Journal of Applied Computational Intelligence

eISSN: 3136-6937

Vol. 4, No. 1, January, 2026

Geometrically Constrained DeepLabv3 Plus for Pantograph Carbon Strip Thickness Estimation

Biljana Todorović[1,*], Mihaila Pavlović[1], Dejana Simić[1]

[1] *Faculty of Computer Science, University of Novi Sad, Novi Sad, 21000, Serbia*

Received 22 November 2025; Revised 12 March 2026; Accepted 14 March 2026; Published 19 March 2026

Abstract

Reliable pantograph carbon strip inspection requires accurate contour localization and a defensible conversion from image coordinates to physical dimensions. This paper proposes a geometrically constrained DeepLabv3+ framework for estimating visible residual thickness in a calibrated observation plane. Boundary and visibility branches distinguish measurable contours from unavailable intervals, while weak ordering and continuity penalties regulate contour relationships without imposing a uniform wear profile. Wear depth is obtained only when a registered unworn reference is available. A numerical benchmark design groups 2400 image-condition records under 120 specimen identifiers, with development and evaluation partitions separated by identifier. The illustrative comparison assigns mean absolute errors of 0.286 mm to the unconstrained baseline and 0.118 mm to the complete configuration, with boundary F1 values of 90.2% and 95.8%, respectively. The corresponding error without the geometric term is 0.173 mm, and the highest-stress example reaches 0.241 mm. These descriptive values demonstrate the proposed evaluation and reporting logic; they do not establish independently measured performance on physical strips or causal improvements from trained model components. Calibration sensitivity, interval undercoverage, and accepted coverage are examined separately because segmentation agreement alone cannot verify dimensional accuracy. The contribution is the integration of contour geometry, visibility-aware reporting, and calibrated dimensional interpretation. Empirical applicability remains to be established through traceable reference measurements, fully specified training and decoding settings, repeated-seed comparisons, and specimen-level validation under contamination, perspective variation, and service damage.

Keywords: Pantograph Carbon Strip, Wear Measurement, Geometric Constraint, Boundary Segmentation, Dimensional Calibration

1. Introduction

Research on pantograph current collection identifies the carbon strip as the consumable contact element linking the overhead conductor to the vehicle power supply [1]. Investigations of strip deterioration distinguish distributed abrasion from localized chipping because these processes produce different residual profiles [2]. Spatially resolved inspection approaches therefore assess local material loss alongside the overall condition of the strip [3]. Controlled optical inspection arrangements combine image acquisition and geometric referencing to recover the exposed material contour at a defined section [4]. The resulting profile supports maintenance assessment by identifying narrow depressions within sections having an acceptable average thickness. Image-based inspection preserves the spatial position of such depressions and supports comparison between successive observations when acquisition geometry is controlled. Measurement planning must consider the orientation of the viewing plane, the distinction between the support datum and the wear boundary, and contamination that may introduce competing edges. A spatially registered dimensional output associates each reported thickness with an observed surface interval and a corresponding datum location. This

* Corresponding author: todor.bija@pf.uns.ac.rs

http://dx.doi.org/10.68406/jaci.2026.vol4iss1n11:145-159

➢ This is an open access article under the CC BY-NC-SA 4.0 license (https://creativecommons.org/licenses/by/4.0/).

© Authors retain all copyrights.

association becomes particularly important near chipped ends, brackets, and deposits, where material identity and surface visibility can differ. The inspection problem consequently extends from recognizing carbon material to establishing which parts of its contour support a calibrated distance measurement. An integrated treatment of segmentation, dimensional calibration, and sample validity provides the basis for investigating this relationship under controlled acquisition conditions.

Traditional contour extraction depends on intensity transitions, local gradients, and geometric fitting. These methods depend on a clear material silhouette; dark deposits and specular support edges can introduce competing responses stronger than the target boundary. Wang et al. [5] used restricted search regions in task-specific contour extraction to reduce this ambiguity. An occluded boundary requires an explicit visibility label because curve fitting across its hidden interval introduces an inferred surface. Learned segmentation uses neighborhood context to distinguish material identity from competing local gradients. Hou et al. [6] used encoder and decoder networks to recover spatial detail from features computed at different resolutions. Multiscale segmentation methods further use context at different receptive fields to handle changes in object extent [7]. For carbon strip inspection, the principal engineering quantity is the distance between two physically defined contours. A small displacement of either curve may have little influence on the overlap of a large mask while materially changing the minimum residual thickness. Boundary-oriented segmentation research addresses this difference between region agreement and contour localization [8]. The methodological challenge is to translate improved contour localization into calibrated dimensional accuracy while preserving the observed wear profile. The dimensional task therefore requires contour evaluation at the same registered locations used for reference measurement.

Geometry can support the interpretation of segmentation when its assumptions are tied to the inspection arrangement. Studies of shape-constrained learning have used structural priors to discourage predictions that violate known spatial relationships [9]. For a suitably rectified carbon strip image, the support datum should remain below the exposed wear contour, and neighboring samples ordinarily belong to the same visible section. These relationships constrain contour ordering while allowing longitudinal variation in residual thickness. Research on robust dimensional vision has also emphasized the need to separate image localization uncertainty from calibration uncertainty [10]. Pixel-scale error and depth displacement contribute dimensional uncertainty independently of segmentation confidence. Similarly, methods for selective prediction provide a framework for withholding outputs when available evidence is inadequate [11]. Local validity decisions preserve measurable intervals when another part of the profile is obscured. Specimen-disjoint assessment is also relevant when repeated observations of the same physical object share texture and damage features [12]. Shared specimen texture and damage features can inflate apparent generalization when partitions overlap at the specimen level. These issues define a research gap between segmentation benchmarks and a complete procedure for geometry-aware carbon strip metrology. The evaluation must also retain acquisition records so that changes in surface condition can be separated from changes in measurement geometry.

To connect the visual appearance of an unevenly worn strip with its remaining material thickness, this study integrates geometrically constrained DeepLabv3+ segmentation with visibility-aware contour extraction and calibration in a common physical reference frame. The methodological contribution is the integration of weak contour ordering, local continuity, visibility gating, and calibrated dimensional reporting. The constituent encoder, decoder, boundary supervision, and planar mapping are established elements; novelty is claimed for their task-specific integration rather than for a new segmentation backbone. Geometric consistency regularizes contour relationships while retaining the uneven shape of the exposed surface. It also distinguishes residual thickness from material loss: wear depth requires a corresponding unworn reference profile, whereas a single calibrated image can support only the remaining visible thickness under the stated planar assumptions. The numerical demonstration examines the relationship between boundary localization, calibration sensitivity, and dimensional error across a defined evaluation benchmark. Its purpose is to examine dimensional sensitivity, evaluation consistency, and reporting logic before empirical validation. Section 2 reviews visual inspection, boundary segmentation, and geometry-aware measurement. Section 3 specifies the acquisition assumptions, network objective, and conversion from contours to dimensional outputs. Section 4 presents the evaluation design, comparative results, component analysis, and failure behavior, with numerical observations linked to the planned data panels. Section 5 states the methodological contribution, identifies its present limitations, and outlines the empirical work required for railway inspection.

2. Related Work

2.1. Visual inspection and dimensional wear estimation

Image-based carbon strip inspection spans component recognition, defect localization, and dimensional assessment. These tasks have related inputs but different annotation requirements. Studies directed at defect detection typically associate a region with a damage category, whereas a thickness task requires the location of a surface relative to a physically defined datum [13]. A bounding box locates the strip, whereas section-specific thickness requires registered contour coordinates. An image-level wear grade supports maintenance screening; a traceable profile additionally records

location and dimensional reference. A measurement pipeline therefore requires explicit correspondence between the image section, the reference section, and the reported dimensional quantity. This correspondence is particularly important when a registered unworn profile is used to infer material loss from residual thickness.

