 244 false flags, and three genuine transposed training pairs were excluded. The retained split was 247 training, 50 validation, and 200 test pairs.

Shape agreement was not treated as proof of image-mask alignment. Seventy-six retained JPEGs carried shape-preserving EXIF orientation metadata and were manually confirmed under raw/identity decoding; no mask was automatically transformed. Preparation-gate and fingerprint procedures are documented in Supplementary Methods S6.

Training was patch based using 256 x 256 pixel crops. A crop was labeled crack-positive when its reference mask contained at least one crack pixel. Every valid top-left coordinate in the retained training set was enumerated before model fitting together with its status, providing the natural positive-crop prevalence and the common sampling frame for all conditional samplers. The audited test set contained no crack-free scenes, so pixel-level average precision (AP) still penalizes false positives on non-crack pixels within each scene, but the study does not evaluate image-level behavior on crack-free scenes or deployment prevalence.

3.2 Sampling Distributions

Let Ω denote the set of all valid 256 x 256 crop positions in the retained training set, Ω_+ the subset whose mask contains at least one crack pixel, and $\Omega_- = \Omega \setminus \Omega_+$ the crack-negative subset, with $N = |\Omega|$ and $N_{\text{pos}} = |\Omega_+|$. The base distribution was uniform over Ω . Exhaustive enumeration yielded $N = 686,632,695$ valid positions, of which $N_{\text{pos}} = 254,540,526$ were crack-positive, giving the exact natural positive-crop prevalence p_{pos} as

$$p_{\text{pos}} = \frac{N_{\text{pos}}}{N} = 0.370708$$

1

Forced-positive and forced-negative sampling were implemented by rejection sampling from this base distribution: each attempt sampled uniformly from Ω and was accepted only when the requested class was obtained, so accepted forced-positive positions were uniform over Ω_+ and accepted forced-negative positions uniform over Ω_- . This avoids image-level double weighting and makes the intervention a controlled change in crop composition relative to a common base. Table 2 summarizes the six conditions.

Table 2

Sampling conditions and reachable training-image support at the primary 256-pixel crop size. A-D are the primary screening conditions; E_spatial_support and F_random_support_matched are targeted mechanism controls (Sections 3.8–3.9).

Condition	Sampling rule	Expected positive fraction	Training-image area reachable	Role
A_low	Forced-negative + free random	0.100	100%	Low foreground exposure
B_random	Uniform random over Ω	0.371	100%	Natural baseline
C_moderate	0.33 forced-positive + 0.67 random	0.578	100%	Moderate enrichment
D_aggressive	Forced-positive only	1.000	53%	Aggressive conditioning
E_spatial_support	Positive from Ω + at p_{pos} ; restricted negatives at $1 - p_{\text{pos}}$	0.371	53%	Natural prevalence, restricted support
F_random_support_matched	Positive from Ω + at p_{pos} ; negatives from exact-size random Ω -subset	0.371	99.98%	Cardinality-matched spatial-release control

These conditions should not be interpreted as varying foreground prevalence alone. Conditioning on crack-positive crops also constrains the source-image pixels reachable by sampled footprints, so D_aggressive differs from B_random in both prevalence and spatial support; this coupling is quantified in Section 3.7. Condition E restores natural prevalence while retaining D-like restriction (Section 3.8), and Condition F holds E's eligible negative-support cardinality fixed while restoring near-complete reachability (Section 3.9).

For C_moderate, the expected positive fraction was

$$\Pr(+ \mid C) = 0.33 + 0.67p_{\text{pos}} = 0.5784$$

2

For A_low, the free-sampling probability was set to $0.10/p_{\text{pos}} = 0.26975$, with the remaining 0.73025 of draws forced negative. This yielded an expected positive-crop fraction of 0.10, compared with the natural prevalence $p_{\text{pos}} = 0.3707$ under B_random. The four conditions therefore span insufficient foreground exposure, unmanipulated natural distribution, moderate enrichment, and complete conditioning on crack-containing crops.

3.3 Network and Training

A fixed two-dimensional U-Net isolated sampling effects from architecture selection. The network had base width 32, depth 4, 7,763,041 trainable parameters, and 18 BatchNorm2d layers (Ronneberger et al., 2015; Ioffe and Szegedy, 2015), and conditions were trained with paired seeds 42, 1337, and 2026 using identical initialization within each seed. Because one architecture and one dataset were used, geometry-dependent quantities such as unreachable training-image area are specific to this configuration.

All primary models were trained for 20,000 updates with batch size 16. The loss was equally weighted binary cross-entropy (BCE) plus soft Dice: BCE used BCEWithLogitsLoss, while Dice was computed per image after sigmoid activation as $1 - (2 \sum p_i y_i + 1) / (\sum p_i + \sum y_i + 1)$ and then averaged across the batch. AdamW used an initial learning rate of 3×10^{-4} , weight decay 1×10^{-4} , 500-update linear warmup, and cosine decay (Milletari et al., 2016; Loshchilov and Hutter, 2017, 2019). Horizontal and vertical flips and random 90° rotations were applied identically across conditions. Automatic mixed precision was enabled on CUDA hardware.

The final checkpoint at 20,000 updates was used for testing; no early stopping or validation-based checkpoint selection was permitted. Training used four persistent DataLoader workers, treated as part of the fixed protocol because worker identity entered the sampler's deterministic random-number stream. Frozen random validation crops were monitored only for convergence and never influenced training duration, sampling, or checkpoint selection.

3.4 Inference and Evaluation

All trained models were evaluated on the same 200 held-out full-resolution test scenes. Prediction used 256 x 256 tiles at a 128-pixel stride with reflect padding only as needed. Overlapping tile predictions were multiplied by the same two-dimensional Hann window, accumulated, normalized pixelwise by the accumulated weights, and trimmed to native dimensions; the whole pipeline was frozen across conditions. The test set was not used for training, calibrator fitting, checkpoint selection, or threshold selection, the sole exception being the explicitly designated transductive diagnostic of Section 3.6.2, which used unlabeled test tiles only to update BatchNorm state. Probability maps were retained for the twelve native BatchNorm A-D runs and reused for later analyses without retraining; other runs retained evaluation summaries and checkpoints.

The primary endpoint was tolerance-aware pixel AP at a two-pixel Euclidean matching tolerance (AP2), with AP0, AP1, and AP3 forming a prespecified tolerance sweep. Let $s(x)$ be the predicted foreground score at pixel x , $P_\tau = \{x: s(x) \geq \tau\}$ the predicted-positive set at threshold τ , G the ground-truth foreground set, and $D_t(A)$ the dilation of A by the discrete Euclidean disk of integer offsets satisfying $\delta x^2 + \delta y^2 \leq t^2$. Tolerant precision and recall were defined as

$$\text{Precision}_t(\tau) = \frac{|P_\tau \cap D_t(G)|}{|P_\tau|}$$

3

$$\text{Recall}_t(\tau) = \frac{|G \cap D_t(P_\tau)|}{|G|}$$

4

A single grey-scale maximum filter with the Euclidean-disk footprint therefore produced the score at each ground-truth pixel, replacing a threshold-by-threshold dilation; the equivalence is derived in Supplementary Methods S10. Predicted probabilities were clipped to [0,1], quantized to unsigned 16-bit integers, and mapped linearly to 2,048 histogram bins. For each image and tolerance, histograms stored all prediction scores, scores inside $D_t(G)$, and neighborhood-maximum scores at ground-truth pixels. Because pooling preceded integration, the endpoint is pixel-weighted. Identical bin edges were used for every condition and seed, and histograms were pooled across the 200 scenes, accumulated from high to low score, and integrated as

$$AP_t = \sum_k \text{Precision}_{t,k} \Delta \text{Recall}_{t,k}$$

5

where k indexes descending histogram thresholds and ΔRecall is the nonnegative recall increment between adjacent bins. At $t = 0$ the disk contains only the center pixel and the implementation reduces to ordinary pooled pixel average precision, an identity used as an internal implementation check. AP_2 was designated as the primary endpoint before inspection of the primary experimental results.

The 2,048-bin mapping and bin edges were fixed across all conditions and seeds. For the twelve native BatchNorm A-D runs, a post hoc check rescored the retained 16-bit probability maps at 8,192 and 65,536 bins; reconstructed 2,048-bin values matched the archived results exactly, and because the stored maps carry 65,536 score levels the 65,536-bin analysis is the finest possible discretization. The D-B contrast was -0.02283 , -0.02222 , and -0.02287 across the three resolutions; across all six prespecified contrasts the maximum change was 0.00282 and none crossed the 0.020 floor. Full results are in Supplementary Section S8.

3.5 Statistical Analysis

Between-condition uncertainty was estimated with a paired test-image bootstrap using 10,000 resamples (Efron and Tibshirani, 1993). Within each draw the same image multiplicities were applied to both conditions and all three paired seeds, and seed-specific contrasts were averaged within the draw. Percentile 95% confidence intervals were taken from the paired contrast distribution. Two-sided p-values used the finite-resample correction $p = 2 \min\{(n_- + 1)/(B + 1), (n_+ + 1)/(B + 1)\}$ capped at 1, with $B = 10,000$; values at the resulting Monte Carlo floor are reported as $p < 0.001$. Multiplicity within the screening family was controlled with the Holm procedure (Holm, 1979).

Only AP2 determined the initial screening decision; AP0, AP1, and AP3 were sensitivity analyses that could not override it. GO required at least one predeclared contrast c to satisfy all three criteria

$$GO \iff \exists c : |\Delta AP2_c| \geq 0.020 \wedge 0 \notin CI_{95,c} \wedge p_{Holm,c} < 0.05$$

6

STOP required every predeclared contrast to fall below the 0.020 practical effect floor and every unadjusted paired-image 95% interval to include zero:

$$STOP \iff \forall c : |\Delta AP2_c| < 0.020 \wedge 0 \in CI_{95,c}$$

7

Any result satisfying neither rule was classified as INCONCLUSIVE, a prospectively defined outcome. The paired bootstrap intervals quantify test-image resampling uncertainty conditional on the three trained initializations and do not capture variability across the broader population of seeds, so seed-level consistency was interpreted descriptively; complementary post hoc two-sided 95% t intervals from the three paired seed differences are constrained by two degrees of freedom. One prospective amendment, replacing an original three-of-three paired-seed sign requirement with a descriptive consistency check, was made before any outcomes were inspected and is documented in Supplementary Methods S9.

