

Reliability-Aware Physics-Informed Perceiver Resampler for Multimodal Defect Inspection of Prefabricated Components

Dominik Szewczyk^{1,*} and Wiktoria Lech¹

¹ Faculty of Electronics and Information Technology, Gdansk University of Technology, Gdansk, 80-233, Poland

*Corresponding author: dominik.sz@pg.edu.pl

Abstract. Automated inspection of prefabricated concrete components remains unreliable when cracks, honeycombing, spalling, exposed reinforcement, and dimensional deviations occur under variable illumination or resemble legitimate manufacturing features. This study proposes a reliability-aware, physics-informed Perceiver Resampler that fuses RGB texture with depth geometry, thermal diffusion residuals, and design-tolerance fields before fixed-length latent resampling. The principal methodological contribution is to use calibrated physical reliability to regulate token selection rather than append sensor channels without quality control. Reliability-weighted physical tokens, defect-adaptive latent queries, and a boundary-consistency decoder enable high-resolution multisensor observations to be processed with a stable computational budget. A replaceable experimental protocol is defined on 18,640 annotated component views collected from factory and construction-site scenes. Illustrative, non-final results show 91.8% mAP50, 84.6% mask mIoU, and 89.3% defect recall, exceeding a visual-only Perceiver baseline by 4.9, 5.7, and 6.8 percentage points, respectively. Under low illumination and partial depth loss, recall decreases by 3.7 percentage points, compared with 9.8 points for the baseline. The resampler processes a 2048 x 1536 multisensor sample in 28.4 ms with 3.1 GB peak memory. Component-level splitting, controlled sensor perturbations, module ablations, confidence calibration, and edge-device profiling are used to examine accuracy and operational reliability. The findings suggest that physical residuals are most effective when they regulate evidence selection. All reported values are illustrative placeholders and must be replaced with verified measurements before submission.

Keywords: *Prefabricated Components, Defect Detection, Physics-Informed Fusion, Perceiver Resampler, Multimodal Inspection*

Received on 03 June 2022, Accepted on 20 October 2022, Published on 30 October 2022

Copyright © 2022 Author(s), licensed to DEA. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

Introduction

A large portion of the quality formation process has been transferred from the construction site to a controlled manufacturing environment thanks to prefabricated concrete construction; nonetheless, the risks have been concentrated in molds, lifting points, embedded parts, joints, transportation edges, etc. Honeycombing denotes a loss of cover and decreased durability, whereas a thin crack could be the result of early shrinking or damage during transit. A noticeable discontinuity in texture or shape that deviates from the component's specified condition is called a surface defect [1]. Manual inspection is still somewhat prevalent; depending on their experience and level of fatigue, technicians can relate the appearance to the type of component, casting orientation, and permitted tolerance [2]. Conventional machine vision creates defect candidates from hand-crafted textures, color contrast, and edges [3]. These indicators are susceptible to damage when exposed to oblique light, in wet concrete, or in close proximity to formwork marks and chamfers [4]. Although deep convolutional detectors enhance representation learning, material state is still primarily inferred from statistical appearance [5]. For safety-related examinations, this flaw is not an issue because a dark patch is probably innocuous and a tiny debonding area may be nearly undetected but generates a detectable thermal or sonic signal.

Erkent et al. [6] By using complementary observation, many inspection modalities reduce ambiguity. Geometric discontinuity and dimensional deviation are provided by depth sensing. By using the transient heat-flow approach, infrared thermography can identify near-surface delamination; however, it is susceptible to environmental and emissivity fluctuations [7]. Tolerance testing and edge-damage measuring are supported by 3D Point Clouds [8]. Although Vision Transformer models have a high token cost at high-resolution images, they have long-range context [9]. By using a fixed number of learnt queries for cross-attention across a large input set, latent-array architectures lower this cost [10]. Because it transforms variable-resolution and variable-modality inputs into a stable latent interface, a Perceiver Resampler is appropriate for component examination [11]. Nevertheless, the majority of existing fusion solutions only concatenate sensor features without assessing the accuracy, consistency, or registration of the physical observations in relation to component requirements. As a result, they will intensify a backdrop gradient caused by heat or a misaligned depth edge.

This study investigates whether reliability-controlled physical residuals can guide latent token selection more effectively than direct multimodal concatenation. The work makes three contributions. First, it represents depth, thermal, and design information as residuals relative to engineering expectations and weights them according to calibration and local signal quality. Second, it introduces global, boundary, and defect-prototype query banks that allocate a fixed latent budget to component context and defect-sensitive regions. Third, it combines multiscale boundary decoding with confidence-aware training and an operational rejection policy for incomplete or unreliable sensing. Accuracy, ablation, robustness, calibration, and edge efficiency are evaluated within one integrated experimental framework.

Related Work

Visual and Geometric Defect Inspection

To identify fractures or porous areas, the earliest automated inspection pipelines employed thresholding, morphology, edge linking, and texture descriptors [12]. The decision boundary will vary if the aggregate color or light changes because most of the parameters are set for a particular surface finish [13]. By learning hierarchical filters, convolutional neural networks lessened this reliance and have been applied to pixel-level segmentation, bounding-box detection, and classification [14]. The detection of faults with varying widths is improved by an order of magnitude using feature-pyramid designs [15]. However, casting seams, handwritten inscriptions, release-agent traces, and shadows are still mistaken for damage by image-only systems. Incorporate a third dimension into the examination and measure warpage, corner loss, and insert displacement using plane fitting and point-to-design comparison [16]. A few millimeters of pose mismatch can be regarded as a tolerance violation, and registration mistake is still a practical issue [17].

Images and point clouds or depth maps have recently been integrated into the component-inspection system. Point-based fusion retains irregular space samples, while projection-based fusion translates geometric information to image pixels. Both approaches enhance edge localization, but instead of treating geometry as a limitation on what constitutes a defect, they typically regard it as another learnt attribute. An intended chamfer, for instance, is a depth discontinuity but not damage; its interpretation depends on the planned component shape. As a result, raw-channel concatenation is replaced by residual physical encoding.

Multimodal Transformers and Physics-Informed Learning

Global connections for remote regions and modalities are provided via transformer fusion [18]. Although dense attention across megapixel inspection images is expensive, cross-attention can probe one modality with another [19]. By condensing numerous observations into a small array, latent bottlenecks lower this expense [20]. For the next language or decision models, a resampling module is used to generate fixed-length visual tokens [21]. Nevertheless, the chosen evidence may be physically impossible but visually predictive because the learnt inquiries are typically only optimized by the task loss.

In data-driven models, physics-informed learning incorporates admissible-state penalties, conservation restrictions, or governing equations [22]. Due to lacking material qualities, boundary conditions, and sensor calibration, precise field equations at the pixel level are rarely available in civil inspection. Weak physics—local planarity of manufactured surfaces, diffusion-consistent thermal background, tolerance envelopes from design

specifications, and monotonic relations between crack opening and depth discontinuity—is a workable substitute. The research gap is that latent token generation can be guided by these flawed but quantifiable correlations without turning the network into an unrealistic deterministic model. Thus, this paper's defect-adaptive queries and reliability-gated residual tokens solve the issue.