Optical dimensional inspection has used calibration targets and constrained imaging arrangements to convert image distances into physical units [14]. A planar transformation corrects perspective within its reference plane; depth displacement requires a separate geometric assessment. Studies of profile acquisition consequently distinguish errors of camera calibration from errors of surface localization [15]. For a carbon strip, contour sharpness and alignment with the calibrated plane therefore require separate verification. A dimensional inspection protocol must specify the observed section, the support features used for alignment, and the validity rules applied to obscured reference features. These requirements place segmentation performance within a defined metrological context. The homography acts on coordinates after lens-distortion correction; it cannot itself remove lens distortion or reconstruct an out-of-plane surface.

2.2. Semantic segmentation and contour localization

Encoder and decoder architectures combine semantic features with progressively recovered spatial detail. DeepLabv3+ uses multiscale context and a decoder to refine dense predictions, including predictions near object boundaries [16]. This architecture is a reasonable starting point for an elongated strip because the model must distinguish material from supports and background while retaining the narrow surface edge. Standard region objectives primarily reward agreement over image areas. A prediction can achieve strong overlap while shifting a small portion of the wear contour. Boundary losses were introduced in segmentation research to make contour errors more visible to optimization [17]. For wear measurement, boundary supervision must identify the specific contour used in dimensional estimation. The cited work concerns a DeepLabv3+ variant; the present description uses the established architecture as a starting point and does not attribute its original invention to that variant.

Distance-transform supervision and auxiliary boundary heads offer different ways of representing contour evidence [18]. A boundary head preserves local detail diluted by coarse features; appropriate supervision distinguishes the external material edge from scratches and deposits. For dimensional use, annotation policy must identify the exposed contour, the support datum, and an explicit state for invisible intervals. Work on segmentation evaluation has examined boundary-sensitive measures alongside region overlap [19]. These measures characterize localized errors, while pixel-to-millimeter conversion contributes a separate calibration-uncertainty component. Their role is therefore diagnostic: dimensional error remains the primary quantity when the system is intended to support thickness measurement.

2.3. Geometric priors and reliability of measurements

Geometric priors range from rigid templates to weak structural constraints. Shape-regularized segmentation can penalize configurations that are implausible for the target object [20]. Rigid templates can constrain the local profile deviations that constitute the wear measurement target. Weak ordering constraints preserve positive material intervals while allowing variation in the wear profile. Local smoothness regularization requires reduced weighting near material discontinuities. Excessive regularization at a chip or sharp groove can attenuate the local material-loss feature and bias the reconstructed surface.

Confidence-aware vision methods distinguish the quality of an estimate from the decision to issue that estimate [21]. In a profile task, this distinction allows visible intervals to be retained while occluded intervals are marked unavailable. Measurement uncertainty studies further motivate the separate treatment of scale, datum, and edge-localization contributions [22]. Combining these contributions requires attention to shared errors: a scale perturbation introduces dependence among all profile samples from the same image. Evaluation methodology for repeated specimen measurements therefore favors grouping and resampling at the physical specimen level [23]. A shared calibration shift can produce correlated profile deviations across repeated views, making the specimen and acquisition configuration relevant grouping variables for assessment. These concepts support an assessment design in which contour accuracy, dimensional agreement, accepted coverage, and failure modes are reported as complementary outcomes.

3. Geometrically Constrained Segmentation and Measurement

3.1. Acquisition geometry and measurement definition

The proposed arrangement uses a fixed side-view camera, a diffuse background, and a planar calibration target positioned in the same section as the exposed carbon strip profile. The measurable surface is the silhouette of that section. The recoverable quantity is the visible profile within the designated observation plane. The support datum is defined by a visible reference feature whose relationship to the material base is independently established. Images with an unobservable reference feature receive an invalid-datum flag and enter the reacquisition queue. An unworn reference profile is optional and must belong to the same registered section. The direct output is residual thickness; cumulative wear depth additionally requires the unworn reference.

Figure 1 presents the acquisition and measurement workflow, including the calibration target, rectification, segmentation, contour validity gate, and dimensional report. The two output branches deliberately separate a visible residual-thickness profile from wear depth computed with an additional unworn reference. Calibration is repeated when the camera or reference plane is moved. A checkerboard or coded target should cover the measurement region to limit amplification of registration error through extrapolation. The rectification map is established before learning and is held constant within each acquisition configuration. An independently withheld verification length provides a separate check on dimensional consistency.

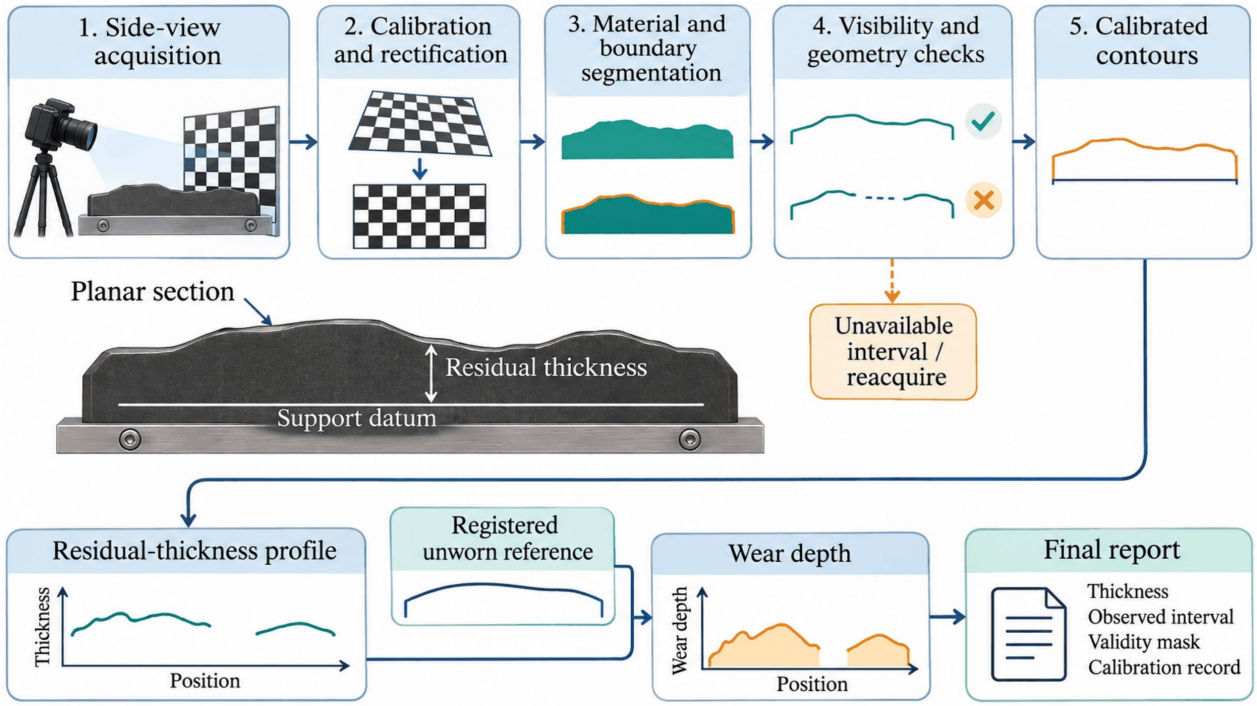


Figure 1. Acquisition and calibrated measurement workflow

The relationship between image coordinates and the calibrated observation plane is expressed by the following equation. Here, u and v denote image coordinates, X and Y denote physical coordinates in the rectified plane, H is the calibrated homography, and κ is the homogeneous normalization factor. Physical coordinates are expressed in millimeters after normalization.

$$\kappa \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} = H \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{1}$$

Applying the mapping equation to boundary coordinates corrects moderate perspective within the calibrated plane. The datum and upper contour are subsequently sampled at corresponding rectified longitudinal positions. Thickness denotes contour separation along the designated rectified normal. For a curved component base, a locally defined datum normal replaces the global vertical direction. Reference registration uses support features outside the worn surface to preserve the displacement associated with material loss. Each valid sample additionally requires an unobstructed boundary, positive material ordering, and a calibration residual below the configured acceptance limit.