The 0.020 threshold served different purposes in the two settings. The screening GO rule required the point estimate, with evidence against zero, to exceed the practical-effect floor, but not the entire interval to lie beyond it. A practical-compatibility claim instead required the entire interval to lie inside ± 0.020 , a deliberately stricter criterion. Individual AP2 values for every condition and seed are in Supplementary Table S1.

Multiplicity was handled by analysis family rather than study-wide. The six predeclared A-D contrasts at AP2 formed the primary screening family and were Holm-adjusted; the same contrasts at AP0, AP1, and AP3 formed sensitivity families, Holm-adjusted within each tolerance, that could not override the AP2 gate. The eight exploratory scene-feature associations were adjusted with the Benjamini-Hochberg false-discovery-rate procedure (Benjamini and Hochberg, 1995).

For native-BatchNorm Condition F, F-B and F-E were locked before any F outcome and interpreted against the ± 0.020 margin rather than by a nonsignificant p-value alone. A later GroupNorm stage locked E-B, F-B, and F-E as a new Holm-adjusted family, with GroupNorm D-B retained only as a reference. Familywise control therefore applies within the families listed in Table 3; no study-wide rate is claimed, and the full chronology of locks is in Supplementary Methods S9.

Table 3 summarizes the inferential families, multiplicity handling, and decision roles used throughout the staged design.

Table 3
Inferential families, multiplicity handling, and decision roles.

Analysis family	Contrasts/endpoints	Multiplicity handling	Decision role
Primary A-D screening	Six AP2 pairwise contrasts among A, B, C, and D	Holm within family	Primary GO/STOP/INCONCLUSIVE screening decision
A-D tolerance sensitivity	Same six contrasts at AP0, AP1, and AP3	Holm within each tolerance family	Sensitivity only; cannot override the AP2 decision
Original GroupNorm B/D interaction	Cross-normalizer D-B interaction, Eq. (8)	Single planned primary test	Tests whether replacing BatchNorm attenuates the D-B deficit
Exploratory scene heterogeneity	Eight scene-feature associations	Benjamini-Hochberg FDR	Exploratory description of heterogeneity
GroupNorm A-D extension	Six AP2 pairwise contrasts among A, B, C, and D under GroupNorm	Holm within family	Later protocol-locked response-control extension
Native BatchNorm mechanism controls	E-B planned primary contrast; F-B and F-E locked before F; E-D secondary	Single planned E-B test; separate locked two-contrast F family	Mechanism testing within the native BatchNorm setting
GroupNorm E/F replication	E-B, F-B, and F-E at AP2; D-B retained only as reference	Holm within the three-contrast family	Normalization replication of the E/F mechanism pattern
Secondary or post-hoc checks	Fixed-threshold calibration contrasts; scene-level influence audit; GroupNorm AP0-AP3 tolerance sweep; native BatchNorm A-D histogram-resolution sensitivity	Unadjusted or descriptive as stated	Secondary or sensitivity interpretation only

3.6 Mechanism Controls

After the primary sampling response was established, targeted analyses evaluated candidate explanations for the aggressive-sampling deficit: probability calibration, BatchNorm-dependent normalization state, localized scene/background difficulty, restriction of spatial support, and reduced eligible negative-support cardinality. Condition E tested whether restricted spatial support was sufficient to reproduce a deficit at natural prevalence, Condition F then separated spatial identity from cardinality under native BatchNorm, and a later GroupNorm E/F replication tested whether that pattern was normalization-invariant.

3.6.1 Calibration

Temperature scaling and positive-slope Platt scaling were fitted to validation full-scene predictions only (C. Guo et al., 2017; Platt, 1999). Both mappings were constrained to be strictly monotonic and therefore preserved score ordering, so the diagnostic retained the original ranking-based AP values rather than recomputing AP after requantizing transformed scores; calibration errors and fixed-threshold metrics were computed from the transformed probabilities. Full procedures are in Supplementary Methods S1.

3.6.2 Normalization Controls

A post-hoc BatchNorm source audit tested whether sampling-dependent running statistics could account for the aggressive-sampling deficit. Native checkpoint statistics remained primary. For sensitivity analysis, running means, variances, and counters were reset and recomputed by cumulative averaging from validation-only calibration crops while model weights and BatchNorm affine parameters remained fixed.

Four validation-only sources matched the A-D training prevalence targets (realized positive fractions 0.1001, 0.3709, 0.5784, and 1.0000). A transductive recalibration using unlabeled test tiles was diagnostic-only and excluded from mechanistic inference. Full recalibration checks and source construction are in Supplementary Methods S2.

A separate architectural control replaced all 18 BatchNorm2d layers with affine GroupNorm (eight groups) while preserving the 7,763,041 trainable-parameter count and the frozen training/evaluation protocol (Y. Wu and He, 2018). The original endpoint control covered B_random and D_aggressive and tested whether a BatchNorm-specific mechanism was necessary for the aggressive-sampling deficit. Because GroupNorm changes both running-state dependence and the normalization operator, it cannot isolate which normalization component is causal. The prespecified B/D interaction was

$$\Delta_{GN} = (AP2_{D,GN} - AP2_{B,GN}) - (AP2_{D,BN} - AP2_{B,BN})$$

8

where the subscripts BN and GN denote BatchNorm and GroupNorm and a positive interaction would indicate attenuation of the D-B deficit under GroupNorm. After the B/D result was known, a protocol-locked extension added GroupNorm A_low and C_moderate under the identical protocol. Before A/C outcomes were inspected, the extension locked the expectations that B_random would remain the peak, that B-A would remain practically large, that C-B would remain practically small and compatible with zero, and that A_low seed standard deviation (SD) would decrease; the complete curve was to be reported regardless of outcome.

3.6.3 Convergence and Protocol Audit

Training convergence and protocol integrity were audited from frozen-validation loss and recorded run metadata. These diagnostics were not used for checkpoint selection or to redefine AP2 decisions; full

procedures and results are provided in Supplementary Methods S3.

3.6.4 Scene-Level Diagnostics

Exploratory scene-level analyses related D_aggressive - B_random differences to background texture, crack-network structure, and near- versus far-background features. These analyses were descriptive rather than causal; feature definitions, multiplicity handling, and detailed results are provided in Supplementary Methods S4.

A separate post-hoc influence audit recomputed pooled AP2 after removing $k = 1, 2, 5, 10, 20$, and 40 scenes ranked by leave-one-out influence for D-B, E-B, and F-B, compared targeted with random removal, and contrasted pooled estimates with per-image medians and ground-truth-pixel-weighted means. The 52 IMG_* and 148 date-stamped filename/resolution families were compared with two-sided Mann-Whitney U tests. Because pooled AP is renormalized after removal, the influence-share statistic is distinct from direct deletion attenuation.

3.7 Spatial Exposure

Spatial support denotes source-image pixels reachable by eligible crops, whereas eligible negative support denotes available crack-negative crop coordinates; its size is the eligible-support cardinality. The analysis characterized training exposure at the pixel level rather than by the nominal number of distinct crop coordinates, because adjacent 256×256 crops overlap heavily and different top-left coordinates can contain nearly the same visual content.

For each training pixel p , let $n_{\text{all}}(p)$ denote the number of valid crop positions whose footprint contains p and $n_{\text{pos}}(p)$ the number of crack-positive crop positions whose footprint contains p . The local positive-support fraction was defined as

$$r(p) = \frac{n_{\text{pos}}(p)}{n_{\text{all}}(p)}$$

9

Under B_random, every valid crop position has probability $1/|Q|$. Under D_aggressive, every positive crop position has probability $1/|Q^+|$. The expected D-to-B exposure ratio for pixel p is therefore

$$q(p) = \frac{n_{\text{pos}}(p) / N_{\text{pos}}}{n_{\text{all}}(p) / N} = \frac{r(p)}{p_{\text{pos}}}$$

10

Thus $r(p) = 0$ identifies area that positive-conditioned cropping can never expose, $q(p) = 1$ identifies equal expected exposure under B and D, and $r(p) = 1$ yields the maximum ratio $q_{\text{max}} = 1/p_{\text{pos}}$. Positive-position bitmaps from data preparation gave $n_{\text{pos}}(p)$ by box summation over the crop footprint, and

$n_{\text{all}}(p)$ followed from valid-crop geometry. Distributions were summarized under two weightings: area weighting, which treats every pixel equally and quantifies the physical fraction of training imagery unreachable or underexposed, and B-random exposure weighting, which weights each pixel by $n_{\text{all}}(p)$. The B-exposure-weighted mean of $q(p)$ equals one by construction and served only as an internal identity check.

The prespecified spatial summaries included the fraction of training-image area with $r(p) = 0$, the fraction with $r(p) < 0.10$, the fraction receiving less than one quarter of B_random exposure, and distributional summaries of $r(p)$.

Expected distinct coordinate counts were retained only as a secondary preflight diagnostic, because coordinate cardinality does not measure independent content when crop overlap is high. The same exposure analysis was repeated at 128 x 128, 256 x 256, and 512 x 512 pixels without retraining; procedures and results are in Supplementary Methods S5 and Supplementary Results S7.

3.8 Spatial-Support Control

A targeted condition, E_spatial_support, tested whether restricted spatial support was sufficient to reproduce the aggressive-sampling deficit after restoring natural positive-crop prevalence. The E design, seeds, primary endpoint, and primary contrast were locked before the three E outcomes were inspected, and the model used the native BatchNorm U-Net with the same training and evaluation protocol as the primary experiment.

Let the D-reachable pixel support be the union of training pixels contained in at least one crack-positive crop. In E, positive crops were drawn uniformly from $\Omega +$ with probability $p_{\text{pos}} = 0.370708$; otherwise a crack-negative crop was drawn uniformly from the subset of $\Omega -$ whose entire footprint lay inside that support. The eligible E negative support contained 9,921,175 positions against 432,092,169 available to B_random. E therefore restored B_random's natural mixture while preventing exposure outside the D-reachable region. The footprint requirement is stricter than matching D's average within-support exposure field, and because E draws its positive component from the same $\Omega +$ set the combined sampler retains the same 53% pixel union.

The prespecified primary estimand was

$$\Delta_{EB} = AP2_E - AP2_B$$

11

with E - D retained as a secondary bounded contrast. The same three paired seeds, optimizer, loss, augmentation, batch size, worker count, update budget, full-scene inference procedure, and paired image-level bootstrap were used.