Physics-Informed Perceiver Resampler

Problem Formulation and Physical Residual Encoding

An inspection sample contains an RGB image, a registered depth field, a thermal frame, a design-derived tolerance map, and component metadata. The model predicts defect classes, boxes, and masks while keeping the latent budget independent of sensor resolution. Raw modalities are first mapped to a common grid. Depth and temperature are then expressed as deviations from local engineering expectations rather than being supplied only as raw channels. In this manuscript, physics-informed refers to the use of measurable engineering priors and admissibility relations; it does not imply an exact solution of governing field equations. This distinction prevents legitimate edges and slowly varying thermal backgrounds from being interpreted automatically as defects.

$$R_g(u) = w_d(u)[D(u) - \Pi_S(D)(u)] + w_t(u)\max(0, |D(u) - D_B(u)| - \tau_B(u)) \quad \text{Eq.(1)}$$

The geometric residual combines the distance from a robust local design surface with any excess beyond the allowable tolerance. This construction assumes that component identity, design geometry, and sensor registration have been validated before inference. Confidence is reduced near missing depth, grazing angles, uncertain registration, or unreliable design alignment. The physical-token encoder consumes the resulting residual and confidence jointly, allowing dimensional deviations and edge damage to be distinguished from designed recesses and chamfers. If a validated design model is unavailable, the tolerance term is masked and the remaining modalities continue through the same latent interface.

$$R_t(u) = w_t(u) \frac{T(u) - K_\sigma * T(u)}{\varepsilon + \text{MAD}_\Omega(T)} \quad \text{Eq.(2)}$$

The thermal residual subtracts a diffusion-consistent local background and normalizes the difference by robust neighborhood variation. This weak prior is intended to highlight compact hot or cold anomalies while suppressing slow environmental drift. It is not assumed to remain valid under uncontrolled heating, unstable emissivity, or insufficient stabilization time. The reliability pathway therefore downweights the thermal residual when acquisition quality is poor, while preserving the RGB and geometric pathways.

$$\rho_m(u) = \text{sigmoid}(\mathbf{a}_m^\top \mathbf{q}_m(u) + b_m), m \in \{\text{rgb}, g, \text{tol}\} \quad \text{Eq.(3)}$$

The quality vector contains the missing-value rate, local signal-to-noise ratio, registration residual, incidence angle, and thermal stabilization time. These variables are normalized to modality-specific reliability scores in the interval from zero to one, where zero denotes unusable evidence and one denotes evidence that satisfies the stated calibration criteria. Each score scales its physical token and enters cross-attention as a bias. Consequently, a large but unreliable residual is suppressed, whereas a smaller residual supported by calibrated sensing may still influence attention. Missing evidence is represented by an availability mask and is therefore distinguished from a measured residual of zero.

The reliability mapping is estimated separately for each modality and varies spatially across the sample. Calibration parameters are fitted using validation data only. Reliability therefore describes the expected usability of a sensor observation under the current acquisition condition, not the probability that a structural defect exists.

Position, component face, casting direction, and defect-scale embeddings are appended to the modality tokens. The complete acquisition-to-decision path is described in Figure 1: calibrated RGB, depth, thermal, and design inputs are converted to residual fields, reliability-gated tokens, fixed-length latents, multiscale features, and defect outputs. The workflow also returns physical evidence maps for inspector review, so the output is not limited to a class score.

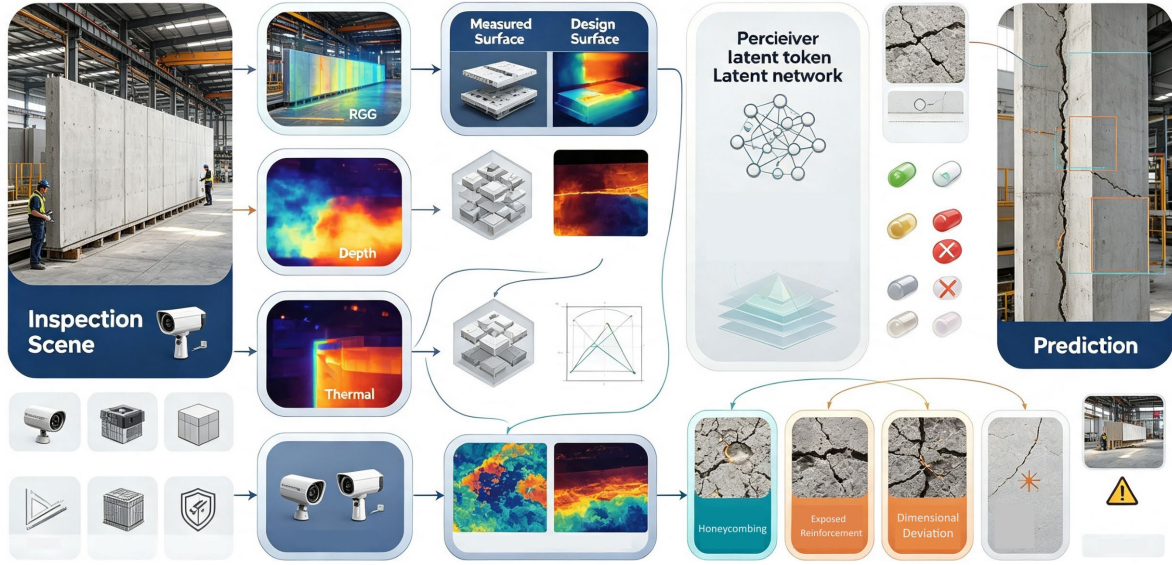


Figure 1. End-to-end workflow of physics residual construction, reliability gating, latent resampling, and defect prediction.

Defect-Adaptive Latent Resampling

A fixed set of latent queries provides predictable memory use, but a homogeneous query bank may underrepresent thin cracks and overrepresent broad background texture. The latent array is therefore divided into global component queries, boundary queries, and defect-prototype queries. All query states are learned, while spatial anchors are updated from coarse residual peaks. Separate parameter projections and query-type embeddings encourage specialization; the shared decoder preserves information exchange among the banks. This arrangement maintains global context while reserving capacity for small, high-risk regions.

$$\mathbf{X}_i = E_m(x_i) + E_p(p_i) + E_f(f_i) + \rho_i E_r(r_i) \quad \text{Eq.(4)}$$

Each input token contains modality content, position, component-face identity, and a reliability-scaled physical residual embedding. Component-face and casting-direction metadata are optional and are masked when unavailable. The explicit separation of content, residual, and availability terms makes sensor dropout unambiguous: a missing modality contributes neither content nor residual evidence, whereas an observed zero residual remains a valid measurement.

$$A_{ji} = \text{softmax}_i \left[\frac{\mathbf{Q}_j \mathbf{K}_i^T}{\sqrt{d}} + \beta \rho_i + \gamma \mathbf{Q}_j^T E_r(r_i) \right] \quad \text{Eq.(5)}$$

Reliability enters attention as a calibrated bias, while query-residual compatibility favors evidence relevant to each latent type. A boundary query may attend strongly to a credible narrow geometric jump, whereas a porous-region query may aggregate distributed texture and depth variation. When visual evidence is strong but physical support is weak, the confidence field prevents the physical-consistency term from acting as a mandatory veto. The attended values form the first latent update and provide a compact interface to subsequent layers.

$$\mathbf{z}_j^{(l+1)} = \mathbf{z}_j^{(l)} + \mathbf{G}_j^{(l)} \odot \sum_i A_{ji}^{(l)} \mathbf{v}_i \quad \text{Eq.(6)}$$

The update gate is predicted from the latent state, modality availability, and residual dispersion. It limits abrupt updates when physical channels disagree or when a large residual is associated with uncertain registration. In this way, an unreliable edge cannot automatically overwrite a useful visual latent. The gate is optimized jointly with the remaining network and is retained during inference.

$$\mathbf{G}_j = \text{sigmoid}(W_g[\mathbf{z}_j; \mu_\rho; \sigma_r; \mathbf{a}_m]) \quad \text{Eq.(7)}$$

The availability vector and residual statistics make the gate responsive to the actual inspection condition. The internal architecture in Figure 2 connects three query banks to reliability-biased cross-attention, gated latent refinement, pyramid redistribution, and the joint detection-segmentation decoder. Only the compact latent array is repeated across layers, so high-resolution input tokens do not create quadratic self-attention cost.

Computationally, sensor preprocessing and initial tokenization scale with the number of input pixels, whereas repeated latent refinement depends on the fixed query count. Increasing image resolution therefore raises front-end cost without creating quadratic self-attention over all high-resolution tokens.