3.2. Network architecture and geometric objective

The proposed architecture retains a DeepLabv3+ encoder, atrous spatial pyramid pooling, and a decoder that combines coarse semantic context with low-level spatial features. An auxiliary boundary branch predicts the exposed upper contour and support datum as distinct channels. A visibility branch estimates whether each longitudinal interval contains sufficient evidence for measurement. These branches share features and receive task-specific supervision. Material masks supervise region classification, contour annotations supervise boundary localization, and visibility labels identify intervals where an edge is unobservable. Contour intervals behind an obstruction receive a hidden-state label, establishing a distinction between observed surface evidence and inferred continuation.

Figure 2 describes the network architecture and the supervision supplied by material masks, contour labels, and visibility annotations. During training, the decoder output enters a differentiable contour extractor; inference uses a confidence-weighted path within the predicted material region. Low-confidence intervals retain their unavailable status during contour decoding. Feature dimensions are defined for an input resolution of 1024×512, an encoder output

stride of 16, and decoder features at one quarter of the input resolution. Upsampling restores the sampling grid before contour extraction. These settings specify the proposed implementation and the spatial resolutions used for boundary recovery.

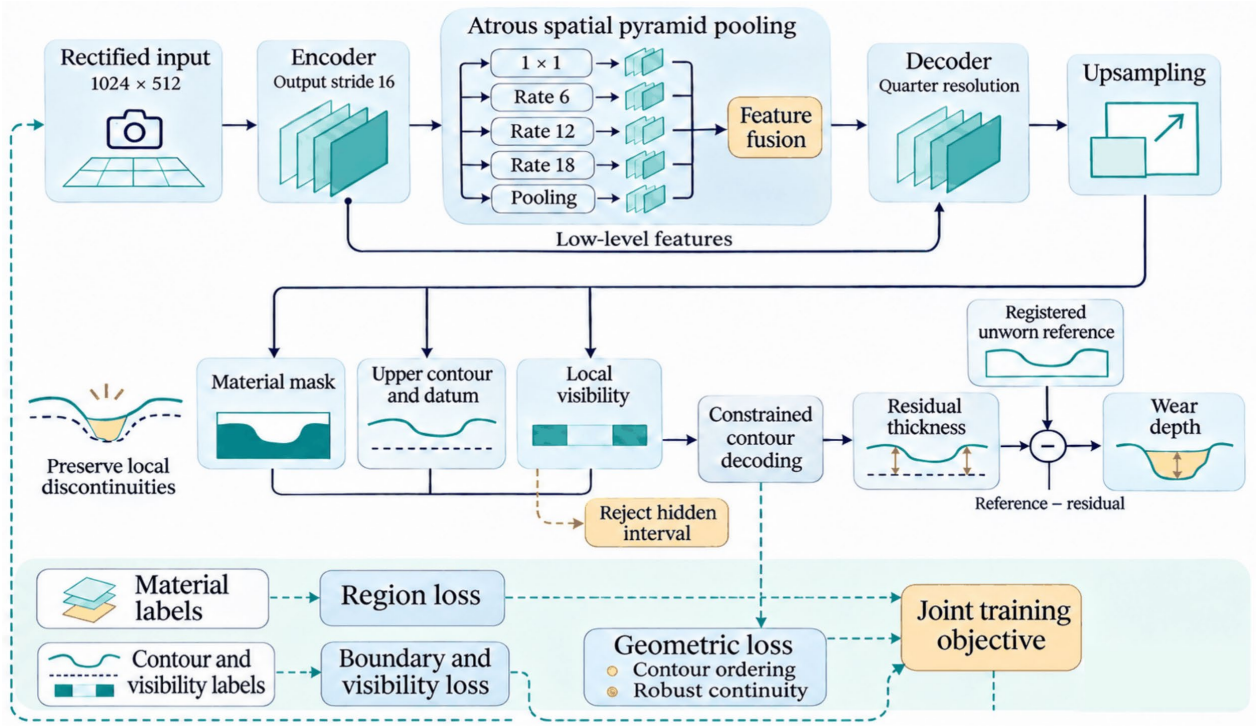


Figure 2. Geometry-aware DeepLabv3+ architecture with boundary and visibility branches and constrained contour decoding

The training objective combines region classification, boundary supervision, and geometric regularization through a weighted loss equation. The region term, L_{scg}, combines cross-entropy and soft Dice losses. The auxiliary term, L_{edge}, incorporates class-balanced boundary supervision and a separately normalized binary visibility loss. The geometric term, L_{geo}, penalizes structural inconsistencies in the recovered contours. The weights λ_b and λ_g control the contributions of auxiliary supervision and geometric regularization, respectively.

$$L = L_{seg} + \lambda_b L_{edge} + \lambda_g L_{geo} \tag{2}$$

The proposed initial values are 0.5 for the boundary weight and 0.2 for the geometric weight, with subsequent adjustment based on the validation partition. Boundary targets occupy a narrow band around annotated contours to moderate the influence of isolated one-pixel annotation differences during early optimization. Region and boundary terms are normalized by valid supervised pixels, while the geometric term is normalized by valid longitudinal samples. This normalization maintains the relative contribution of geometric supervision when occlusion reduces the observable contour length. The stated 0.5 auxiliary weight applies to the combined boundary and visibility term, not solely to boundary supervision. The internal boundary-to-visibility weighting and boundary-band width remain implementation parameters that must be specified before an empirical replication.

Within the geometric penalty equation, the ordering term discourages inverted material intervals, while the continuity term regulates abrupt variations in local contour slope. The upper contour and base datum are denoted by $y_{u,i}$ and $y_{b,i}$, respectively. The parameter m defines a positive separation margin, Δ^2 denotes the second longitudinal finite difference, and ρ is a robust penalty function. The coefficient η controls continuity regularization. The summation covers the N valid samples identified by the visibility mask; continuity evaluation uses neighboring samples within the same visible interval.

$$L_{geo} = \frac{1}{N}\sum_{i=1}^{N}\left[max\left(0, m + y_{u,i} - y_{b,i}\right) + \eta\rho\left(\Delta^2 y_{u,i}\right)\right] \tag{3}$$

The ordering term assumes that rectified vertical coordinates increase toward the base. Reversing the coordinate convention requires a corresponding reversal of the ordering condition. The robust continuity term bounds the influence of large second differences while retaining discontinuities supported by boundary evidence. Its local weight is reduced near a high-confidence material discontinuity, allowing preservation of a sharp chip. Observable edge evidence determines this adaptive weighting so that regularization follows local contour support. The initial 0.2 mm margin

provides a positivity safeguard within contour decoding. The reference scale used to convert the margin remains fixed for each training configuration. Because the margin is expressed in millimeters, it must be evaluated in calibrated coordinates or converted to pixels using the local mapping. Robust continuity reduces sensitivity to large differences but does not guarantee preservation of a chip; chip-depth bias requires separate verification.

3.3. Profile extraction and uncertainty reporting

At inference, the exposed upper contour and base datum are extracted where boundary evidence and visibility exceed validation-selected thresholds. Multiple disconnected candidates are resolved using a path cost that combines boundary probability, material support, and bounded local displacement. Occlusions longer than the specified gap tolerance terminate the decoded contour interval. An isolated invalid sample receives an unavailable-state label, and the reported minimum thickness is restricted to the observed interval. Unobserved intervals retain an unresolved inspection status. This distinction is relevant when the inspection criterion depends on the smallest remaining section. The gap tolerance controls contour connectivity only: a hidden sample is never converted into an accepted measurement, even within a short gap. The minimum is therefore the minimum over accepted visible positions, not a certified minimum over the whole strip.