This intervention tests the sufficiency of restricted background support, not the identity of excluded regions in isolation. Relative to B_random, E simultaneously changes which background locations can be

sampled and sharply reduces negative-support cardinality, so E alone cannot separate exclusion of particular regions from generic loss of diversity; Condition F (Section 3.9) resolves that ambiguity. The design is not a complete prevalence-by-coverage factorial, because the prevalence-1.0 with full-support corner is non-constructible under foreground-presence conditioning.

3.9 Cardinality-Matched Control

Condition F_random_support_matched was designed after the E result to separate cardinality from spatial identity, with its protocol locked before any F outcome. F retained $p_{\text{pos}} = 0.370708$ and drew positive crops uniformly from Ω^+ , as in E. Negative crops were drawn uniformly from an exact-size random subset of Ω^- containing 9,921,175 positions. An independent subset was generated for each seed by exact multivariate-hypergeometric allocation across training images followed by uniform sampling without replacement within each image.

Unlike E, F did not confine those negative positions to D-reachable footprints. A pre-outcome audit measured the area actually reachable by each realized subset rather than assuming full coverage: 99.9812%, 99.9813%, and 99.9820% across seeds, with corresponding fractions of B_random's expected exposure on unreachable area below 2.3×10^{-7} . F therefore matched E in prevalence and eligible cardinality while restoring essentially complete spatial access.

Training, inference, AP computation, and the paired bootstrap were otherwise unchanged. The two locked contrasts were F-B and F-E. Compatibility of F with B together with a materially positive F-E would show that reduced cardinality alone was insufficient and would support the spatial identity of excluded background as the dominant component of the E deficit. Because that inference could depend on normalization, a protocol-locked GroupNorm replication was added afterward (Section 3.10).

3.10 GroupNorm E/F Replication

After the native-BatchNorm E/F results and the GroupNorm A-D curve were known, a protocol-locked replication tested E and F under the GroupNorm architecture of Section 3.6.2. Six new models were trained while the previously trained GroupNorm B_random and D_aggressive models were reused as references; only BatchNorm was replaced, by the same affine eight-group operator.

Before the six GroupNorm outcomes were inspected, E-B, F-B, and F-E were locked as a new replication family, Holm-adjusted within it, retaining the 0.020 scale. The locked expectations were that E-B would remain materially negative, that the full interval for F-B would remain inside ± 0.020 , and that F-E would remain materially positive.

After the AP2 replication decision, AP0, AP1, and AP3 were recomputed for GroupNorm B, D, E, and F from stored predictions without retraining. This tolerance sweep was explicitly post hoc, used to assess whether the contrasts changed qualitatively with matching tolerance, and could not override the locked AP2 interpretation.

4. Results

The analysis proceeds from the A-D screening experiment to the BatchNorm E/F mechanism controls and then to the GroupNorm replication. The support controls are interpreted by separating three properties of the 256-pixel sampler: positive-crop prevalence, eligible negative-support cardinality, and spatial reachability (Table 4).

Table 4

Mechanism-control design at the primary 256-pixel crop size. F used an exact-size negative subset matching E's eligible negative-support cardinality of 9,921,175 positions while restoring near-complete measured spatial reachability.

Condition	Positive prevalence	Eligible negative support	training-image area reachable	B-weighted unreachable
B_random	0.371	432,092,169	100%	0%
E_spatial_support	0.371	9,921,175	53.00%	39.06%
F_random_support_matched	0.371	9,921,175	99.98%	0.00002%
D_aggressive	1.000	not sampled	53.00%	39.06%

4.1 Sampling Response

The initial A-D experiment was a prespecified screening stage used to determine whether targeted mechanism testing was warranted. The D-B contrast therefore serves as the trigger for the mechanism programme rather than as the final mechanistic estimate; Section 4.4 shows that it is strongly scene-concentrated, and Sections 4.6 and 4.7 report the interventions that provide more direct evidence.

AP was evaluated at the four matching tolerances AP0-AP3. The four primary conditions produced a non-monotonic response rather than a consistent benefit from foreground enrichment. At AP2, B_random achieved the highest mean performance (0.8119 ± 0.0020 across seeds), followed by C_moderate (0.8075 ± 0.0068), D_aggressive (0.7890 ± 0.0029), and A_low (0.7299 ± 0.0150). The D_aggressive-minus-B_random contrast met the screening criterion at -0.02283 (paired-image 95% confidence interval (CI) [-0.04014, -0.00437], Holm-adjusted $p = 0.0360$). Moderate enrichment did not improve on random cropping (C-B = -0.00439, 95% CI [-0.01198, 0.00411], $p = 0.293$), whereas reducing positive-crop exposure to 10% produced the largest adjacent deficit (B-A = 0.08195, 95% CI [0.06430, 0.10160], $p < 0.01$). Tables 5 and 6 report the tolerance means and all six predeclared contrasts.

Table 5

Mean AP $\pm$ 1 SD across three seeds by sampling condition and tolerance under the native BatchNorm architecture.

Condition	AP0	AP1	AP2 (primary)	AP3
A_low	0.6603 $\pm$ 0.0158	0.6991 $\pm$ 0.0155	0.7299 $\pm$ 0.0150	0.7579 $\pm$ 0.0145
B_random	0.7396 $\pm$ 0.0023	0.7804 $\pm$ 0.0021	0.8119 $\pm$ 0.0020	0.8394 $\pm$ 0.0019
C_moderate	0.7339 $\pm$ 0.0077	0.7753 $\pm$ 0.0072	0.8075 $\pm$ 0.0068	0.8359 $\pm$ 0.0063
D_aggressive	0.7135 $\pm$ 0.0026	0.7557 $\pm$ 0.0027	0.7890 $\pm$ 0.0029	0.8191 $\pm$ 0.0033
E_spatial_support	0.7102 $\pm$ 0.0038	0.7523 $\pm$ 0.0037	0.7853 $\pm$ 0.0036	0.8147 $\pm$ 0.0035
F_random_support_matched	0.7379 $\pm$ 0.0076	0.7794 $\pm$ 0.0068	0.8116 $\pm$ 0.0062	0.8400 $\pm$ 0.0055
Note. E_spatial_support and F_random_support_matched were later targeted mechanism controls and were not part of the initial A-D screening family.				

Table 6

All prespecified AP2 contrasts (second minus first); A, B, C, and D denote the four primary BatchNorm conditions.

Contrast	AP2 diff.	95% CI	Holm p	Floor met?	GO met?
A -> B	+ 0.08195	[0.06430, 0.10160]	< 0.01	Yes	Yes
B -> C	-0.00439	[-0.01198, 0.00411]	0.2934	No	No
C -> D	-0.01843	[-0.03161, -0.00561]	0.0198	No	No
A -> C	+ 0.07756	[0.05758, 0.09982]	< 0.01	Yes	Yes
B -> D	-0.02283	[-0.04014, -0.00437]	0.0360	Yes	Yes
A -> D	+ 0.05912	[0.03863, 0.08109]	< 0.01	Yes	Yes

One contrast illustrates why the practical effect floor was included. C_moderate outperformed D_aggressive by 0.01843 AP2 in magnitude, with a 95% CI excluding zero and Holm-adjusted p-value 0.0198, but the magnitude remained below the prespecified 0.020 practical floor and therefore did not satisfy the GO gate.

Tolerance relaxation affected the two sides of the response differently. The B_random-to-D_aggressive deficit decreased from 0.02606 at AP0 to 0.02037 at AP3, a 21.8% reduction, with Holm-adjusted p-values of 0.0140, 0.0228, 0.0360, and 0.0728. At AP3 the contrast still met the 0.020 floor and its unadjusted 95% CI [-0.03729, -0.00130] excluded zero but no longer met the Holm-adjusted criterion;

because AP2 was fixed in advance, the screening decision is unchanged. The attenuation indicates that part of the contrast depends on the matching criterion, although with resolution-heterogeneous scenes it is not interpreted as a constant physical-distance localization effect. The B_random-to-A_low deficit was essentially unchanged, rising slightly from 0.07929 to 0.08153. Figure 1 summarizes the AP0-AP3 response across the four primary sampling conditions and the later E_spatial_support control.

4.2 Calibration

Validation-only recalibration improved probability quality while preserving the underlying score ordering. AP2 was retained from the original ranking metric and was not recomputed from transformed, requantized probability maps. For D_aggressive, Platt scaling reduced negative log-likelihood from 0.05368 to 0.04398 and expected calibration error from 0.01003 to 0.00487; at the fixed 0.5 operating point, changes in the strict and relaxed D-B harmonic mean of precision and recall (F1) gaps were only + 0.00229 and + 0.00256, with confidence intervals including zero. Calibration therefore did not explain the ranking deficit.

4.3 Normalization Effects

The exact-source BatchNorm audit showed that normalization state materially changed model behavior, especially for the two extreme conditions. Recalibration on the matched source changed AP2 by -0.00136, -0.00491, -0.00811, and + 0.00566 for A through D. Across all four common sources the AP2 range was 0.2104 for A_low and 0.1072 for D_aggressive, against 0.0246 and 0.0211 for B_random and C_moderate. The common-source D-B contrast ranged from - 0.1168 to + 0.0123, but the D_exact source raised D by only 0.00566 while lowering B by 0.02949, so the reversal was not a substantive rescue of the aggressive model.

The GroupNorm intervention provided the stronger necessity test for the B_random-versus-D_aggressive endpoint. Under GroupNorm, B_random reached AP2 = 0.84446 and D_aggressive 0.78613, giving D-B = -0.05833. Relative to the BatchNorm contrast of -0.02283, the prespecified normalization interaction was - 0.03551 (95% CI [-0.05213, -0.02057], $p < 0.001$), so replacing BatchNorm did not attenuate the deficit. The later A/C extension completed the curve: A_low = 0.78263 ± 0.03247 , B_random = 0.84446 ± 0.00255 , C_moderate = 0.83638 ± 0.00567 , and D_aggressive = 0.78613 ± 0.00421 , preserving the ordering B_random > C_moderate > D_aggressive > A_low under both normalizers although D and A were not statistically separated under GroupNorm.