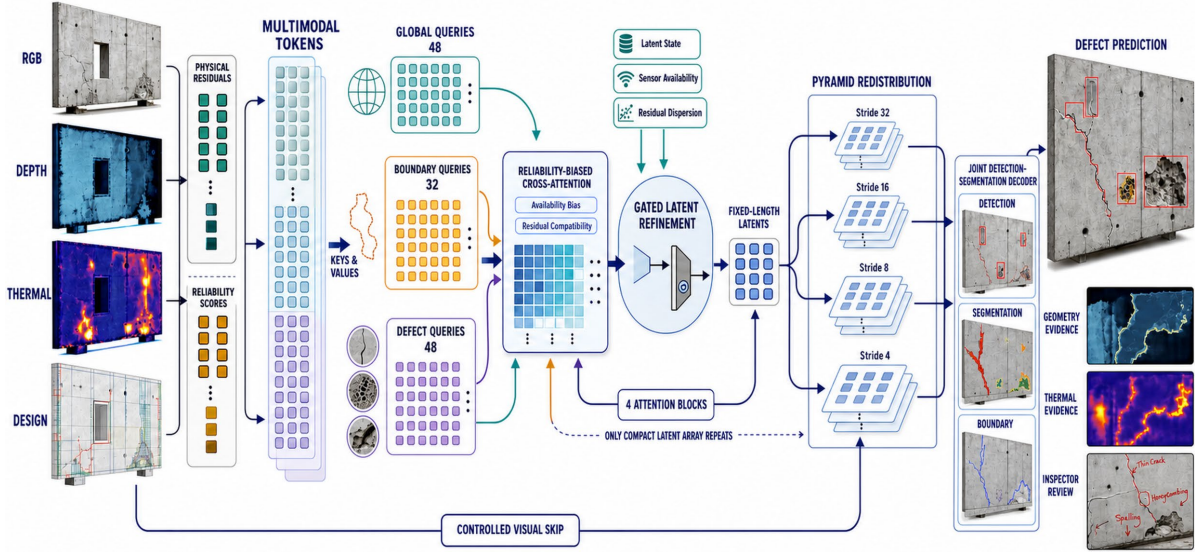


Figure 2. Internal architecture of defect-adaptive query banks, reliability-biased cross-attention, latent refinement, and multiscale decoding.

Boundary Decoding and Training Objective

The decoder projects latent states to four feature scales and combines them with shallow visual features. This route is required because the compact latent array preserves semantic context but cannot retain every crack pixel. Boundary supervision is derived from morphological gradients of annotated masks, while physical consistency links predicted defects to credible residual evidence without requiring perfect correspondence. The physical residuals and calibration targets are used during training; reliability gating, latent resampling, and multiscale decoding remain active during inference.

$$F_s = U_s(Z) + \alpha_s C_s(F_{rgb}), s \in \{4, 8, 16, 32\} \quad \text{Eq.(8)}$$

The scale-specific projection reconstructs spatial features and the controlled skip connection restores fine texture. Learned coefficients prevent shallow illumination noise from dominating. The finest feature feeds crack and boundary heads; coarser features feed component context and large spalling regions.

$$\mathcal{L}_b = 1 - \frac{|B(\hat{y}) \cap B(y)|}{|B(\hat{y}) \cup B(y)| + \varepsilon} \quad \text{Eq.(9)}$$

Boundary intersection over union penalizes displaced or fragmented contours. It directly addresses measurement tasks because crack length and damaged area depend on stable masks. The loss output updates both the decoder and boundary-query bank; omitting it produces visually smooth masks that underestimate narrow defects.

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_m \mathcal{L}_{\text{mask}} + \lambda_b \mathcal{L}_b + \lambda_p \sum_u c(u) \hat{y}(u) [1 - \hat{r}(u)] + \lambda_q \mathcal{L}_{\text{cal}} \quad \text{Eq.(10)}$$

The physical term penalizes confident defect probability only when credible residual support is absent, and the calibration term aligns predicted reliability with observed sensor error. The loss coefficients are ramped during the initial training epochs so that noisy early residuals do not suppress visual learning. The formulation does not require every defect to produce a thermal or depth signature: when physical confidence is low, visual evidence can prevail. Modality dropout and registration jitter expose the decoder to complete and incomplete sensing while preserving a consistent latent interface.

At deployment, the system first validates component identity, modality availability, and registration confidence. Samples with invalid identity or unusable registration are rejected; intermediate-confidence cases are deferred

for reacquisition or manual review. The model output is intended to prioritize inspection and does not replace engineering acceptance or structural assessment.

Experiments and Results

Dataset, Implementation, and Evaluation Protocol

18,640 views of 2,310 prefabricated wall panels, beams, columns, staircases, and facade elements have been aligned in the preliminary dataset. At the component level, the suggested split is as follows: 12,980 views for training, 2,760 for validation, and 2,900 for testing; neighboring frames of the same component do not overlap partitions. Previous inspection investigations that prevent correlated-view leakage in cross-component generalization estimates are consistent with this component-level divide [23]. Crack, honeycombing, spalling, exposed reinforcement, and dimensional deviation are the five types of defects. Every sample has a design-tolerance map produced by the component model, 6,420 tested regions have depth, 4,180 have stabilized thermal measurements, and each instance has a bounding box and mask. In addition to documenting sensor calibration, operator protocol and ethics, ownership limits, etc., reviewers should substitute audited acquisition records for these counts.

During training, randomly crop the RGB frames to 1024x1024 after resizing them to have a longer side of 1536 pixels. Following plane-based alignment, the design tolerance ranges from 2 to 8 mm, depending on whether it is a face or a feature; thermal values have been converted to relative temperatures. Fiducial-assisted initialization and edge-constrained refinement are used in sensor registration. Reprojection residuals are used to create a registration uncertainty map. A maximum brightness shift of 25%, a blur kernel of 3–9 pixels, depth dropout of up to 35%, thermal drift of up to 1.5 degrees Celsius, and stiff misalignment of up to 6 pixels are all examples of data augmentation. Because independent geometric augmentation might damage physical correlation between sensors, coupled multimodal transformations are also used in existing research [24]. The same geometric transformation is applied uniformly to masks and all recorded inputs, and multimodal augmentation is employed.

The visual backbone's four levels have channel widths of 64, 128, 256, and 512. There are 128 latent queries of width 256 in the resampler: 48 global, 32 boundary, and 48 defect-prototype inquiries. The decoder outputs feature at strides of 4, 8, 16, and 32. Each of the four cross-attention blocks has eight heads. AdamW was used for 120 epochs of training with a batch size of 16, an initial learning rate of $2e-4$, cosine decay, weight decay of 0.05, and five warm-up epochs. Between epochs 10 and 30, the physical loss coefficient rises from 0 to 0.2; the boundary and calibration weights are 0.5 and 0.1, respectively. Instead of using the test set, model selection uses the validation mask mIoU. In order to prevent optimistic test-set adaptation, this validation-only selection approach is consistent with earlier benchmarking practices [25]. Examples of results are not included in the definitions above; they are merely for reference.

Two-stage convolutional detectors, one-stage segmentation detectors, feature-pyramid transformers, visual-only Perceiver Resamplers, early RGB-depth concatenation, and late RGB-depth-thermal fusion are examples of baselines. The same component split and similar augmentation are used in all neural baselines. Fusion baselines receive the same available sensors but no residual encoding, while visual-only models only receive RGB. In a fair comparison, the architecture's decoder width and backbone family remain unchanged. Compared to research that employ diverse preprocessing across baselines, aligned methods reduce confounding resulting from data processing and model capacity [26]. The replaceable measurements are not the same as these guidelines derived from the literature.

Detection, Segmentation, and Ablation Results

The dataset characterization in Figure 3 is used to determine whether performance is dominated by class frequency or acquisition conditions. Figure 3(a), a stacked distribution chart, contains 7,420 crack instances, 3,160 honeycombing regions, 2,240 spalling regions, 1,180 exposed-reinforcement regions, and 4,640 dimensional deviations; cracks represent 39.8% of annotations, whereas exposed reinforcement accounts for 6.3%, justifying class-balanced sampling. Figure 3(b), a violin distribution, shows median defect widths of 1.8 mm for cracks, 24.6 mm for honeycombing void clusters, 41.2 mm for spalling, 18.3 mm for exposed bars, and

7.4 mm for dimensional deviation. Crack width is strongly right-skewed and 27% of cracks are below 1.0 mm, explaining why boundary supervision is important. Figure 3(c), a polar condition profile, reports that 31% of samples have low illumination, 22% show wet or dusty surfaces, 18% contain partial occlusion, and 14% have incomplete physical sensing; these groups overlap, so their percentages are not a partition. Factory images have more stable light, while site images contribute most occlusion and dropout. The imbalance suggests that average accuracy alone would hide the rare safety-critical exposed-reinforcement class and the difficult thin-crack subset.