Residual thickness and reference-dependent wear depth are related by the measurement equation below. In a rectified, locally parallel section, residual thickness t_i is the separation between the base datum and exposed upper contour. Wear depth w_i is the difference between the registered unworn thickness $t_{0,i}$ and residual thickness. A measured unworn profile provides the reference when longitudinal manufacturing variation is relevant.

$$t_i = y_{b,i} - y_{u,i}, w_i = t_{0,i} - t_i \tag{4}$$

The thickness equation is applied at registered longitudinal positions, with a validity record accompanying each accepted interval. This record contains the boundary-localization contribution, datum contribution, shared calibration contribution, and a flag for unresolved depth displacement. The depth flag identifies geometric model inadequacy and triggers verification or reacquisition. Repeated observations characterize repeatability, while independent reference measurements establish dimensional agreement. Reference measurements acquired under repositioning should be compared with image estimates at corresponding locations. The mean profile retains its original validity mask to distinguish measured and unavailable positions.

The proposed training protocol uses specimen-disjoint partitions, consistent augmentation of images and geometric targets, and validation-based early stopping. Geometric targets undergo the same transformation as the image to preserve spatial correspondence. Photometric augmentation includes exposure changes and localized glare; geometric augmentation is restricted to transformations with an available calibration mapping. The proposed optimizer settings comprise an initial learning rate of 0.0001, a batch size of eight, and a maximum of 120 epochs. These settings define the implementation protocol. Evaluation separates dimensional error, boundary agreement, visibility selection, and calibration sensitivity to examine their distinct contributions to inspection performance. These are proposed settings rather than a record of completed training. An executable implementation additionally requires the backbone, optimizer identity, initialization, random seeds, augmentation distributions, and decoding thresholds; these are not established by the numerical panels.

For reproducible annotation, each contour label records its coordinate frame and whether its position was directly observed. Intervals obscured by a support bracket receive an unavailable contour state. Visible endpoints delimit the measurable interval and anchor the visibility labels. When annotators disagree about an edge obscured by a deposit, adjudication records an uncertain contour state and retains both candidate interpretations. Datum annotation follows a comparable policy because datum displacement propagates directly into residual thickness. An independent reviewer verifies a subset of datum labels against the acquisition fixture and reference section before training. These procedures specify the information required beyond a binary material mask. During model development, images with valid material segmentation and invalid measurement geometry remain in the dataset, with their dimensional targets masked appropriately. Retaining these cases supports learning of the distinction between component recognition and dimensional validity. The inspection record preserves these outputs as separate diagnostic fields. Versioned annotation guidelines and a held-out adjudicated subset support assessment of model changes independently of changes in labeling convention.

4. Experimental Design and Quantitative Analysis

4.1. Benchmark design and comparative accuracy

The numerical benchmark design contains 120 distinct specimen identifiers, each associated with 20 imaging conditions, giving 2400 image records. A fixed partition assigns 72 specimens to training design, 24 to validation design, and 24 to evaluation, corresponding to 1440, 480, and 480 records. Records are grouped by specimen identifier for

partitioning and comparison. The identifiers provide a grouping structure for reproducible demonstrations of specimen-level analysis. Each evaluation specimen contributes smooth reference thickness variation and controlled localized depressions. Imaging-condition variables encode exposure, blur, partial occlusion, and calibration perturbation. Factorial robustness protocols vary acquisition factors independently to separate main effects from interactions [24]. The evaluation tracks these factors separately to examine changes in contour visibility and dimensional sensitivity. Identifier separation prevents overlap within the demonstration design, but does not establish the existence of 120 independently acquired physical strips. The numerical panels are descriptive illustrations of the assessment protocol, not a documented physical acquisition campaign.

Reference thickness values range across a nominal 8–22 mm inspection interval, with a 30 mm unworn reference used only to demonstrate subtraction for wear depth. The reference profile includes broad undulation and local depressions whose positions vary across specimen identifiers. The error analysis distinguishes signed residuals from absolute error and compares their distribution across the defined acquisition conditions. The aggregate comparison retains physical units and a common set of evaluation metrics. Five method labels represent a threshold contour baseline, U-Net, DeepLabv3+, a boundary-enhanced variant, and the complete geometry-aware configuration. The labels distinguish the region-based, boundary-enhanced, and geometry-aware configurations. Reference protocols for calibrated vision distinguish a dimensional comparator from an annotation used to train the model [25]. A physical study should acquire independent reference profiles to establish the dimensional comparator.

The primary metric is mean absolute residual-thickness error across registered valid positions. Root means square error and the upper tail of absolute error describe sensitivity to large excursions. Boundary F1 is defined with a fixed two-pixel matching tolerance, while intersection over union is reported only as a region diagnostic. The segmentation summary scores accompany the dimensional error distributions. The two metric groups address region agreement and dimensional deviation, respectively. Grouped evaluation protocols assign repeated specimen observations to a common partition to control dependence [26]. The figure code exports the values used in the plots so that dimensions and errors can be checked independently of the document. For physical validation, sample-level errors should first be averaged within each specimen and specimen means then weighted equally. Confidence intervals should resample specimens, not treat correlated longitudinal samples as independent. A two-pixel boundary tolerance corresponds to different physical distances when image scale changes.

Uncertainty and selection are evaluated separately. An engineering error target of 0.25 mm defines the dimensional comparison. The proposed validity gate accepts 92% of records under nominal conditions and labels the remaining records as unavailable. Selective prediction protocols assess accepted coverage alongside retained error and choose thresholds using validation data [27]. The benchmark also includes a fixed low-to-high stress sequence, five calibration offsets, and a set of continuity strengths. The comparison uses descriptive metrics to examine differences among the defined configurations. The evaluation record retains dimensional errors, acquisition conditions, and validity labels for examining their effects on the reported profile. The 0.25 mm value is an illustrative engineering target, not a cited railway acceptance standard. Record acceptance and longitudinal sample coverage are different quantities and should be reported separately; the stated 92% concerns record only. All comparative errors require the same accepted positions, or a separate error-versus-coverage comparison.

Table 1 gives the nominal performance comparison for the five configurations. Mean absolute error decreases from 0.462 mm for threshold extraction to 0.351 mm for U-Net, 0.286 mm for DeepLabv3+, 0.173 mm for the boundary-enhanced variant, and 0.118 mm for the complete configuration. Relative to the unconstrained DeepLabv3+ configuration, the last value represents a 58.7% reduction, calculated from the displayed means. Root means square error follows the same ordering, with values of 0.610, 0.468, 0.389, 0.238, and 0.164 mm. The error metrics retain the same configuration ordering across the comparison. Evaluation methods for dimensional vision report errors in physical units alongside image-space metrics [28]. Consequently, the percentage reduction is secondary to the absolute millimeter error and the behavior of the largest deviations. The 58.7% reduction is descriptive arithmetic on the illustrated means. Without paired empirical records and repeated training runs, it is neither a significance estimates nor evidence that geometric regularization causes this reduction.

Table 1. Nominal dimensional and segmentation performance across five configurations

Configuration	MAE (mm)	RMSE (mm)	IoU (%)	Boundary F1 (%)
Threshold	0.462	0.610	88.6	82.4
U-Net	0.351	0.468	92.7	87.6
DeepLabv3+	0.286	0.389	94.6	90.2
Boundary enhanced	0.173	0.238	96.1	94.2

Complete geometry	0.118	0.164	96.8	95.8

Agreement analysis evaluates individual dimensional discrepancies alongside mean displacement [29]. Figure 3(a) compares the five mean errors with a 0.25 mm target, placing the complete configuration 0.132 mm below that limit. Figure 3(b) shows the specimen error distributions, including the heavier tail of threshold extraction. Figure 3(c) compares 120 reference and predicted thickness samples over 8–22 mm using identity and ±0.25 mm guides. Figure 3(d) places the complete configuration residuals against their paired mean thickness, with the same tolerance band. The panels quantify complementary aspects of dimensional performance. The bar chart compares method-level errors, the distributions characterize dispersion, the agreement scatter reveals scale-dependent displacement, and the residual plot assesses error variation across thickness. The residual pattern provides a basis for examining thickness-dependent displacement across the evaluated dimensional range. The 120 plotted pairs are thickness samples and must not be equated with 120 evaluation specimens: the evaluation partition contains 24 specimen identifiers. A mean error below 0.25 mm does not imply that every sample meets that allowance.