Across the new six-contrast GroupNorm family, A_low -> B_random was + 0.06182 (95% CI [0.04728, 0.07762], Holm $p < 0.01$), A_low -> C_moderate + 0.05374 ($p < 0.01$), A_low -> D_aggressive + 0.00349 ([-0.01402, 0.02221], $p = 0.7223$), B_random -> C_moderate - 0.00808 ([-0.01289, -0.00357], $p < 0.01$), B_random -> D_aggressive - 0.05833 ([-0.07364, -0.04473], $p < 0.01$), and C_moderate -> D_aggressive - 0.05025 ($p < 0.01$). B_random and C_moderate were statistically separated but their difference stayed below the practical floor, and A_low and D_aggressive were not detectably different. Normalization

therefore changed absolute performance and arm asymmetry but not the qualitative inverted-U response.

The prespecified predictions for the GroupNorm extension were partly supported. That B_random would remain the peak and B-A would remain practically large was supported. C-B remained practically small but its interval excluded zero, refuting the expected compatibility with zero. The prediction that A_low seed SD would decrease was also refuted: it rose from 0.0150 under BatchNorm to 0.03247 under GroupNorm, with a seed range of 0.06286, and the spread persisted despite stable late-training diagnostics. Figure 2 compares the complete A-D sampling response under BatchNorm and GroupNorm.

The GroupNorm convergence audit did not identify an obvious endpoint-convergence explanation for the elevated A_low seed variability; full diagnostics are reported in Supplementary Methods S3.

4.4 Scene-Level Heterogeneity

The pooled D-B AP2 deficit was not uniformly expressed at the individual-scene level. D_aggressive performed worse than B_random on 105 of 200 test scenes and better on 95, while the median per-image D-B difference was only -0.00189 despite the pooled contrast of -0.02283 . Across test scenes, the mean per-image localization recovery from AP0 to AP2 was 0.00290 AP.

The influence audit showed substantial concentration of the screening effect. The normalized leave-one-out influence-share statistic assigned 22.8% to the single most influential scene; removing that scene changed D-B from -0.02283 to -0.02029 , an 11.1% attenuation. Removing the 10 most deficit-driving scenes attenuated the contrast to -0.00664 , a 70.9% reduction, and removing 20 changed it to $+0.00319$, whereas median random removal left it essentially unchanged. Ground-truth-pixel weighting moved D-B from a per-image median of -0.00189 to -0.00770 , closing only 27.8% of the gap to the pooled value.

The concentration was also aligned with filename/resolution family: the median D-B contrast was -0.0358 among the 52 IMG_* scenes and $+0.0021$ among the 148 date-stamped scenes (two-sided Mann-Whitney $p = 1.1 \times 10^{-5}$). This comparison used one per-image contrast per scene and so was not itself pixel-weighted, although the pooled endpoint does weight scenes by contributed pixels. Because these labels are not verified acquisition metadata, the result is interpreted only as descriptive resolution-associated heterogeneity, and the -0.02283 screening contrast should not be read as a uniform penalty across scenes. Figure 3 summarizes the removal trajectories.

Exploratory scene-feature analyses did not support a simple near-versus-far background explanation for the D-B deficit; detailed associations are reported in Supplementary Methods S4.

4.5 Spatial Exposure

The pixel-level exposure audit identified a much larger structural change than aggregate crop counts revealed. At the 256-pixel crop, natural positive-crop prevalence was 0.370708. Under D_aggressive, 46.997% of training-image area had zero positive-conditioned exposure and could never enter any sampled crop, and when pixels were weighted by the exposure they would receive under B_random,

39.061% of expected training exposure fell on unreachable pixels. A further 52.558% of area, or 45.570% under exposure weighting, received less than one quarter of the B_random rate.

At the opposite extreme, pixels contained in every positive crop that could cover them attained the maximum possible D-to-B exposure ratio implied by the natural positive-crop prevalence:

$$q_{max} = \frac{1}{p_{pos}} = 2.6975$$

The B-exposure-weighted mean exposure ratio was 1.000 by construction, showing why an aggregate mean does not reveal the redistribution. Thus, 100% positive-conditioned cropping simultaneously removes a large portion of training-image context and overweights portions of the remaining D-reachable support. Figure 4 illustrates the corresponding spatial-exposure geometry in a representative retained training scene.

4.6 BatchNorm Spatial-Support Control

Condition E was constructed using the natural positive-crop probability while restricting negative crops to D-reachable footprints. Its mean AP2 was 0.78528, compared with 0.81187 for B_random and 0.78905 for D_aggressive. The prespecified primary E-B contrast was -0.02659 (95% CI $[-0.03608, -0.01723]$, p -value < 0.001), establishing that the performance loss persisted under natural positive-crop probability when negative/background support remained restricted.

The secondary E-D contrast was -0.00377 (95% CI $[-0.02005, 0.01190]$, $p = 0.651$), which is not an equivalence result but bounds the detectable difference between E and D. Across AP0-AP3, E reproduced 112.8% to 121.3% of the corresponding B-D gap and was slightly below D at every tolerance. E therefore established that restricted background support was sufficient to reproduce the deficit at natural prevalence, but could not by itself distinguish spatial exclusion from the much smaller eligible negative set. Condition F tested that directly.

4.7 BatchNorm Cardinality-Matched Control

Condition F matched E's natural positive-crop prevalence and its eligible negative-support cardinality of 9,921,175 positions, but independently drew a random subset of that same size from the full Ω -domain for each seed. The spatial audit confirmed that this manipulation restored near-complete access: F reached 99.9812%-99.9820% of training-image area across seeds, while only about 0.000020%-0.000022% of B_random's expected exposure fell on F-unreachable pixels. By comparison, E reached 53.00% of training-image area and excluded 39.06% of B_random-weighted exposure.

F achieved mean AP2 = 0.81164 ± 0.00618 , essentially matching B_random (0.81187 ± 0.00197) and exceeding E_spatial_support (0.78528 ± 0.00361). The locked F-B contrast was -0.00024 with 95% CI $[-0.00500, 0.00387]$, entirely inside the ± 0.020 margin; the three paired seed differences gave a 95% t interval of $[-0.01575, 0.01527]$, also inside it, though conditional on the realized seeds and subsets. F-E

was + 0.02636 (95% CI [0.01920, 0.03389], $p < 0.001$), so F recovered 99.1% of the E-to-B deficit. The lower endpoint of the F-B interval corresponds to a residual deficit of 18.8% of the E-B point estimate, conditional on this experiment rather than a universal bound.

Across paired seeds the F-B differences were - 0.00745, + 0.00333, and + 0.00340. Because initialization and the independently drawn negative subset both varied by seed, the across-seed variation combines two sources that were not separately identified, and no consistent directional penalty appeared. F outperformed B on 116 of 200 scenes, with a median per-image difference of + 0.00094 AP2. Under native BatchNorm the approximately 44-fold cardinality reduction was therefore insufficient to reproduce E's deficit when negative positions remained dispersed. This is conditional on the fixed 20,000-update budget, under which the reduced pool still far exceeded the number of realized negative draws.

The native-BatchNorm E/F controls did not inherit the D-B concentration pattern to the same degree. For E-B, 140 of 200 scenes favored B, and ground-truth-pixel weighting moved the contrast from a per-image median of -0.00490 to -0.02293, closing 83.1% of the gap to the pooled - 0.02659. Targeted removal of the 10 most deficit-driving E scenes attenuated the deficit by 47.4%, and removal of 20 left - 0.00815 rather than reversing sign. E-B showed no detectable family difference ($p = 0.458$), and F-B's small residual family shift was far below the D-B split. D-B is therefore best treated as the heterogeneous screening signal, while the E/F interventions carry the more direct evidence.

4.8 GroupNorm E/F Replication

The later GroupNorm E/F replication produced a different mechanism pattern from the native BatchNorm experiment. At AP2, GroupNorm B_random achieved 0.84446 ± 0.00255 , D_aggressive 0.78613 ± 0.00421 , E_spatial_support 0.83426 ± 0.00286 , and F_random_support_matched 0.84314 ± 0.00165 across the three seeds. Thus, the previously established GroupNorm D-B deficit remained large at -0.05833, while the GroupNorm E and F controls clustered much closer to B_random. Table 7 reports the complete AP0-AP3 means.

Within the locked three-contrast GroupNorm replication family, E-B was - 0.01019 (95% CI [-0.01423, -0.00656], Holm $p < 0.001$); the interval excluded zero but lay entirely inside the ± 0.020 practical region, so the prediction that E would remain materially below B was not supported. F-B was - 0.00132 ([-0.00342, 0.00062], $p = 0.1984$), well inside the compatibility region, and the seed-level interval [-0.01022, 0.00758] was also inside ± 0.020 . F-E was + 0.00887 ([0.00551, 0.01252], $p < 0.001$), statistically positive but with its full interval below 0.020, so the prediction that F would remain materially above E was not supported. The locked replication was therefore mixed rather than a confirmation of the BatchNorm pattern.

The cross-normalizer comparison is informative because the contrasts moved in opposite directions. Relative to BatchNorm, the D-B deficit grew from 0.02283 to 0.05833, whereas E-B fell from 0.02659 to 0.01019 and F-E from 0.02636 to 0.00887; F-B remained practically compatible with zero under both. GroupNorm E exceeded D by 0.04814 against - 0.00377 under BatchNorm, although E and D differ in

within-support sampling density as well as prevalence. Restricted support was therefore sufficient to reproduce the BatchNorm deficit but not a normalization-invariant explanation for the larger GroupNorm deficit.

As a descriptive seed-level sensitivity check, the paired GroupNorm E-D differences were + 0.04426, + 0.05382, and + 0.04634 for seeds 42, 1337, and 2026, giving a mean of + 0.04814 and a two-sided 95% t interval of [0.03565, 0.06063]. This n = 3 interval is not a replacement for the paired test-image bootstrap used for the primary contrast families and is interpreted only as seed-level consistency evidence.

The post-hoc GroupNorm tolerance sweep showed no hidden reversal. D-B changed from - 0.06048 at AP0 to -0.05656 at AP3, only 6.49% attenuation, against 21.8% for the BatchNorm contrast; E-B changed from - 0.00908 to -0.01064, F-B from - 0.00105 to -0.00149, and F-E from + 0.00803 to + 0.00916. The GroupNorm contrasts were therefore broadly tolerance-stable, although as a post-hoc sweep this is sensitivity evidence rather than a causal decomposition. Figure 5 visualizes the AP2 pattern; a consolidated summary is in Supplementary Table S2.

Table 7

Mean AP $\pm$ 1 SD across three seeds for the GroupNorm B, D, E, and F conditions. B and D reuse the previously trained GroupNorm reference models.