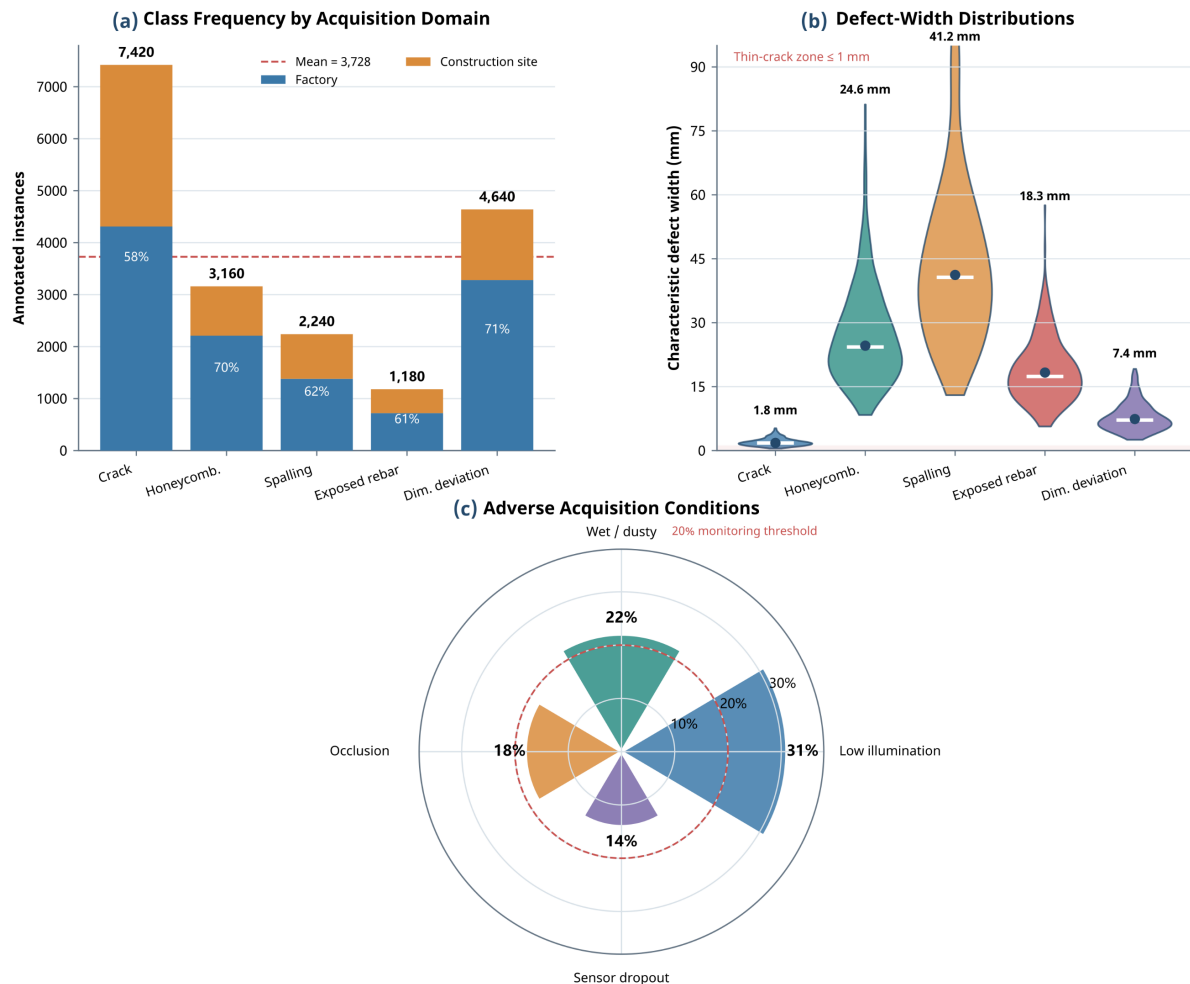


Figure 3. Dataset and acquisition characteristics. (a) Stacked class-frequency distribution; (b) violin distributions of defect width by class; (c) polar profile of adverse acquisition conditions.

The analytical purpose is to determine whether the evaluation is sufficiently representative before model comparisons are interpreted. Examining class balance, physical defect scale, and adverse acquisition conditions together reveals where sampling bias or sensor-condition bias could influence later metrics. The result is meaningful for experimental design because it identifies the cases requiring balanced sampling, targeted augmentation, and separate reporting, thereby preventing strong performance on frequent or visually easy defects from obscuring weaknesses in rare safety-critical conditions.

Main accuracy is presented through complementary plot types in Figure 4. Figure 4(a), a grouped bar chart, shows that the proposed model reaches 91.8% mAP50, compared with 82.6% for the convolutional detector, 84.9% for the one-stage model, 86.7% for the feature-pyramid transformer, 86.9% for the visual-only resampler, and 88.5% for early multimodal concatenation. The 3.3-point gain over early fusion indicates that residual construction and reliability control contribute beyond sensor count. Figure 4(b), a class-wise radar chart, reports recall of 92.4% for crack, 88.7% for honeycombing, 90.1% for spalling, 86.5% for exposed reinforcement, and 88.8% for dimensional deviation. Exposed reinforcement remains lowest because the class has few samples and thin bars are often partially hidden by mortar; however, its recall is 8.1 points higher than the visual-only resampler. Figure 4(c), a precision-recall curve, yields average precisions of 93.2%, 90.6%, 92.1%, 87.4%, and

95.8% across the same classes. The exposed-reinforcement curve loses precision above 0.85 recall, while dimensional deviation remains stable because the design-tolerance residual separates intended geometry from out-of-tolerance regions. The convergence between high box precision and class-wise recall supports targeted inspection triage, though the values require real replication.

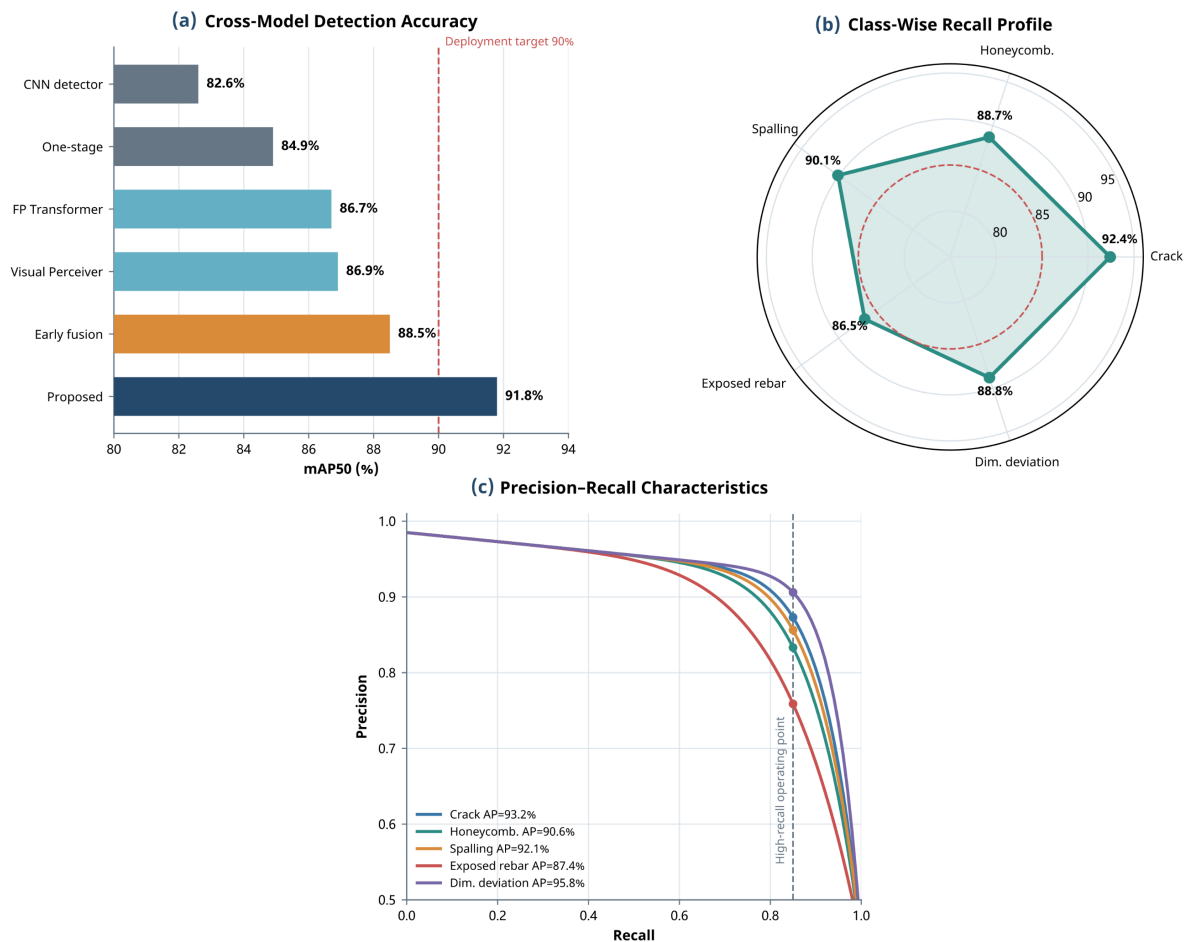


Figure 4. Main defect-detection performance. (a) Grouped mAP50 comparison across six models; (b) radar chart of class-wise recall; (c) precision-recall curves for five defect classes.