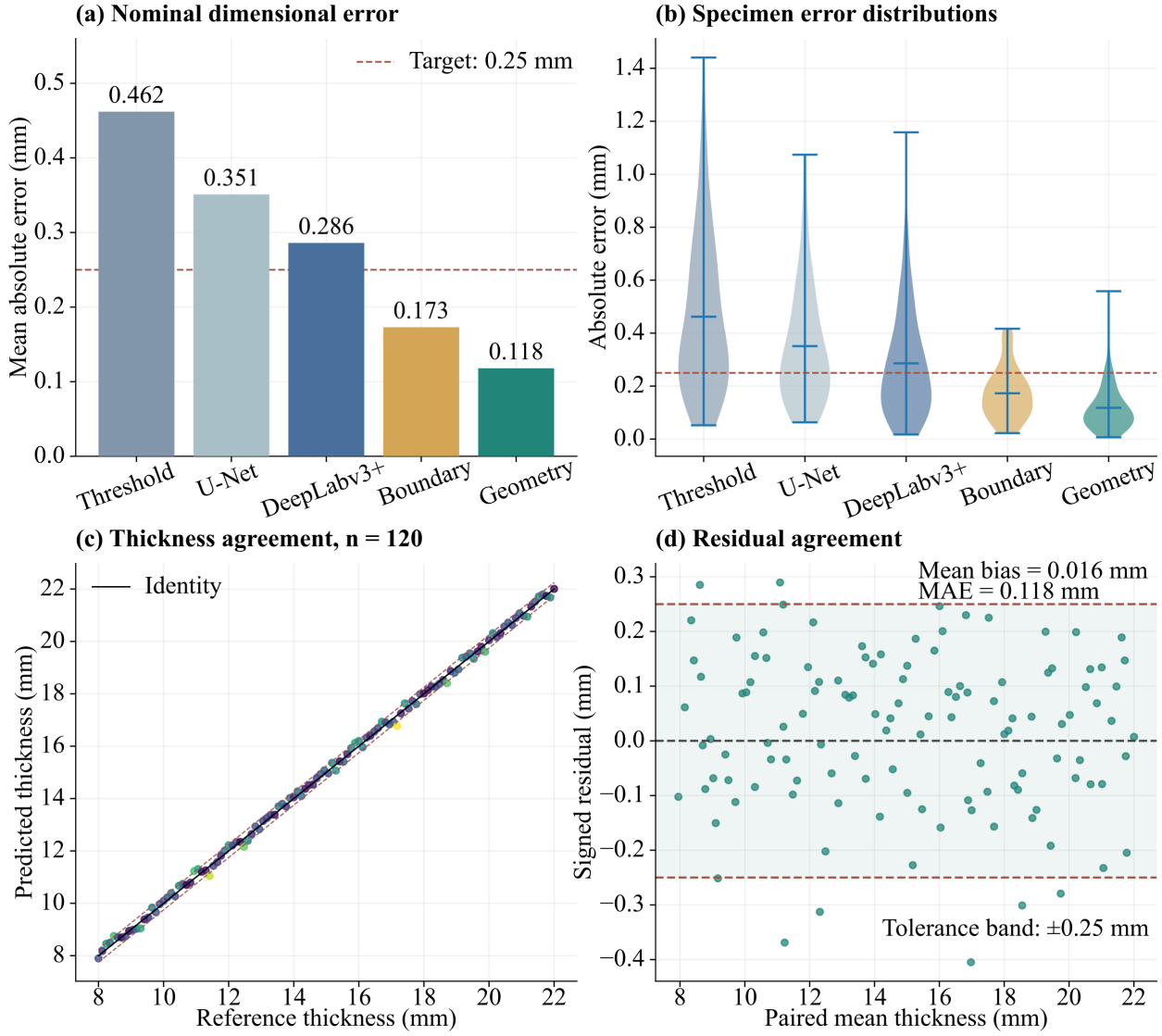


Figure 3. Nominal performance comparison with (a) mean absolute error, (b) absolute-error distributions, (c) reference and predicted thickness, and (d) residual agreement

The boundary-enhanced and complete configurations have intersection-over-union values of 96.1% and 96.8%, respectively, whereas their mean absolute errors differ by 0.055 mm. The paired metrics associate a modest change in region overlap with a larger proportional difference in dimensional error. Material pixels far from the measurement edge dominate region overlap, whereas dimensional error depends on displacement at the selected contours. The boundary F1 values are 94.2% and 95.8% for the same two configurations. The score ordering is consistent across the two configurations; the relationship between boundary F1 and dimensional error should be examined using paired specimen records. A future empirical analysis should retain per-specimen pairs of the two quantities and evaluate whether improvements remain associated after accounting for scale and visibility. Controlling acquisition conditions helps distinguish a direct metric association from a shared response to image quality.

Interpretation of the predicted-thickness scatter depends on the range of reference thicknesses. A wide between-specimen thickness range can dominate the apparent correlation between reference and predicted values. Vertical deviations from the identity line quantify the dimensional discrepancies. The residual agreement panel uses a much narrower dimensional range and retains the sign of each discrepancy. Positive residuals indicate overestimation of remaining thickness, which can be more consequential than a similarly sized underestimation when a minimum section is used for screening. The sign of each residual preserves the direction of the dimensional discrepancy. An operational report would additionally identify the minimum observed location, register that location to the reference profile, and state whether the validity gate excluded nearby intervals. The inspection report therefore retains location, reference registration, and visibility alongside the image-level metrics.

4.2. Component sensitivity and profile reliability

Matched ablation protocols hold training and evaluation conditions constant when comparing model components [30]. Figure 4(a) shows a component sequence with errors of 0.286, 0.219, 0.173, 0.139, and 0.118 mm for the baseline, edge branch, boundary supervision, ordering constraint, and complete configuration. Figure 4(b) displays a single heatmap across five illumination states and five occlusion levels. Figure 4(c) compares four method responses over six stress levels, with the complete configuration increasing from 0.118 to 0.241 mm. Figure 4(d) displays a threshold sensitivity scatter in which accepted coverage falls as the confidence gate is tightened. The incremental comparison separates the error levels associated with each component configuration. The heatmap applies an explicit interaction term, so simultaneous darkness and occlusion receive more error than either isolated factor. An actual component study would therefore repeat training across seeds and preserve a common validity policy before assigning an improvement to the geometric term. The sequence is cumulative rather than a factorial ablation: each point includes the preceding components. Its successive differences cannot isolate component interactions. The heatmap interaction is prescribed for illustration and is not an empirically estimated illumination-by-occlusion effect.