Condition	AP0	AP1	AP2 (primary)	AP3
B_random	0.77317 $\pm$ 0.00197	0.81334 $\pm$ 0.00225	0.84446 $\pm$ 0.00255	0.87165 $\pm$ 0.00284
D_aggressive	0.71269 $\pm$ 0.00549	0.75370 $\pm$ 0.00485	0.78613 $\pm$ 0.00421	0.81509 $\pm$ 0.00351
E_spatial_support	0.76409 $\pm$ 0.00240	0.80364 $\pm$ 0.00255	0.83426 $\pm$ 0.00286	0.86100 $\pm$ 0.00327
F_random_support_matched	0.77212 $\pm$ 0.00165	0.81217 $\pm$ 0.00158	0.84314 $\pm$ 0.00165	0.87016 $\pm$ 0.00180

5. Discussion

Foreground-conditioned cropping changes both class composition and the set of source-image regions that can enter training. The A-D experiment showed a non-monotonic response: B_random was the best point estimate, moderate enrichment did not improve AP2, and aggressive positive-only sampling was harmful. The GroupNorm curve preserved that ordering, so the qualitative response did not require BatchNorm, but the mechanism evidence differed. Under native BatchNorm, E reproduced the deficit at natural prevalence and F restored near-complete spatial access at the same cardinality and returned to B_random; under GroupNorm, E-B and F-E both fell inside the $\pm$ 0.020 region while D-B grew. Spatial restriction is therefore a demonstrated mechanism for the native-BatchNorm deficit but not a normalization-invariant explanation of aggressive-sampling degradation.

The initial D-B GO contrast was also strongly heterogeneous: removing 10 deficit-driving scenes attenuated it by 70.9%, removing 20 reversed its sign, and the deficit concentrated in one filename/resolution family. This does not invalidate the prespecified screening result but fixes its role. D-B established a reproducible problem worth explaining; the E/F interventions identify a spatial-support mechanism within one normalization setting; and the GroupNorm replication shows the mechanism cannot be assumed to dominate under another.

5.1 Prevalence and Spatial Support

Foreground sampling changes not only class composition but also the spatial support available during training. At the 256-pixel crop size, 37.1% of uniformly sampled positions contained crack, while positive-only conditioning made 47.0% of training-image area unreachable. Under B_random exposure weighting, 39.1% of expected training exposure fell on pixels that D_aggressive could never reach, so the exclusion is not an artifact of counting border pixels equally. The two measures describe different quantities: physical coverage, and the share of baseline training signal removed.

Under native BatchNorm, Condition E restored the natural positive-crop prevalence while retaining restricted spatial support. Condition F held both that prevalence and E's eligible negative-support cardinality fixed, drawing an independent random subset of the same size from the full crack-negative domain, and therefore differed from E in the spatial identity and distribution of eligible negative positions. F restored 99.98% spatial reachability, returned to B_random ($F-B = -0.00024$), and remained materially above E. Within the BatchNorm experiment the E deficit was therefore not explained by reduced cardinality alone; the spatial identity and distribution of excluded background were the dominant tested factors.

The GroupNorm replication limits that interpretation. Under GroupNorm, E-B and F-E were both practically small while D-B more than doubled in magnitude, and E exceeded D rather than matching it. Restricted support therefore did not reproduce the aggressive-sampling deficit at a material magnitude under GroupNorm, although E and D also differ in within-support sampling density, so that reversal is not an isolated prevalence effect. What survives across both normalizers is narrower: an approximately 44-fold reduction in eligible negative-support cardinality was not itself materially harmful when the retained positions remained spatially dispersed, whereas the consequence of spatial exclusion depended on normalization.

The tested conditions do not form a complete prevalence-by-coverage factorial. Prevalence 1.0 with full spatial support is not constructible under crack-positive conditioning, because regions sufficiently far from cracks cannot generate positive crops. The mechanism claims are therefore conditional on the tested sampling and normalization settings.

5.2 Relation to Prior Work

Crack-focused filtering and adaptive sampling can be useful within their intended pipelines, but a positive-patch rate does not identify which source-image regions remain accessible. Table 1 lists the six most relevant prior studies, and the present evidence bears on each. The exposure audit shows that the 1,000-crack-pixel region filter of Yang et al. (2020) is a spatial-selection rule as well as a rebalancing rule, and the crack-patch-only supervision of K. Zhang et al. (2021) avoids an all-background failure mode while restricted support carried a separable cost here under BatchNorm. The contextual-information warning of Nguyen and Thai (2023) is operationalized by the E/F controls, while the crop-size sweep shows that the severity of crack-focused cropping is strongly crop-size dependent. The E/F mechanism interpretation is conditional on normalization.

The closest sampling comparison is Li et al. (2024), who reported adaptive probabilistic sampling outperforming random cropping in their tested configurations. The present samplers are not replications of that strategy, so those gains are not contradicted. The current results instead bound a broader interpretation: stronger target-biased sampling was not uniformly beneficial in this fixed-architecture Crack500 experiment, and the mechanism associated with aggressive conditioning changed with normalization.

The exposure analysis operationalizes the contextual-information warning by measuring which area and which baseline training signal are removed. Under BatchNorm the E/F controls show that spatial restriction, rather than reduced eligible-coordinate count, can reproduce a material deficit; under GroupNorm the same restriction produces only a small effect while aggressive conditioning remains strongly harmful. The contribution is therefore that sampling geometry is measurable, experimentally separable from cardinality, and capable of interacting with the rest of the training system.

5.3 Calibration and Normalization

The negative controls narrow the plausible mechanism space. Monotonic recalibration improved negative log-likelihood, Brier score, and expected calibration error while preserving score ordering, and Platt scaling moved the fixed-threshold operating point while recovering only a small, statistically uncertain fraction of the F1-score gap. The aggressive-sampling effect therefore cannot be reduced to a poor probability scale or an unfortunate decision threshold.

Normalization state was consequential, but its role is more specific than a simple BatchNorm-mismatch explanation. Re-estimating BatchNorm statistics on different exact sources changed the apparent contrast, yet the matched D source improved D only modestly and the common-source sign reversal was driven mainly by degradation of B_random. The GroupNorm replacement preserved the ordering $B_{\text{random}} > C_{\text{moderate}} > D_{\text{aggressive}} > A_{\text{low}}$ while the B-A versus B-D arm-magnitude ratio fell from 3.59 to 1.06. BatchNorm running-state dependence is therefore not necessary for the response around the B_random peak, but normalization materially changes its magnitude and asymmetry.

The GroupNorm replication shows that the mechanism pattern also differed: under native BatchNorm E-B and F-E were materially large, whereas under GroupNorm both fell within the ± 0.020 region even

though D-B increased. Because GroupNorm changes both running-state dependence and the normalization operator, these experiments cannot identify which property interacts with aggressive sampling. They do show that the harm from D_aggressive is robust to replacing BatchNorm, whereas the sufficiency of the E restriction is not.

The scene-level analysis likewise failed to support a localized hard-background explanation. Near-crack and far-background gradient and edge-density associations were nearly identical, and the strongest exploratory patterns were modest correlations with broader scene texture and crack-network fragmentation. These descriptors did not provide a positive mechanistic explanation and do not substitute for the controlled interventions.

This also clarifies what can and cannot be inferred about standard class-imbalance remedies. Monotonic recalibration demonstrably cannot restore a ranking deficit in the learned scores. Loss reweighting was not tested and should not be called ineffective, although a class weight changes the contribution of observed pixels without making an excluded region appear in training at all. Whether reweighting compensates for restricted support remains open.

5.4 Practical Implications

The pixel-level exposure fields of Section 3.7 provide a low-cost pretraining diagnostic, requiring only image dimensions, masks, crop size, and the sampling rule. They identify area that is unreachable or underexposed relative to uniform random cropping and quantify how exposure is redistributed before any model is fitted, turning an often undocumented choice into an auditable part of the protocol. As the GroupNorm replication shows, they characterize the intervention rather than predicting performance magnitude, and because sampling geometry is crop-size dependent, crop size should be reported alongside prevalence and reachability.

Studies using foreground-conditioned crops should therefore report the crop dimensions, the exact sampling rule, the natural positive-crop prevalence under uniform cropping, and a measure of reachable or excluded support. Where the eligible crop set is restricted, cardinality should be reported separately from spatial coverage, because Condition F shows the two are not interchangeable. These recommendations concern transparent characterization and do not imply that the effects observed here generalize unchanged across architectures, datasets, crop sizes, or normalization strategies.

5.5 Limitations

The findings should be interpreted within the tested scope. All training experiments used one U-Net architecture, one audited Crack500 split, a 256-pixel crop size, and three initialization seeds, so the paired bootstrap quantifies test-image uncertainty conditional on the trained models rather than population-level variation across initializations. All 200 held-out scenes contained cracks, so the evaluation characterizes segmentation within crack-containing pavement rather than image-level crack-presence detection under deployment prevalence. Broader replication across architectures, datasets, and foreground sparsity regimes is needed before the effect magnitudes are treated as general.

The spatial quantities are likewise specific to the tested data and crop geometry: the 47.0% unreachable-area and 39.1% exposure figures apply to the retained Crack500 training set at 256 pixels. The geometric sweep showed that these change substantially with crop size, but models were not retrained at 128 or 512 pixels, so the sweep establishes geometric sensitivity rather than a corresponding AP2 trend. AP0-AP3 tolerances are defined in image-pixel rather than physical coordinates.

The E/F controls isolate important components of the intervention but do not form a complete factorial design. F matched E's eligible negative-support cardinality while restoring near-complete reachability, but its random subset also differed from E in the across-image distribution of negative coordinates. The prevalence-1.0 with full-support condition is not constructible under foreground-presence conditioning, so D_aggressive necessarily combines complete conditioning with restricted support. The GroupNorm replication further limits generalization, since restricted support reproduced a material deficit under BatchNorm but not under GroupNorm.

Two evaluation details also bound interpretation. First, AP was computed using a fixed 2,048-bin histogram approximation after 16-bit score quantization; post hoc rescoring through the full 65,536-level stored representation showed that no prespecified contrast crossed the 0.020 floor as resolution increased, but this check was possible only for the twelve native BatchNorm A-D runs and not for the E/F or GroupNorm controls, for which probability maps were not retained. Second, the comparison of image-source and resolution groups used per-image contrasts and is interpreted descriptively rather than as a decomposition of pooled AP.