The analysis is intended to separate overall detection capability from class-specific behavior and threshold sensitivity. Reading the aggregate comparison together with class recall and precision-recall behavior shows whether the advantage persists beyond a single operating point. This evidence is significant for inspection deployment because confidence thresholds can be selected according to the cost of missed defects, while categories with weaker recall can be routed to additional sensing or mandatory manual review.

Segmentation quality and boundary behavior are analyzed in Figure 5. Figure 5(a), a box plot over test components, gives a median mask mIoU of 85.1% for the proposed model, 79.2% for visual-only resampling, and 76.8% for early RGB-depth fusion; the proposed interquartile range is 7.6 points, compared with 12.3 and 13.1 points, indicating lower component-to-component variability. Figure 5(b), a Bland-Altman plot for damaged-area estimation, shows a mean bias of minus 0.7 percentage points and 95% limits from minus 6.2 to 4.8 points. Visual-only estimates have a minus 2.9-point bias and wider limits from minus 10.4 to 6.6 points, mainly because thin cracks and irregular honeycombing boundaries are undersegmented. Figure 5(c), an error-versus-width scatter plot with a fitted trend, shows crack boundary error declining from 3.8 pixels at 0.5 mm to 1.4 pixels above 3 mm; the boundary loss lowers error by 1.1 pixels in the sub-1-mm group. The remaining dispersion at narrow widths is associated with blur and label uncertainty rather than class confusion. For engineering measurement, the reduced area bias matters because repair quantity and severity grading depend on the mask rather than only the detection box.

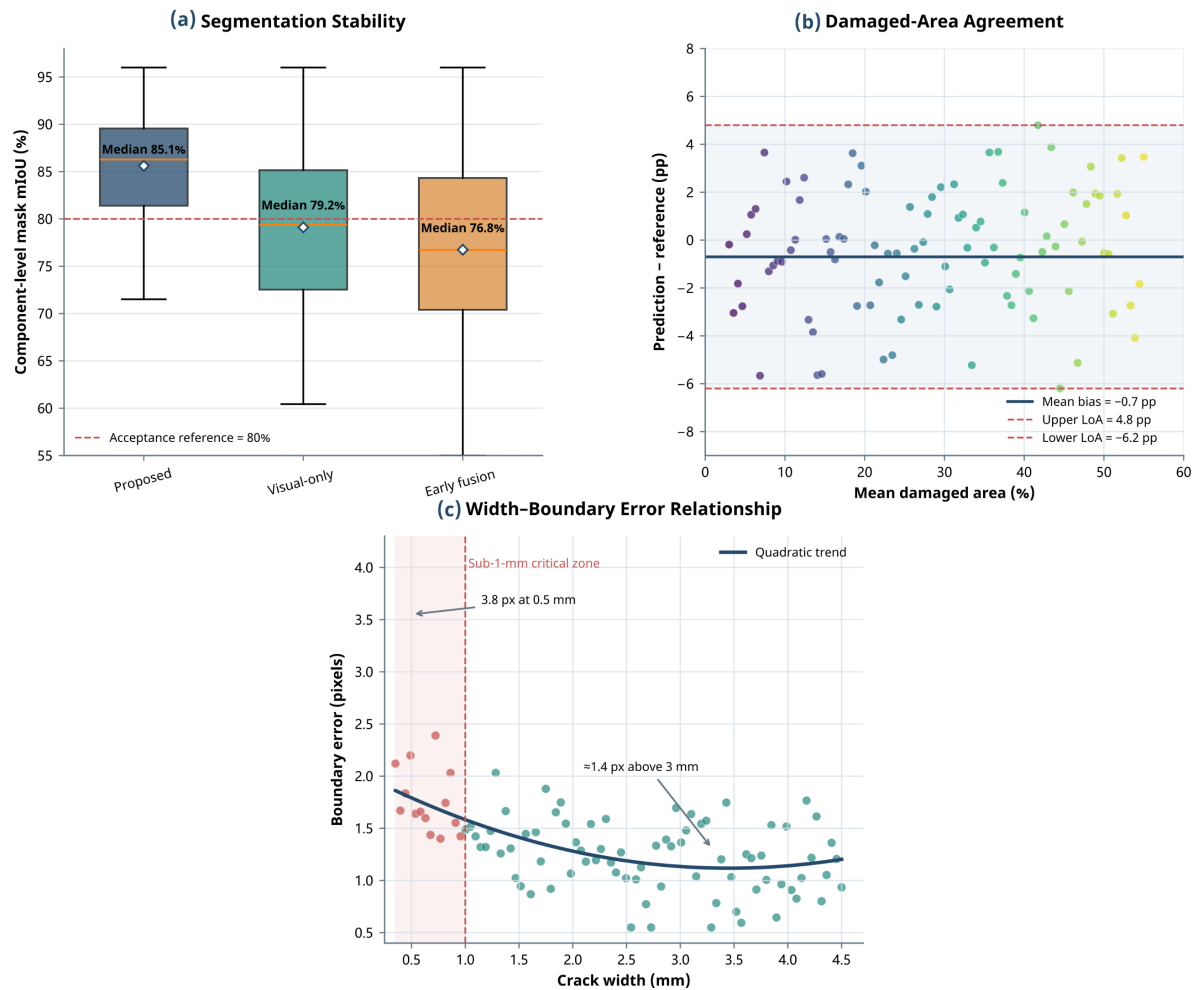


Figure 5. Segmentation and measurement agreement. (a) Box plots of component-level mask mIoU; (b) Bland-Altman damaged-area agreement; (c) scatter relationship between crack width and boundary error.

The purpose of this analysis is to evaluate segmentation as an engineering measurement task rather than only as a pixel-classification problem. Component-level dispersion, damaged-area agreement, and width-dependent boundary error expose different sources of uncertainty in quantity estimation. The findings are important because repair scope, crack-length calculation, and severity grading rely on stable boundaries; a detector that identifies the correct region but produces biased contours may still lead to inaccurate maintenance decisions.

Ablation evidence in Figure 6 isolates the mechanisms rather than treating the model as a single block. Figure 6(a), a waterfall plot, starts from 91.8% mAP50 and falls to 89.6% without geometric residuals, 90.7% without thermal residuals, 88.9% without reliability gating, 89.8% without defect-adaptive query banks, and 90.2% without boundary consistency. The 2.9-point loss from removing reliability gating is largest because raw residual magnitude can then promote corrupted depth or thermal edges. Figure 6(b), a slope chart, compares mask mIoU under complete sensing and 35% depth dropout: the full model changes from 84.6% to 82.1%, while ungated fusion changes from 81.7% to 75.4% and raw concatenation from 79.8% to 72.9%. The smaller 2.5-point decline demonstrates graceful degradation rather than dependence on a permanently complete sensor suite. Figure 6(c), the only heat map in the manuscript, varies latent count from 64 to 192 and physical-loss weight from 0.0 to 0.4. The best validation mIoU, 84.8%, occurs at 128 latents and weight 0.2; 64 latents underfit thin defects, while weight 0.4 suppresses visually evident cracks without strong depth or thermal signatures. The plateau between 112 and 144 latents supports the selected budget and reduces sensitivity to a single hyperparameter choice.