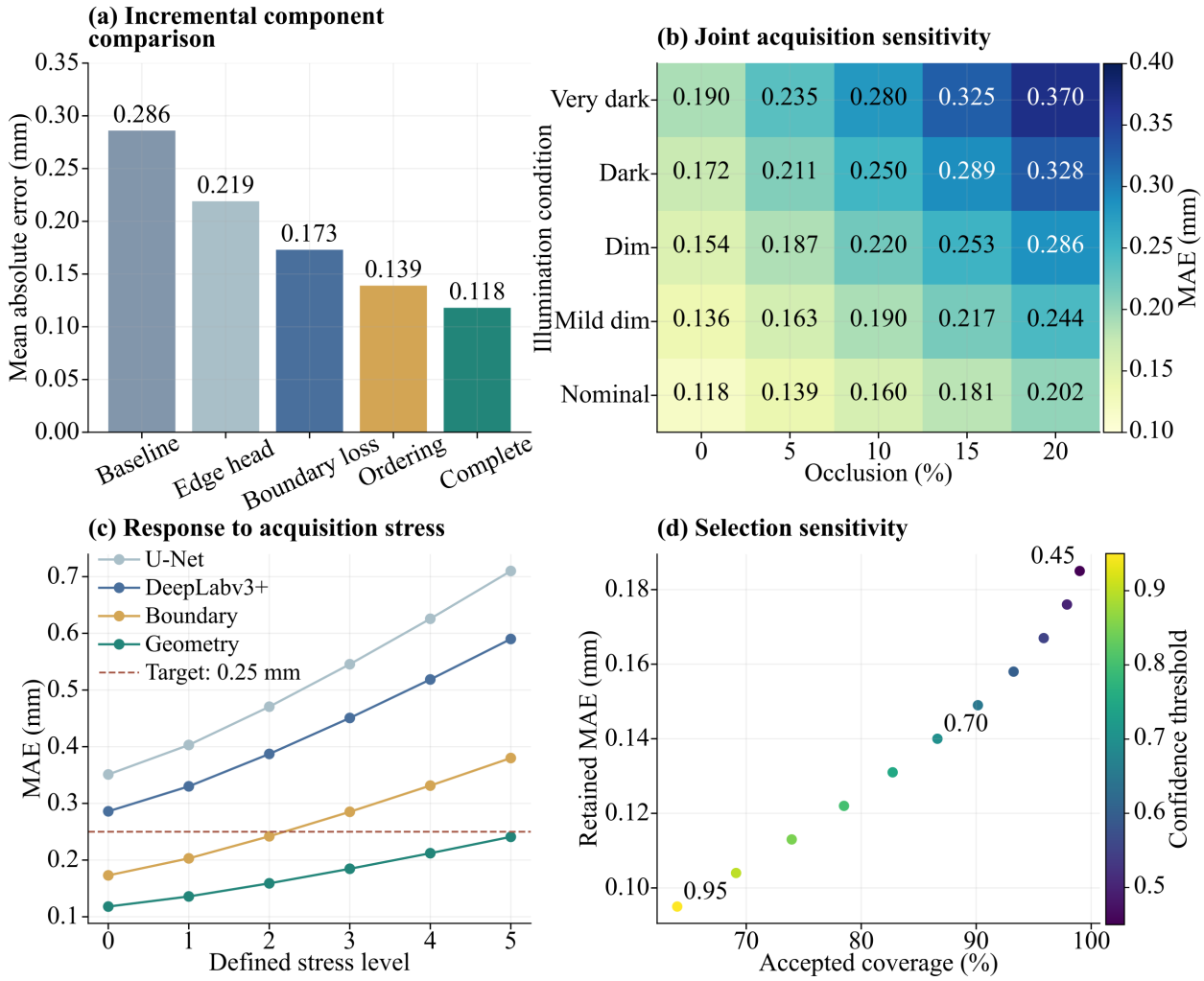


Figure 4. Component and operating-envelope diagnostics with (a) incremental configurations, (b) illumination–occlusion sensitivity, (c) stress response, and (d) confidence-gate selection

The complete configuration error is 0.118 mm under nominal conditions and 0.241 mm at the highest defined stress level, corresponding to a 0.009 mm margin below the specified target. At this margin, calibration and reference uncertainty can become the dominant contributions relative to the remaining allowance. The boundary-enhanced configuration has a nominal mean absolute error of 0.173 mm. The boundary-enhanced point also represents the complete design with its geometric term removed. This maintains consistency between the component sequence and the comparative table. The ordering term is intended to prevent contour inversion, while the robust continuity term addresses isolated oscillation. Physical experiments should distinguish ordering validity from localization error because the two terms address different contour properties. The 0.009 mm difference is a difference between an illustrative mean and a target; it is not a demonstrated safety margin, particularly when reference uncertainty and upper-tail error are unresolved.

The combined illumination and occlusion response distinguishes loss of contrast from loss of observable contour length. At nominal illumination, increasing occlusion from 0% to 20% changes the error from 0.118 to 0.202 mm. Under the darkest condition, the same change increases error from 0.190 to 0.370 mm. The larger increment identifies a stronger combined response to low illumination and occlusion, where weak edge evidence increases sensitivity near the remaining visible contour. A stricter confidence gate lowers retained error while reducing accepted coverage. The displayed threshold sweeps spans 0.45–0.95 and produces coverage from 99% to 64%, with retained error decreasing from 0.185 to 0.095 mm. Threshold selection therefore requires an explicit coverage requirement and a record of the locations rejected at each setting. The sweep describes selective reporting; reduced retained error alone is insufficient to establish better measurement performance when coverage also decreases.

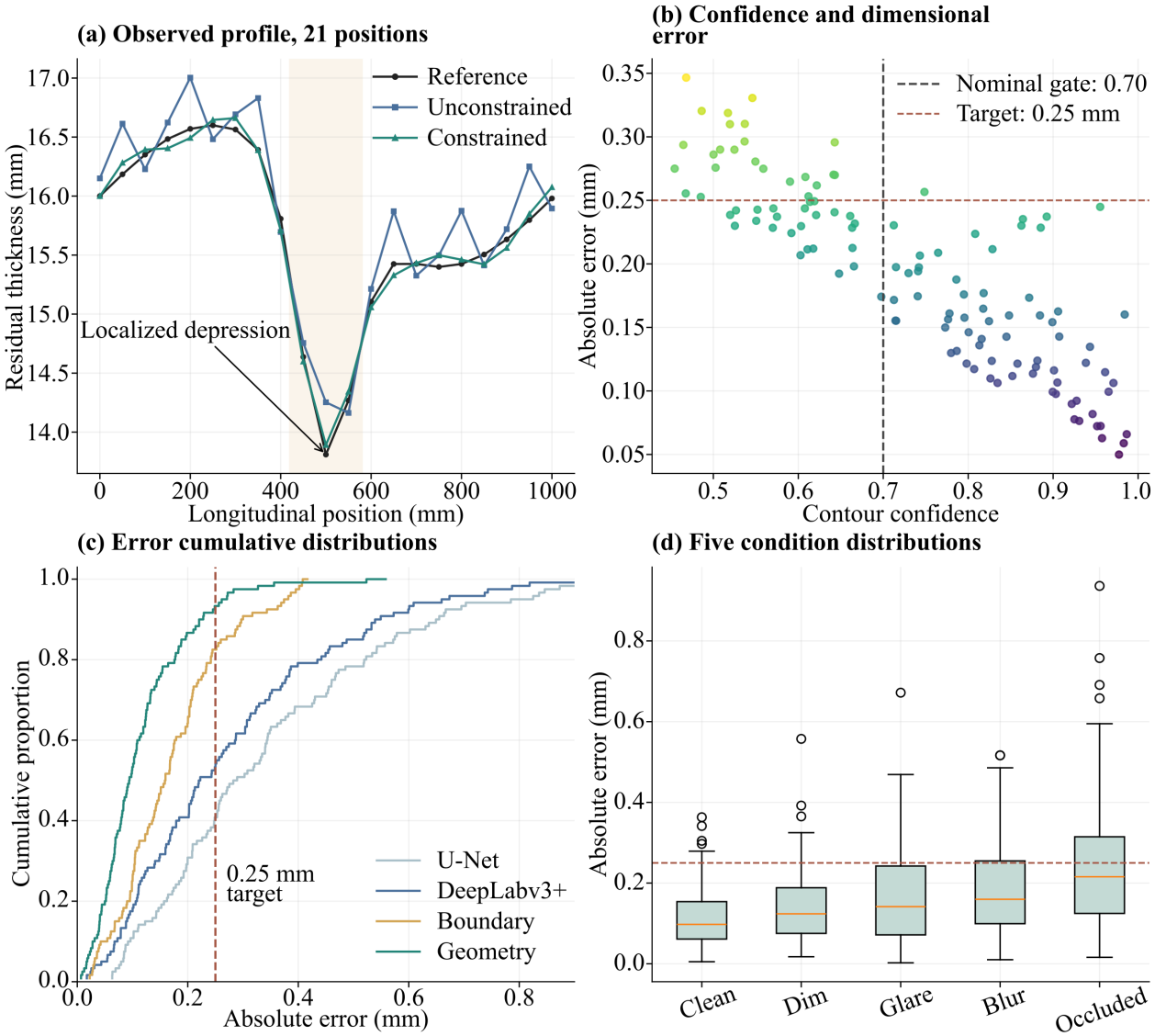


Figure 5. Profile and reliability analysis with (a) longitudinal residual thickness, (b) confidence and error, (c) cumulative error distributions, and (d) condition-dependent dispersion