6. Conclusion

Foreground-conditioned crop sampling is not only a class-composition intervention; it also changes which source-image regions can enter training. In the native-BatchNorm screening experiment, B_random produced the highest AP2, moderate enrichment did not improve on random cropping, and D_aggressive triggered the prespecified GO criterion. A_low was also substantially worse than B_random, so the data do not support simply minimizing foreground exposure. The GroupNorm curve preserved the same ordering, so the non-monotonic response did not require BatchNorm, even though normalization changed effect magnitudes and arm asymmetry.

The strongest mechanistic evidence under the native BatchNorm protocol came from the matched E/F controls. E restored natural positive-crop prevalence while restricting background support and remained materially below B_random. F matched E's eligible negative-support cardinality exactly while restoring 99.98% reachability, returned to B_random performance, and exceeded E. Within BatchNorm and the fixed training budget, reducing cardinality alone was insufficient to reproduce the deficit, while the spatial identity and distribution of excluded background were the dominant tested component.

The GroupNorm replication prevents that mechanism from being generalized uncritically: E-B and F-E both remained inside the prespecified ± 0.020 practical region and F-B remained compatible with B, while D-B was much larger. Restricted spatial support was therefore sufficient under one normalizer but was

not a normalization-invariant explanation for aggressive-sampling degradation, and calibration and scene descriptors likewise did not account for the ranking effect. Foreground-focused sampling should accordingly be reported and evaluated as a spatial data-selection rule, with natural positive-crop prevalence, crop geometry, spatial reachability, and eligible-support cardinality documented separately. The conclusions remain conditional on one U-Net architecture, one retained Crack500 split, 256-pixel crops, three initialization seeds, and a crack-containing test set.

Declarations

Author Contribution

Ehsan Ghaffari: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing - original draft, Writing - review and editing. Kelvin C. P. Wang: Conceptualization, Methodology, Supervision, Resources, Writing - review and editing. Both authors reviewed and approved the final manuscript.

Acknowledgement

The authors acknowledge the institutional support of the Western Transportation Institute and thank Clayton Donally for his collegial support and collaboration throughout this work.

Data Availability

The Crack500 data used in this study are publicly available through the original authors' pavement-crack-detection project repository at <https://github.com/fyangneil/pavement-crack-detection>, which provides the dataset download links. The dataset is not redistributed by the authors. Code for crop-domain enumeration, training, full-scene inference, exposure-field computation, calibration, and paired-bootstrap evaluation, together with protocol-lock documents and derived preparation/alignment artifacts, is available from the corresponding author upon reasonable request.

References

1. Abdollahi, C., & Okte, E. (2026). Evolution of automated crack, pothole, and roughness detection and assessment for flexible pavements: state-of-the-art review. *Automation in Construction*, 191, 107167. <https://doi.org/10.1016/j.autcon.2026.107167>
2. Al-Huda, Z., Peng, B., Algburi, R. N. A., Al-antari, M. A., AL-Jarazi, R., Al-maqtari, O., & Zhai, D. (2023). Asymmetric dual-decoder-U-Net for pavement crack semantic segmentation. *Automation in Construction*, 156, 105138. <https://doi.org/10.1016/j.autcon.2023.105138>
3. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1),

- 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
4. Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106, 249–259. <https://doi.org/10.1016/j.neunet.2018.07.011>
 5. Chen, Z., Lai, Z., Chen, J., & Li, J. (2024). Mind marginal non-crack regions: Clustering-inspired representation learning for crack segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12698–12708. <https://doi.org/10.1109/CVPR52733.2024.01207>
 6. Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.
 7. Fan, L., Tang, S., Ariffin, M. K., A. b., M., Ismail, M. I. S., & Wang, X. (2025). Impact of image preprocessing and crack type distribution on YOLOv8-based road crack detection. *Sensors (Basel, Switzerland)*, 25(7), 2180. <https://doi.org/10.3390/s25072180>
 8. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of Machine Learning Research, 70, 1321–1330.
 9. Guo, X., Tang, W., Wang, H., Wang, J., Wang, S., Qu, X., & Lin, X. (2025). MorFormer: Morphology-aware Transformer for generalized pavement crack segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 26(6), 8219–8232. <https://doi.org/10.1109/TITS.2025.3558782>
 10. Han, C., Ma, T., Huyan, J., Huang, X., & Zhang, Y. (2022). CrackW-Net: A novel pavement crack image segmentation convolutional neural network. *IEEE Transactions on Intelligent Transportation Systems*, 23(11), 22135–22144. <https://doi.org/10.1109/TITS.2021.3095507>
 11. Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
 12. Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of Machine Learning Research, 37, 448–456.
 13. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), 203–211. <https://doi.org/10.1038/s41592-020-01008-z>
 14. Lau, S. L. H., Chong, E. K. P., Yang, X., & Wang, X. (2020). Automated pavement crack segmentation using U-Net-based convolutional neural network. *Ieee Access : Practical Innovations, Open Solutions*, 8, 114892–114899. <https://doi.org/10.1109/ACCESS.2020.3003638>
 15. Li, F., Mou, Y., Zhang, Z., Liu, Q., & Jeschke, S. (2024). A novel model for the pavement distress segmentation based on multi-level attention DeepLabV3+. *Engineering Applications of Artificial Intelligence*, 137, 109175. <https://doi.org/10.1016/j.engappai.2024.109175>
 16. Lim, S., Kim, D., Kim, H., Jeong, J., Choi, C., & Ahn, H. (2026). Assessing the domain gap between public crack datasets and UAV imagery for deep learning-based crack segmentation. *Applied Sciences*, 16(14), 6975. <https://doi.org/10.3390/app16146975>
 17. Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2980–2988.

<https://doi.org/10.1109/ICCV.2017.324>

18. Liu, H., Yang, J., Miao, X., Mertz, C., & Kong, H. (2023). CrackFormer network for pavement crack segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 24(9), 9240–9252. <https://doi.org/10.1109/TITS.2023.3266776>
19. Loshchilov, I., & Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. International Conference on Learning Representations (ICLR). <https://openreview.net/forum?id=Skq89Scxx>
20. Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. International Conference on Learning Representations (ICLR). <https://openreview.net/forum?id=Bkg6RiCqY7>
21. Ma, N., Song, Z., Hu, Q., Liu, C. W., Han, Y., Zhang, Y., Fan, R., & Xie, L. (2027). Vehicular road crack detection with deep learning: An online benchmark for comprehensive evaluation of existing algorithms. *Advanced Engineering Informatics*, 77, 105158. <https://doi.org/10.1016/j.aei.2026.105158>
22. Mao, N., Ding, S., Zhang, Y., Chen, X., Ai, C., & Zhou, B. (2026). Context-guided hard-negative background suppression for crack segmentation on complex-texture farmland roads. *Sensors (Basel, Switzerland)*, 26(16), 5185. <https://doi.org/10.3390/s26165185>
23. Milletari, F., Navab, N., & Ahmadi, S. A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. 2016 Fourth International Conference on 3D Vision (3DV), 565–571. <https://doi.org/10.1109/3DV.2016.79>
24. Mulyanto, A., Sari, R. F., Muis, A., & Harwahu, R. (2025). Vision-based automated pavement distress inspection: A review. *Ieee Access : Practical Innovations, Open Solutions*, 13, 115474–115502. <https://doi.org/10.1109/ACCESS.2025.3582976>
25. Nguyen, Q. D., & Thai, H. T. (2023). Crack segmentation of imbalanced data: The role of loss functions. *Engineering Structures*, 297, 116988. <https://doi.org/10.1016/j.engstruct.2023.116988>
26. Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), *Advances in large margin classifiers* (pp. 61–74). MIT Press.
27. Reményi, I., Kiss, Á., Kárász, Z., & Botzheim, J. (2026). Enhancing recall in asphalt pavement crack segmentation using a HoughNet-extended U-Net architecture with reconstruction head. *Scientific Reports*, 16, 1889. <https://doi.org/10.1038/s41598-025-31492-2>
28. Rill-García, R., Dokládlová, E., & Doklád, P. (2022). Pixel-accurate road crack detection in presence of inaccurate annotations. *Neurocomputing*, 480, 1–13. <https://doi.org/10.1016/j.neucom.2022.01.051>
29. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention, MICCAI 2015 (LNCS 9351, pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28
30. Shrivastava, A., Gupta, A., & Girshick, R. (2016). Training region-based object detectors with online hard example mining. Proceedings of the IEEE Conference on Computer Vision and Pattern

- Recognition (CVPR), 761–769. <https://doi.org/10.1109/CVPR.2016.89>
31. Silva, G., Almegdadi, O., Batista, F., Marcelino, J., & Marques, N. C. (2026). Crack detection in pavement imagery: Evaluating U-Net variants. In *Progress in Artificial Intelligence, EPIA 2025, Part I* (LNAI 16121, pp. 136–148). Springer. https://doi.org/10.1007/978-3-032-05176-9_11
 32. Siriborvornratanakul, T. (2023). Pixel-level thin crack detection on road surface using convolutional neural network for severely imbalanced data. *Computer-Aided Civil and Infrastructure Engineering*, 38(16), 2300–2316. <https://doi.org/10.1111/mice.13010>
 33. Tang, Y., Zhang, A. A., Luo, L., Wang, G., & Yang, E. (2021). Pixel-level pavement crack segmentation with encoder-decoder network. *Measurement*, 184, 109914. <https://doi.org/10.1016/j.measurement.2021.109914>
 34. Tao, H. (2024). Weakly-supervised pavement surface crack segmentation based on dual separation and domain generalization. *IEEE Transactions on Intelligent Transportation Systems*, 25(12), 19729–19743. <https://doi.org/10.1109/TITS.2024.3464528>
 35. Wu, K., Peng, B., & Zhai, D. (2025). Boundary-aware axial attention network for high-quality pavement crack detection. *IEEE Transactions on Neural Networks and Learning Systems*, 36(7), 13555–13566. <https://doi.org/10.1109/TNNLS.2024.3497145>
 36. Wu, Y., & He, K. (2018). Group normalization. *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19. https://doi.org/10.1007/978-3-030-01261-8_1
 37. Yang, F., Zhang, L., Yu, S., Prokhorov, D., Mei, X., & Ling, H. (2020). Feature pyramid and hierarchical boosting network for pavement crack detection. *IEEE Transactions on Intelligent Transportation Systems*, 21(4), 1525–1535. <https://doi.org/10.1109/TITS.2019.2910595>
 38. Yin, T., Zhang, W., Kou, J., & Liu, N. (2025). Promoting automatic detection of road damage: A high-resolution dataset, a new approach, and a new evaluation criterion. *IEEE Transactions on Automation Science and Engineering*, 22, 2472–2484. <https://doi.org/10.1109/TASE.2024.3379945>
 39. Yu, Z., Chen, Q., Shen, Y., & Zhang, Y. (2024). Robust pavement crack segmentation network based on transformer and dual-branch decoder. *Construction and Building Materials*, 453, 139026. <https://doi.org/10.1016/j.conbuildmat.2024.139026>
 40. Zhang, K., Zhang, Y., & Cheng, H. D. (2021). CrackGAN: Pavement crack detection using partially accurate ground truths based on generative adversarial learning. *IEEE Transactions on Intelligent Transportation Systems*, 22(2), 1306–1319. <https://doi.org/10.1109/TITS.2020.2990703>
 41. Zhang, Y., Fan, J., Zhang, M., Shi, Z., Liu, R., & Guo, B. (2022). A recurrent adaptive network: Balanced learning for road crack segmentation with high-resolution images. *Remote Sensing*, 14(14), 3275. <https://doi.org/10.3390/rs14143275>
 42. Zhang, Y., & Zhang, L. (2024). Detection of pavement cracks by deep learning models of Transformer and UNet. *IEEE Transactions on Intelligent Transportation Systems*, 25(11), 15791–15808. <https://doi.org/10.1109/TITS.2024.3420763>
 43. Zhong, J., Ma, Y., Zhang, M., Xiao, R., Cheng, G., & Huang, B. (2024). A pavement crack translator for data augmentation and pixel-level detection based on weakly supervised learning. *IEEE Transactions*