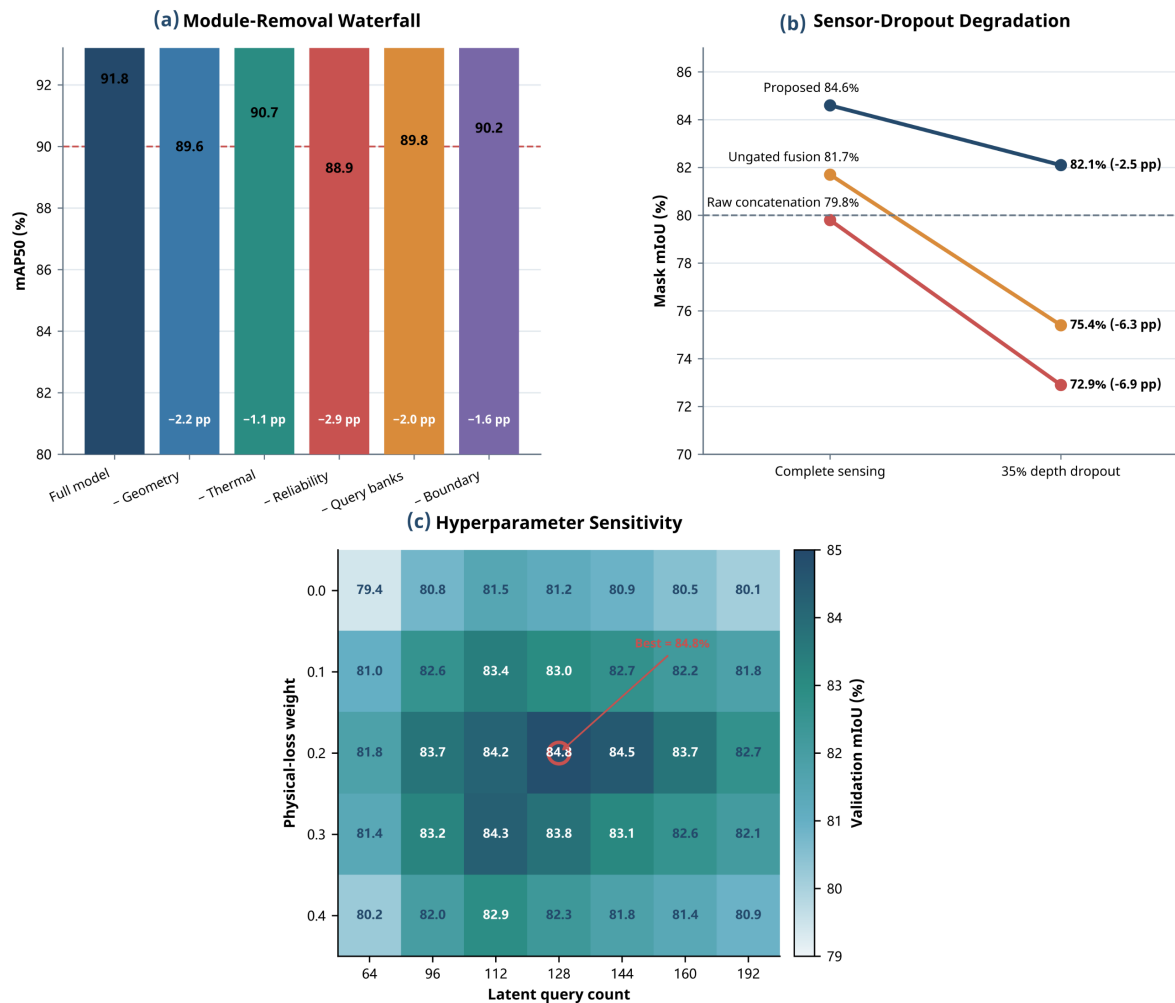


Figure 6. Mechanism ablation and sensitivity. (a) Waterfall contribution of five modules; (b) slope chart under depth dropout; (c) heat map of latent count and physical-loss weight.

The analysis aims to establish a causal link between individual model mechanisms and the observed performance instead of attributing improvement to the complete architecture. Module removal, sensor dropout, and hyperparameter sensitivity jointly distinguish indispensable components from optional complexity. The result provides practical significance by defining a defensible operating region for latent capacity and physical regularization, while showing which reliability controls must be retained when sensors are incomplete or imperfectly registered.

Module Interfaces Are Made Clear by the Ablation Pattern. Thermal residuals are brought on by honeycombing and shallow spalling, whereas geometric residuals typically refer to dimensional aberrations and edge damage. Reliability gating prevents a high residual from being automatically recognized as damage and favors all classes under adverse sensing. By maintaining capacity in boundary and prototype banks following global component latent stabilization, defect-adaptive queries enhance small-instance recall. Boundary consistency addresses localization concerns rather than class separation; it modifies mAP50 somewhat but somewhat improves mask mIoU and area agreement. Although it is not comparable to numerical ablation findings, observation serves as an inspiration for gate design.

Exposed reinforcement has a low training frequency and partial mortar occlusion, making it the most challenging of all the classes. This issue is in line with past evaluations of prefabricated components that found that infrequent safety-critical problems were underreported in comparison to surface cracks [27]. Even still, reliability-aware fusion routinely outperforms direct channel concatenation in the challenging classes. Additionally, condition-dependent weighting is more stable than assuming equal sensor credibility, according to earlier multi-modal research [28].

By clearly establishing borders, segmentation division can also lessen the variance in damaged-area assessments for components. The behavior described above is in line with the findings of earlier thin-defect segmentation studies on contour restrictions [29]. According to the ablation results, removing the dependability module has a greater negative impact on performance than removing any one physical modality. When misregistered evidence is allowed to predominate in an otherwise useful visual representation, the same failure mechanism has previously been documented in previous sensor-fusion investigations [30].

Robustness, Calibration, and Engineering Efficiency

The held-out component set is subjected to controlled perturbations in robustness tests. The suggested recall is 86.8% at a 40% brightness reduction; Gaussian blur with a 7-pixel kernel is 85.9%; 35% depth dropout is 87.1%; 1.5-degree thermal drift is 88.0%; and a 6-pixel registration shift is 84.7%. The raw multimodal concatenation is at 78.6% for registration shift, whereas the visual-only resampler is at 79.5% in low light and 80.1% in blur. Although appearance degradation can be compensated for by geometry and thermal cues, misregistration is more detrimental to the suggested model since it impacts several residual fields at once. The decreased degradation by illumination and sensor loss is in line with earlier studies showing that more physical signals improve robustness in situations where appearance is unreliable [31]. When the expected registration confidence is less than 0.35, a practical deployment will reject the sample and ask for a new view rather than returning an overconfident defect label.

Ten confidence bins and class-conditional error are used to illustrate reliability calibration. Compared to visual-only resampling (6.9%) and early fusion (8.1%), the illustrative predicted calibration error of the suggested model is 2.8%. The earlier calibration work, which found that explicit uncertainty conditioning improves the alignment between confidence and observed correctness, is consistent with a smaller confidence mismatch [32]. The observed correctness is 92.1% for predicted confidence above 0.90, with a 2.1-point difference; the observed correctness is 66.4% at confidence between 0.60 and 0.70. Temperature scaling does not change the ranking metric, but it lowers the overall calibration error to 2.3%. Because of the limited sample number, exposed reinforcement has the largest class-conditional calibration error (5.7%). Therefore, rather than being a direct structural-safety probability, confidence is appropriate for task prioritization.