Confidence-calibration methods assess probability reliability separately from discrimination between easier and harder observations [31]. Figure 5(a) compares a reference wear profile, an unconstrained estimate, and a constrained estimate over 21 longitudinal positions, preserving the position and shape of a local depression. Figure 5(b) plots error

against confidence for 120 observations and marks the nominal selection threshold. Figure 5(c) compares empirical cumulative distributions for four configurations and marks the 0.25 mm target. Figure 5(d) shows box distributions for five image-condition groups. At the profile depression, the constrained trace retains the location and approximate width of the feature while reducing high-frequency excursions elsewhere. The confidence scatter includes substantial dispersion, representing uncertainty in the relationship between contour confidence and dimensional error. After confidence selection, the cumulative distributions describe the retained error tail and the fraction satisfying a specified allowance.

The profile example includes a localized reduction near the middle of the observed section, where an aggressive continuity prior could cause overestimation of the remaining material. The constrained profile retains that reduction while suppressing oscillation over the adjacent visible intervals. A future evaluation should report separate errors for smooth regions, groove flanks, and chip edges alongside the profile-wide mean. The five condition groups represent clean images, dim illumination, glare, blur, and partial occlusion, with target mean errors of 0.118, 0.143, 0.167, 0.192, and 0.241 mm. The overlapping distributions indicate substantial error variability within each condition group. The last group particularly requires a distinction between a large observed error and an unavailable segment; the latter must remain excluded from thickness reporting and explicitly counted in coverage.

At the 0.25 mm target, the cumulative distribution is interpreted as the proportion of observations whose absolute error satisfies the dimensional allowance. The four curves also expose the residual tail beyond this threshold, which remains relevant when a specimen contains several inspection locations. The median and upper-tail deviations describe distinct aspects of the error distribution. In the condition distributions, positive skew indicates that a few difficult observations contribute disproportionately to the mean. Evaluation of upper-tail observations requires review of the corresponding boundary evidence and datum localization. For a groove flank, a displaced upper contour changes the apparent width and depth of local loss; for an occluded segment, the relevant response is an unavailable interval. These cases require separate diagnostic labels even when their scalar errors are similar.

4.3. Calibration sensitivity and inspection diagnostics

Prediction-interval evaluation compares empirical coverage with the stated nominal probability [32]. Figure 6(a) compares nominal 80%, 90%, 95%, 97%, and 99% intervals with empirical coverages of 77%, 87%, 93%, 95%, and 97%. Figure 6(b) shows dimensional bias at scale offsets of −2%, −1%, 0%, 1%, and 2%, using a 15 mm reference section to produce biases of −0.30, −0.15, 0, 0.15, and 0.30 mm. The interval example exhibits undercoverage at every displayed level. A two-percent scale error alone exceeds the 0.25 mm specified target at the stated section thickness. The comparison isolates the shared multiplicative scale contribution from contour-localization error and local profile smoothness. These interval coverages are diagnostic examples, not a validated uncertainty model. A physical evaluation must specify interval construction, the unit of coverage assessment, sample counts, and widths before interpreting nominal probability as a measurement guarantee.

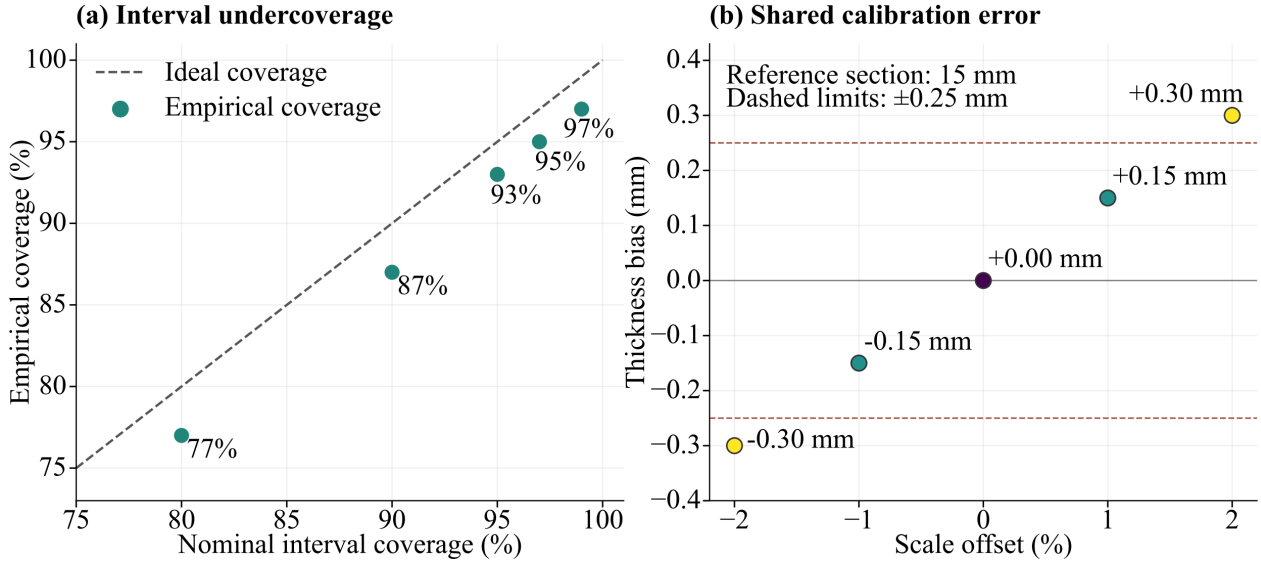


Figure 6. Uncertainty diagnostics with (a) interval coverage calibration and (b) scale-offset sensitivity

Scale sensitivity varies with the actual thickness being measured. At the ends of the 8–22 mm reference interval, a two-percent scale offset would correspond to 0.16 and 0.44 mm, respectively. The 15 mm example defines a section-specific conversion between relative scale error and dimensional bias. Datum displacement contributes an additive error: a common upward or downward shift changes local thickness through the reference position. A practical reference

protocol should vary scale and datum position independently. It should also include out-of-plane displacement, because a planar transformation can appear numerically stable while the observed contour lies away from the calibrated section. An unresolved depth discrepancy requires a qualified measurement or an invalid-geometry flag, with uncertainty tied to the available geometric evidence.

The coverage deficits are three percentage points at the 80% and 90% nominal levels and two percentage points at the three higher levels. The consistent deficits across interval settings indicate a calibration discrepancy that requires assessment across the full coverage curve. A physical evaluation should retain the full coverage curve, interval width, and specimen grouping information. The scale experiment isolates a shared contribution that affects every point in an image. At a 15 mm section, the absolute bias reaches the 0.25 mm specified allowance at approximately 1.67% scale error. This value provides an acquisition-verification trigger for that thickness and allowance. At a different thickness, the trigger changes proportionally, so calibration acceptance should be tied to the inspected dimensional range.