Figures

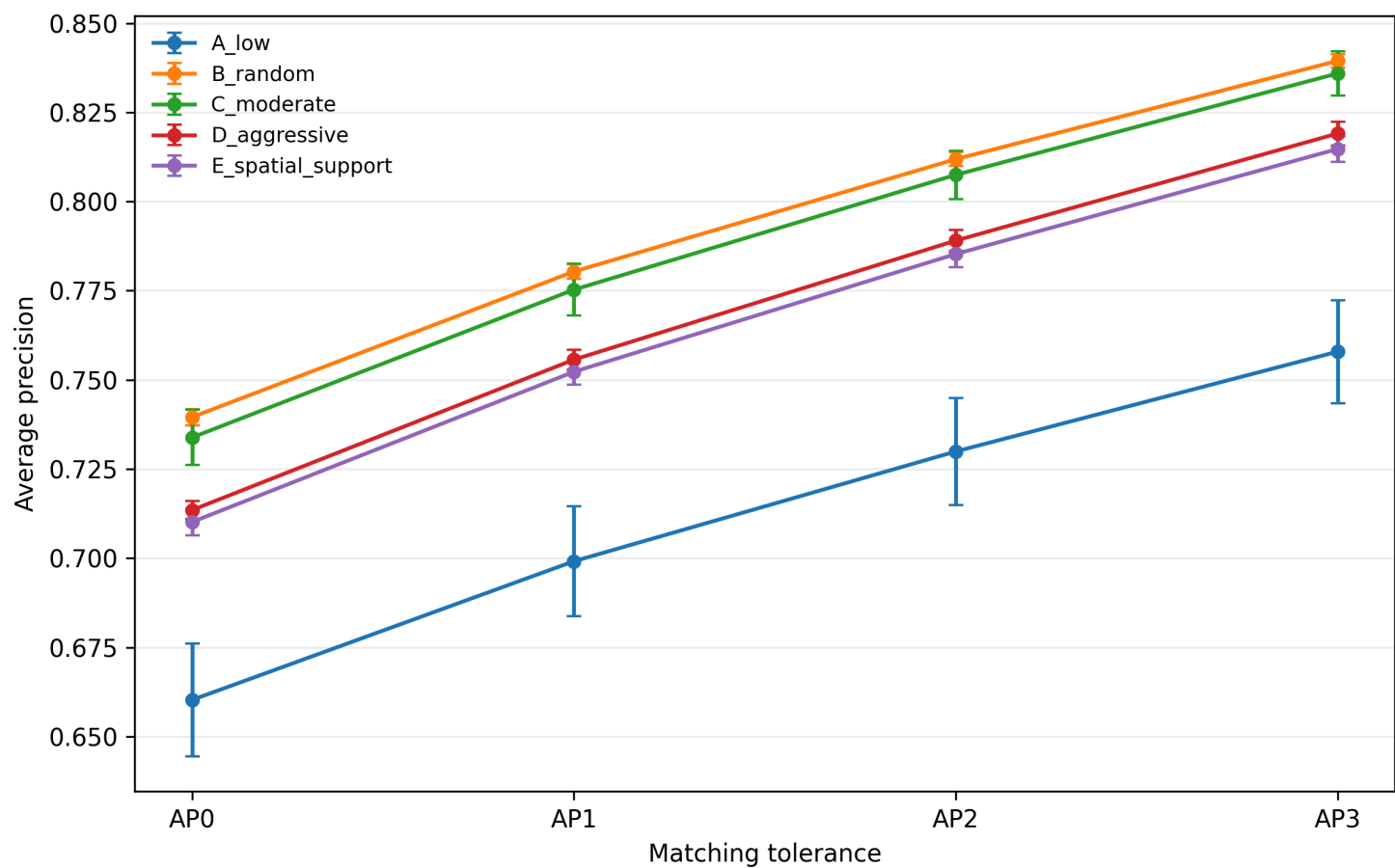


Figure 1

Mean AP0-AP3 ± 1 SD across three seeds for the four primary sampling conditions and the later E_spatial_support control, which was not part of the A-D screening family.

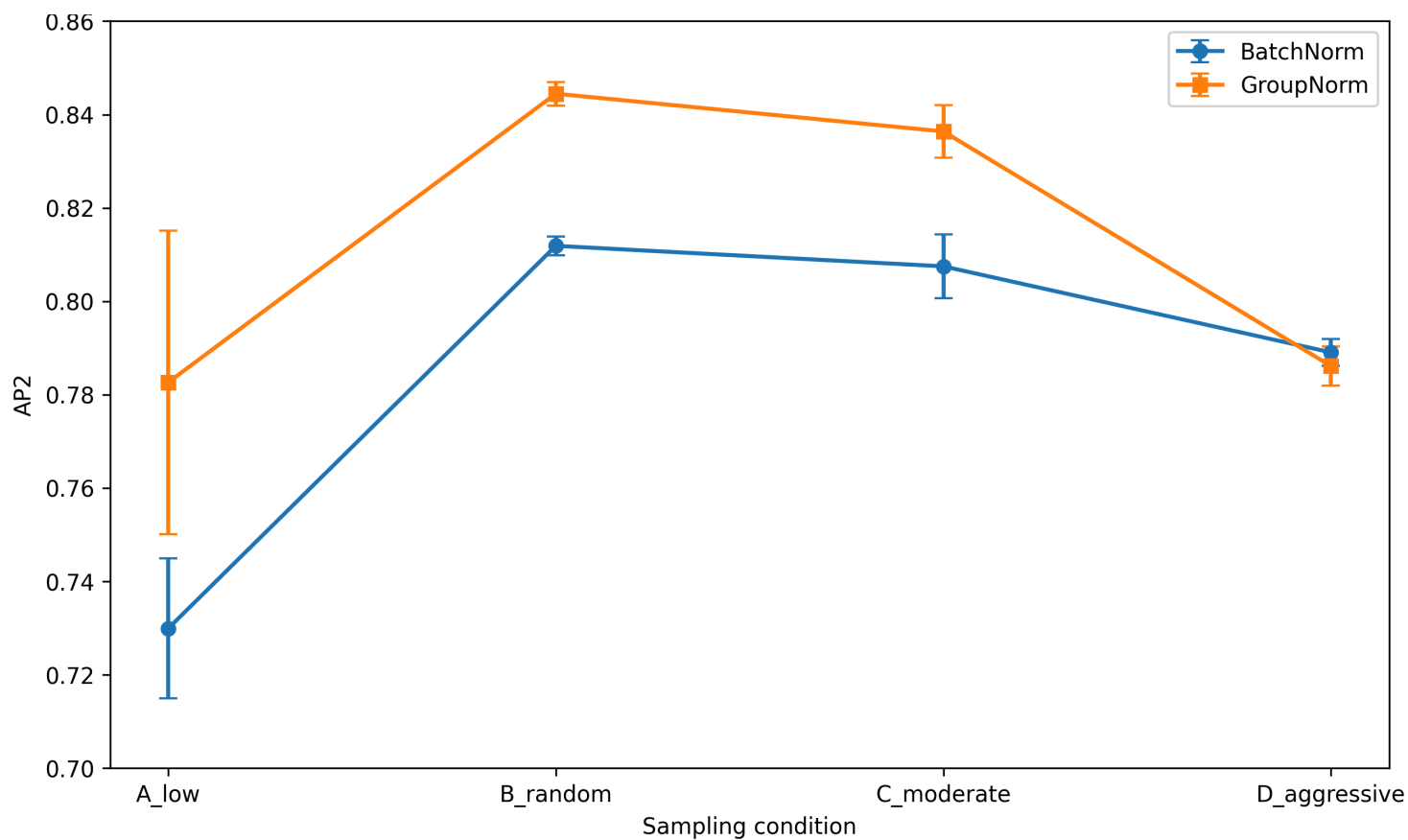


Figure 2

Mean AP2 $\pm$ 1 SD across three seeds for the four sampling conditions under native BatchNorm and GroupNorm. The ordering is B_random > C_moderate > D_aggressive > A_low under both normalizers, although D_aggressive and A_low are not statistically separated under GroupNorm. Error bars show across-seed SD.