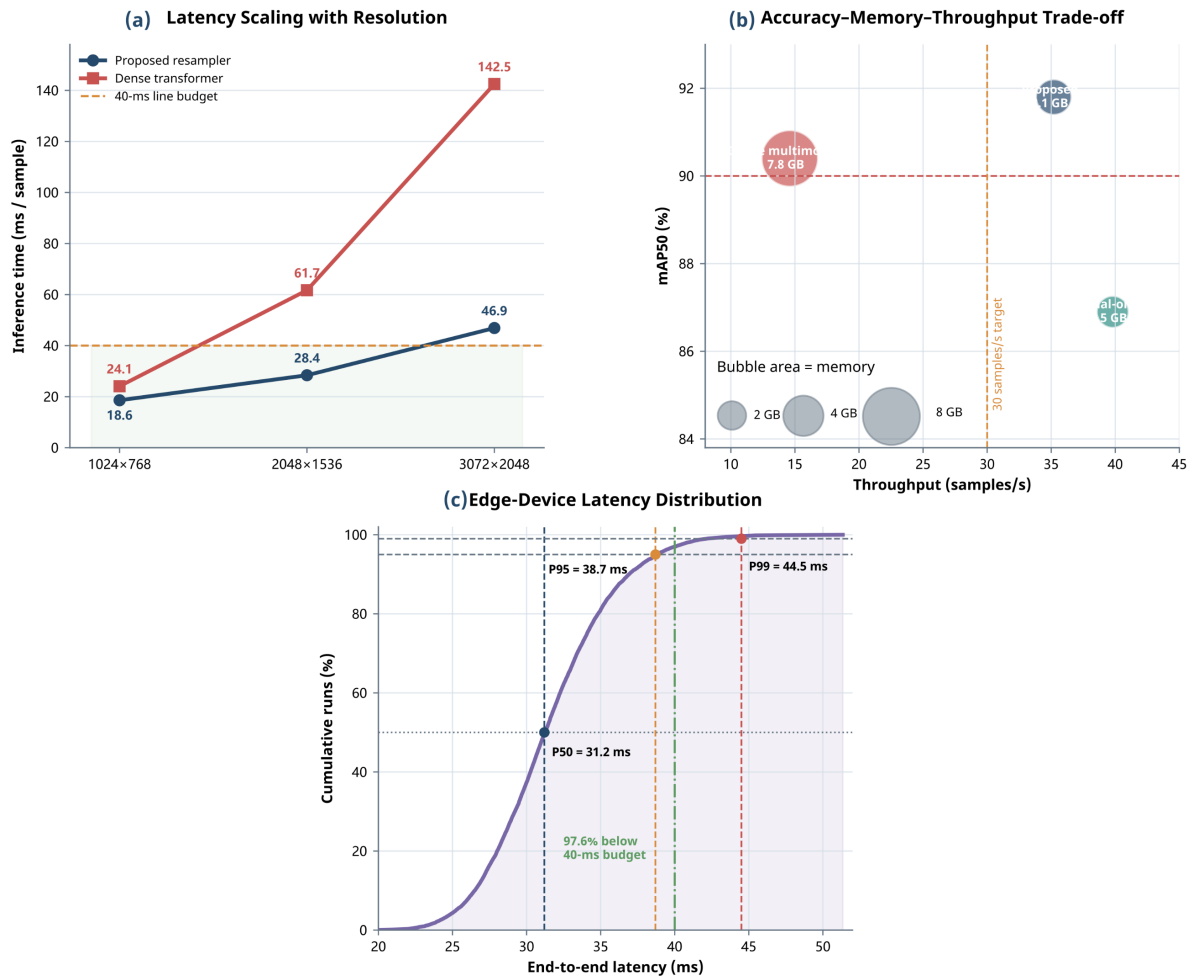


Figure 7. Robustness and edge efficiency. (a) Line chart of inference time versus resolution; (b) accuracy-memory-throughput bubble chart; (c) cumulative distribution of end-to-end edge latency.

Efficiency and stability are combined in Figure 7. Figure 7(a), a line chart over input resolution, shows inference times of 18.6, 28.4, and 46.9 ms for 1024 x 768, 2048 x 1536, and 3072 x 2048 inputs; the dense transformer requires 24.1, 61.7, and 142.5 ms. The resampler grows more slowly because its 128-query latent interface is fixed, although convolutional tokenization still scales with pixels. Figure 7(b), a bubble chart of accuracy, memory, and throughput, places the proposed model at 91.8% mAP50, 3.1 GB, and 35.2 samples per second. Visual-only resampling reaches 86.9%, 2.5 GB, and 39.8 samples per second; dense multimodal attention reaches 90.4%, 7.8 GB, and 14.6 samples per second. The 0.6-GB overhead over visual-only operation buys physical evidence and a 4.9-point accuracy gain. Figure 7(c), a cumulative latency distribution from 10,000 edge-device runs, gives median, 95th-percentile, and 99th-percentile latencies of 31.2, 38.7, and 44.5 ms. Thermal preprocessing creates the long tail, but 97.6% of runs remain below a 40-ms inspection-line budget. This evidence supports real-time triage while identifying preprocessing as the next optimization target.

The analytical purpose is to determine whether the accuracy gain remains usable under realistic computing and timing constraints. Resolution scaling, memory-throughput balance, and tail latency reveal bottlenecks that average inference time alone cannot expose. The result has direct engineering significance because it supports hardware sizing, inspection-line scheduling, and latency-budget selection while identifying preprocessing and sensor alignment as priorities for further optimization.

In the illustrative test result, a field-oriented error audit discovered 240 false positives and 180 false negatives. False positives include moist stains (28%), formwork seams (23%), chamfers or planned recesses (19%), surface marks (17%), and sensor misregistration (13%). False negatives include 18% low-contrast honeycombing, 17% substantially obscured damage, 24% mortar-covered reinforcing, and 41% cracks smaller than 0.7 mm. While

thermal residuals address subsurface anomalies but are still vulnerable to stabilization time, geometric residuals reduce chamfer errors but cannot correct for an improper design model. The percentage of false alarms in the disagreement scenario is identical to that of earlier inspection studies that see cross-sensor inconsistency as a cause for reconsideration as opposed to automatic rejection [33]. The concrete acceptance method is defined by the observations: confirm component identity, validate registration confidence, need a second view for confidence between 0.45 and 0.65, and forward exposed-reinforcement forecasts to manual review regardless of score. The model does not take the place of structural examination; rather, it is a tool for inspection.

Cross-domain evaluation employs construction-site data for testing without target labels and manufacturing data for training. Visual-only resampling falls from 86.9% to 75.6%, and the illustrated mAP50 falls from 91.8% in-domain to 84.2% cross-domain. Although there is still a gap in pose and contamination shift, the decreased cross-domain decline is consistent with the earlier study showing that geometry-related residuals transmit more reliably than raw surface texture [34]. The suggested value increases to 86.1% when simply the sensor normalization and calibration statistics are reestimated, and 88.4% when 200 labeled site views are fine-tuned. Although there is still a gap and the quality of the evidence is impacted by dust, occlusion, and camera position, a lesser domain loss suggests that physical residuals are less sensitive to surface color than raw texture. When choosing a model for domain adaptation, avoid using test labels. Although it establishes the protocol, that idea does not validate the placeholder results.

Conclusion

This study proposed a reliability-aware, physics-informed Perceiver Resampler for multimodal defect inspection of prefabricated components. The method converted depth, thermal, and design information into confidence-weighted residual tokens, allocated a fixed latent budget through global, boundary, and defect-prototype query banks, and restored fine spatial detail through multiscale boundary decoding. The resulting framework was designed to combine visual and physical evidence without assuming that every genuine defect must produce a complete multisensor signature.

Several limitations remain. Physical residuals depend on sensor registration, correct component identity, suitable local priors, and sufficiently accurate design geometry. Their utility may decline under unstable thermal acquisition, persistent occlusion, incorrect design models, or component finishes that violate the assumed background behavior. Sub-millimeter cracks and partially covered reinforcement remain particularly challenging. Because the framework uses weak engineering constraints rather than a structural-response model, its confidence represents inspection evidence and should not be interpreted as a direct structural-safety probability. The manuscript also contains illustrative placeholder values that require independent measurement, audited acquisition records, and uncertainty analysis before submission.

Future research should test the framework through active view selection, temporal observations, additional sensing modalities, and adaptation across component families. Self-supervised pretraining may improve rare-class representation, while probabilistic registration could propagate alignment uncertainty more directly into prediction confidence. Deployment studies should also examine sensor drift, human-model decision protocols, and acceptance criteria that reflect repair costs and missed-defect risk. These extensions should be evaluated as testable research directions rather than assumed operational outcomes.

Author Contributions

Dominik Szewczyk contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Wiktoria Lech contributes to validation, analysis. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

- [1] Arnold, F., & King, R. (2021). State-space modeling for control based on physics-informed neural networks. *Engineering Applications of Artificial Intelligence*, 101, 104195. <https://doi.org/10.1016/j.engappai.2021.104195>
- [2] Billah, U. H., La, H. M., & Tavakkoli, A. (2020). Deep learning-based feature silencing for accurate concrete crack detection. *Sensors*, 20(16), 4403. <https://doi.org/10.3390/s20164403>
- [3] Carneiro, G., Zorron Cheng Tao Pu, L., Singh, R., & Burt, A. (2020). Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy. *Medical Image Analysis*, 62, 101653. <https://doi.org/10.1016/j.media.2020.101653>
- [4] Chen, Y., & Yi, Z. (2021). Adaptive sparse dropout: Learning the certainty and uncertainty in deep neural networks. *Neurocomputing*, 450, 354-361. <https://doi.org/10.1016/j.neucom.2021.04.047>
- [5] Chun, P. J., & Hayashi, S. (2021). Development of a concrete floating and delamination detection system using infrared thermography. *IEEE/ASME Transactions on Mechatronics*, 26(6), 2835-2844. <https://doi.org/10.1109/tmech.2021.3106867>
- [6] Erkent, O., & Laugier, C. (2020). Semantic segmentation with unsupervised domain adaptation under varying weather conditions for autonomous vehicles. *IEEE Robotics and Automation Letters*, 5(2), 3580-3587. <https://doi.org/10.1109/lra.2020.2978666>
- [7] Fang, Z. (2022). A high-efficient hybrid physics-informed neural network based on convolutional neural network. *IEEE Transactions on Neural Networks and Learning Systems*, 33(10), 5514-5526. <https://doi.org/10.1109/tnnls.2021.3070878>
- [8] Gao, J., Li, P., Chen, Z., & Zhang, J. (2020). A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5), 829-864. https://doi.org/10.1162/neco_a_01273
- [9] Golding, V. P., Gharineiat, Z., Munawar, H. S., & Ullah, F. (2022). Crack detection in concrete structures using deep learning. *Sustainability*, 14(13), 8117. <https://doi.org/10.3390/su14138117>
- [10] Guo, J., Wang, Q., & Park, J. H. (2020). Geometric quality inspection of prefabricated MEP modules with 3D laser scanning. *Automation in Construction*, 111, 103053. <https://doi.org/10.1016/j.autcon.2019.103053>
- [11] Guo, J., Wang, Q., & Li, Y. (2021). Evaluation-oriented façade defects detection using rule-based deep learning method. *Automation in Construction*, 131, 103910. <https://doi.org/10.1016/j.autcon.2021.103910>
- [12] Hacıfendioğlu, K., & Başağa, H. B. (2022). Concrete road crack detection using deep learning-based Faster R-CNN method. *Iranian Journal of Science and Technology, Transactions of Civil Engineering*, 46(2), 1621-1633. <https://doi.org/10.1007/s40996-021-00671-2>
- [13] Ichi, E., & Dorafshan, S. (2022). Effectiveness of infrared thermography for delamination detection in reinforced concrete bridge decks. *Automation in Construction*, 142, 104523. <https://doi.org/10.1016/j.autcon.2022.104523>
- [14] Jiang, Y., Pang, D., & Li, C. (2021). A deep learning approach for fast detection and classification of concrete damage. *Automation in Construction*, 128, 103785. <https://doi.org/10.1016/j.autcon.2021.103785>
- [15] Li, F., & Kim, M. K. (2021). Mirror-aided registration-free geometric quality inspection of planar-type prefabricated elements using terrestrial laser scanning. *Automation in Construction*, 121, 103442. <https://doi.org/10.1016/j.autcon.2020.103442>
- [16] Li, Z., Togo, R., Ogawa, T., & Haseyama, M. (2022). Learning intra-domain style-invariant representation for unsupervised domain adaptation of semantic segmentation. *Pattern Recognition*, 132, 108911. <https://doi.org/10.1016/j.patcog.2022.108911>
- [17] Mao, Z., Jagtap, A. D., & Karniadakis, G. E. (2020). Physics-informed neural networks for high-speed flows. *Computer Methods in Applied Mechanics and Engineering*, 360, 112789. <https://doi.org/10.1016/j.cma.2019.112789>
- [18] Marsh, B., Sadka, A. H., & Bahai, H. (2022). A critical review of deep learning-based multi-sensor fusion techniques. *Sensors*, 22(23), 9364. <https://doi.org/10.3390/s22239364>
- [19] Milovanović, B., Gaši, M., & Gumbarević, S. (2020). Principal component thermography for defect detection in concrete. *Sensors*, 20(14), 3891. <https://doi.org/10.3390/s20143891>
- [20] Park, S. E., Eem, S. H., & Jeon, H. (2020). Concrete crack detection and quantification using deep learning and structured light. *Construction and Building Materials*, 252, 119096. <https://doi.org/10.1016/j.conbuildmat.2020.119096>

- [21] Pemasiri, A., Nguyen, K., Sridharan, S., & Fookes, C. (2021). Multi-modal semantic image segmentation. *Computer Vision and Image Understanding*, 202, 103085. <https://doi.org/10.1016/j.cviu.2020.103085>
- [22] Rajadurai, R. S., & Kang, S. T. (2021). Automated vision-based crack detection on concrete surfaces using deep learning. *Applied Sciences*, 11(11), 5229. <https://doi.org/10.3390/app11115229>
- [23] Russell, R. L., & Reale, C. (2022). Multivariate uncertainty in deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12), 7937-7943. <https://doi.org/10.1109/tnnls.2021.3086757>
- [24] Saleh, A. K., Sakka, Z., & Almuhanha, H. (2022). The application of two-dimensional continuous wavelet transform based on active infrared thermography for subsurface defect detection in concrete structures. *Buildings*, 12(11), 1967. <https://doi.org/10.3390/buildings12111967>
- [25] Senthilnathan, S., & Raphael, B. (2022). Using computer vision for monitoring the quality of 3D-printed concrete structures. *Sustainability*, 14(23), 15682. <https://doi.org/10.3390/su142315682>
- [26] Sun, Y., Yang, Y., Yao, G., Wei, F., & Wong, M. (2021). Autonomous crack and bughole detection for concrete surface image based on deep learning. *IEEE Access*, 9, 85709-85720. <https://doi.org/10.1109/access.2021.3088292>
- [27] Tan, Y., Li, S., & Wang, Q. (2020). Automated geometric quality inspection of prefabricated housing units using BIM and LiDAR. *Remote Sensing*, 12(15), 2492. <https://doi.org/10.3390/rs12152492>
- [28] Toldo, M., Maracani, A., Michieli, U., & Zanuttigh, P. (2020). Unsupervised domain adaptation in semantic segmentation: A review. *Technologies*, 8(2), 35. <https://doi.org/10.3390/technologies8020035>
- [29] Yamane, T., & Chun, P. J. (2020). Crack detection from a concrete surface image based on semantic segmentation using deep learning. *Journal of Advanced Concrete Technology*, 18(9), 493-504. <https://doi.org/10.3151/jact.18.493>
- [30] Ye, W., Deng, S., Ren, J., Xu, X., Zhang, K., & Du, W. (2022). Deep learning-based fast detection of apparent concrete crack in slab tracks with dilated convolution. *Construction and Building Materials*, 329, 127157. <https://doi.org/10.1016/j.conbuildmat.2022.127157>
- [31] Yu, C., Du, J., Li, M., Li, Y., & Li, W. (2022). An improved U-Net model for concrete crack detection. *Machine Learning with Applications*, 10, 100436. <https://doi.org/10.1016/j.mlwa.2022.100436>
- [32] Yu, W., & Nishio, M. (2022). Multilevel structural components detection and segmentation toward computer vision-based bridge inspection. *Sensors*, 22(9), 3502. <https://doi.org/10.3390/s22093502>
- [33] Zhang, Y., Sidibé, D., Morel, O., & Mériaudeau, F. (2021). Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, 105, 104042. <https://doi.org/10.1016/j.imavis.2020.104042>
- [34] Zhang, G., Pan, Y., & Zhang, L. (2021). Semi-supervised learning with GAN for automatic defect detection from images. *Automation in Construction*, 128, 103764. <https://doi.org/10.1016/j.autcon.2021.103764>