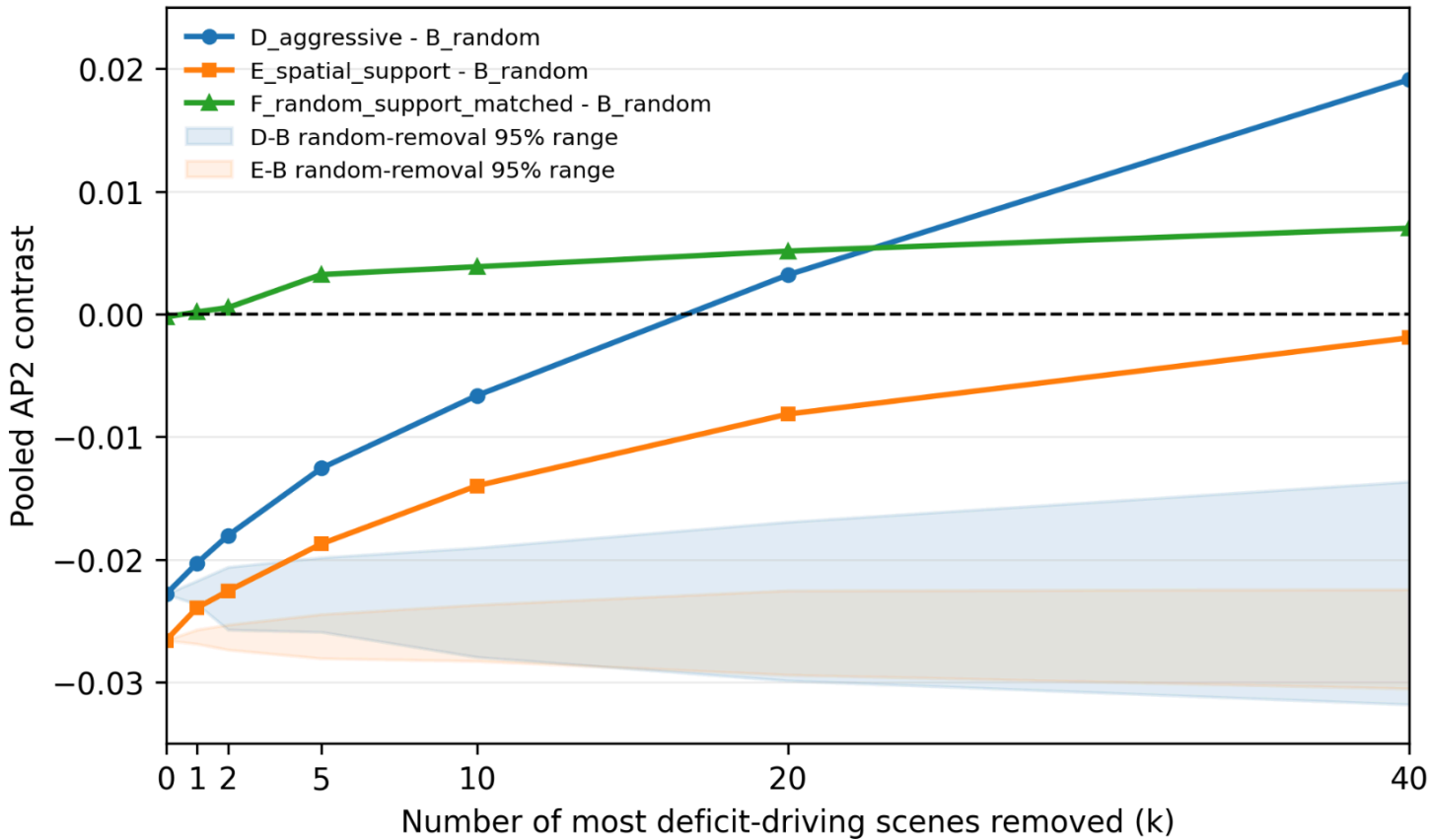


Figure 3

Influence-removal analysis at AP2. Curves show the pooled contrast after removing the k most deficit-driving scenes for D-B, E-B, and F-B. D-B crosses zero after 20 targeted removals, E-B remains negative through 40, and F-B remains near zero. Shaded envelopes show the 95% random-removal reference ranges for D-B and E-B.

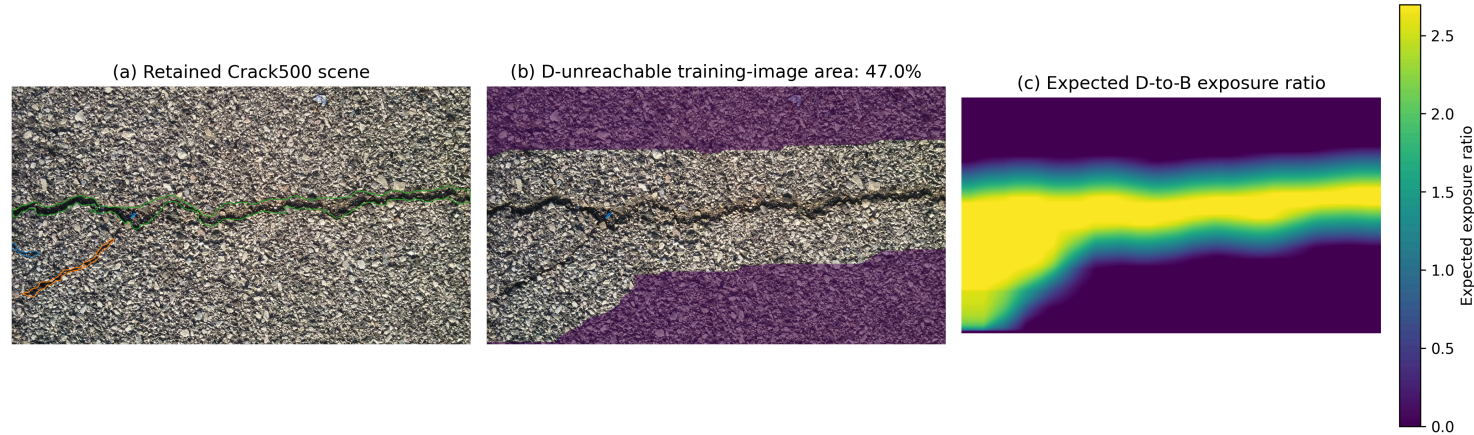


Figure 4

Representative retained Crack500 training scene, selected as the frame whose per-image unreachable fraction was closest to the dataset-wide value. The middle panel shades area with zero exposure

probability under D_aggressive at the 256-pixel crop; the right panel shows the expected D-to-B exposure ratio. The 47.0% and 39.1% values reported in the text are dataset-level quantities.

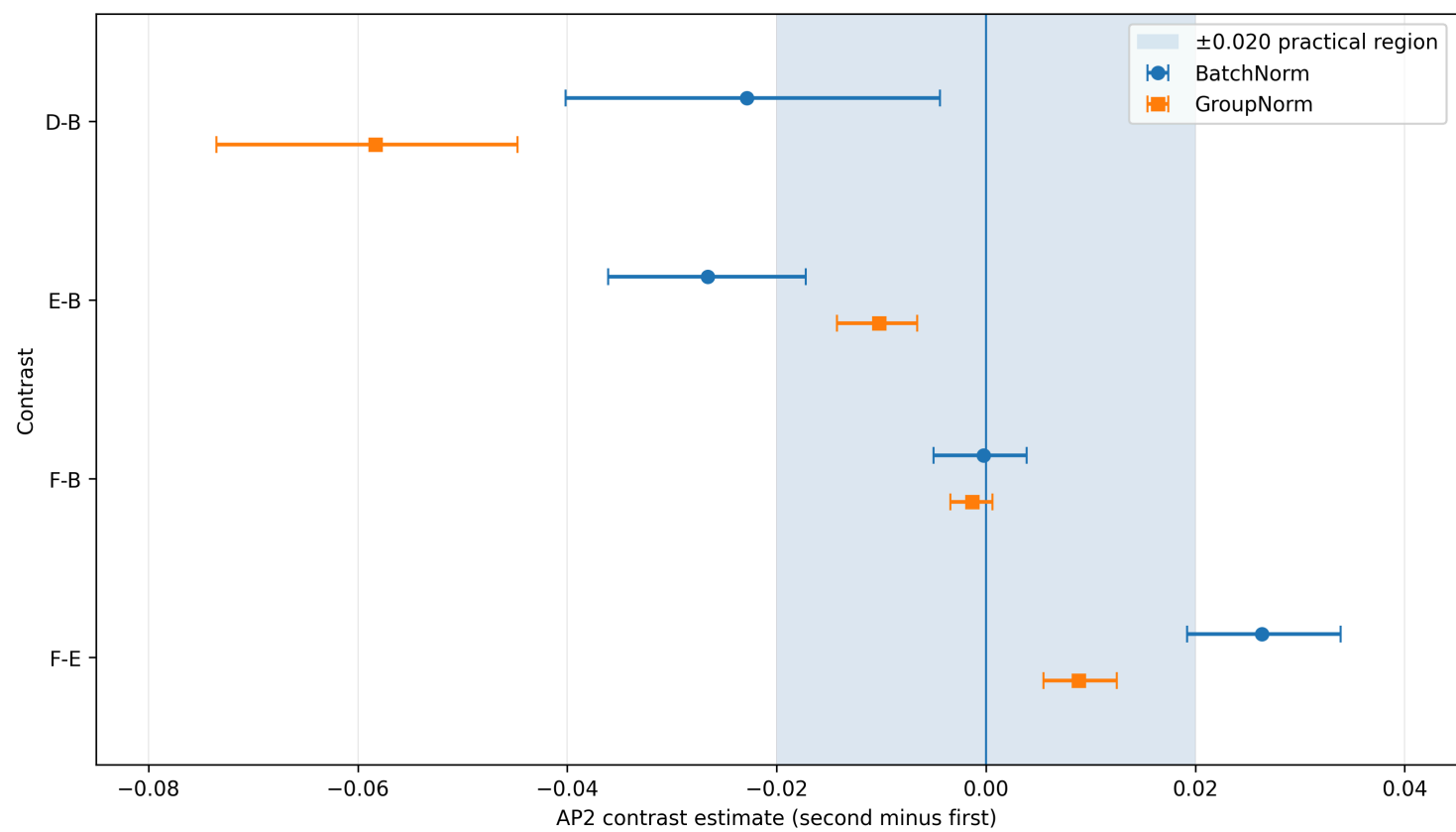


Figure 5

AP2 mechanism contrasts under native BatchNorm and GroupNorm. Points show pooled three-seed estimates and horizontal bars paired-image 95% confidence intervals; the shaded region denotes the prespecified ± 0.020 practical region.

Supplementary Files

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

- [ForegroundSamplingSpatialSupportSupplementary.docx](#)