Inference benchmarking specifies the device, batch size, numerical precision, synchronization procedure, and preprocessing scope [33]. Figure 7(a) displays five operating configurations in an error–latency plane, with latency coordinates of 24, 31, 38, 43, and 49 ms per record. Figure 7(b) shows 100 failure records distributed among occlusion, glare, datum loss, severe chips, and calibration displacement with counts of 34, 23, 18, 15, and 10. The latency coordinates associate each configuration with a separate processing budget. The resulting points describe the error–latency tradeoff across the configurations. The failure categories separate visibility, datum, material-damage, and calibration issues. Occlusion occupies the largest category because the proposed validity policy prioritizes direct contour evidence over completion of a visually plausible surface. The displayed latency coordinates are processing-budget examples, not reproducible device measurements. Likewise, the 100 failure records illustrate category allocation rather than an observed deployment failure rate; their denominator is the displayed failure set, not all acquisitions.

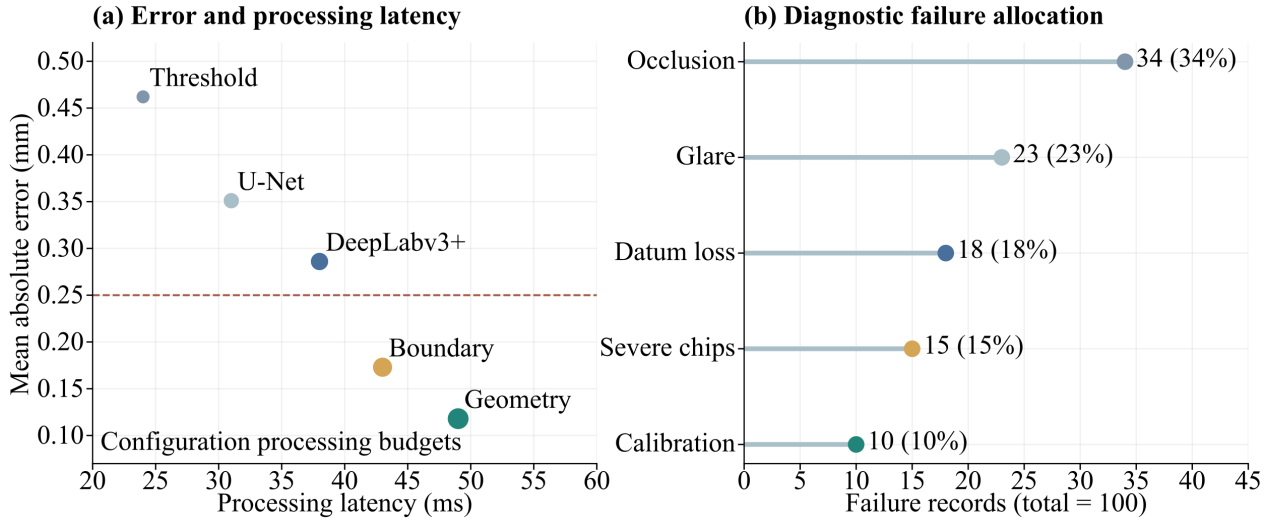


Figure 7. Deployment diagnostics with (a) error and processing latency and (b) failure-category allocation

The proposed full pipeline would time rectification, network inference, contour extraction, validity checks, and report assembly separately. Camera transfer and reference registration should be included whenever they are required for the deployed inspection cycle. End-to-end latency includes acquisition transfer, rectification, inference, contour extraction, reference registration, and report generation. Rejection handling also has a measurable operational consequence: a record with a missing datum may require reacquisition, whereas glare may be corrected through exposure or illumination adjustment. Severe chips require retention of the observed discontinuity and a material-damage flag in the dimensional report. The diagnostic categories are therefore attached to different corrective actions. Before field use, each category should be assigned reproducible labeling criteria, reviewed against reference measurements, and examined for disagreements between inspectors. The resulting records support analysis of configuration-specific processing demands and corrective actions.

Each retained inspection record should preserve the acquisition configuration, calibration identifier, specimen identifier, validity mask, and reference registration version. Repeated inspections should compare mutually visible registered positions and separately record intervals visible on only one occasion. A newly visible minimum must be distinguished from temporal material loss. The report should display the original contour evidence at the minimum and retain unresolved datum candidates. Reacquisition should record changes in exposure, camera position, and illumination, together with the applicable calibration identifier. For datum loss, a second view should verify the support feature at the registered section; for glare, matched contour locations should be compared before and after illumination adjustment. The reference profile should remain fixed during this comparison. Unsuccessful attempts should be retained to support

future estimates of reacquisition burden and reference disagreement. These records are prerequisites for an operational assessment rather than evidence of deployment effectiveness.

5. Conclusion

This paper formulated pantograph carbon strip wear measurement as a coupled problem of semantic segmentation, contour validity, and calibrated dimensional interpretation. The proposed DeepLabv3+ framework adds boundary and visibility supervision and applies weak geometric constraints to preserve physically meaningful contour ordering. Residual thickness is distinguished from reference-dependent wear depth, and unsupported intervals remain unavailable with an explicit unavailable-state label. The evaluation framework connects region and boundary diagnostics with dimensional agreement and explicitly retained coverage. The numerical examples illustrate this assessment structure and do not constitute independent experimental verification of accuracy, runtime, or maintenance suitability.

The present framework is limited by its planar observation geometry and its dependence on a clearly identifiable support datum. Generalization across acquisition environments, component designs, and damage morphologies requires broader specimen-level assessment. The planar observation assumption also excludes hidden transverse damage and unresolved depth variation. Excessive geometric regularization and inaccurate visibility labels can distort sharp material-loss features.

Future work should acquire registered images and traceable reference profiles across independent specimens, acquisition sessions, and operating environments. The network should then be trained with repeated seeds and evaluated using specimen-level uncertainty estimates and fixed validity rules. Additional investigation should address discontinuity-preserving regularization, independent calibration verification, and acquisition strategies for obscured surfaces. A field study should connect measurement confidence and rejected intervals to reacquisition procedures before any maintenance decision is automated.

6. Declarations

6.1 Author Contributions

Biljana Todorović contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Mihaila Pavlović contributes to validation, analysis, investigation, data collection, draft preparation. Dejana Simić contributes to conceptualization, validation, analysis. All authors have read and agreed with the manuscript before its submission and publication.

6.2 Funding and Acknowledgments

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

6.3 Institutional Review Board Statement

Not applicable.

6.4 Informed Consent Statement

Not applicable.

6.5 Date Availability

Not applicable.

6.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

6.7 Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors utilized Grammarly for linguistic polishing, logical organization and draft revision. The generative AI tool was not involved in research design, data collection, data analysis, or the equation of core conclusions. All outputs generated by AI were reviewed, revised and verified manually by the authors. The authors take full responsibility for all content of this manuscript.

7. References

[1] Zhou, N., Zhi, X., Cheng, Y., Sun, Y., Wang, J., Gu, Z., Li, Z., & Zhang, W. (2024). Contact strip of pantograph heuristic wear model and its application. Tribology International, 194, Article 109546. https://doi.org/10.1016/j.triboint.2024.109546

[2] Judek, S. (2025). Defect detection of railway pantograph contact strips based on laser scanning measurements. Measurement Science and Technology, 36(11), Article 116110. https://doi.org/10.1088/1361-6501/ae15af

[3] Song, Y., Mei, G., Liu, Z., & Gao, S. (2024). Assessment of railway pantograph-catenary interaction performance with realistic pantograph strip imperfection. Vehicle System Dynamics, 62(9), 2355–2374. https://doi.org/10.1080/00423114.2023.2289662

[4] Yang, K., Liu, Y., Li, X., Bai, Z., Wan, Y., Xiao, Y., & Li, J. (2024). A high-accuracy single-frame 3D reconstruction method with color speckle projection for pantograph sliders. Measurement, 237, Article 115192. https://doi.org/10.1016/j.measurement.2024.115192

[5] Wang, J., Xie, C., Zhang, X., Zhang, Z., & Wang, X. (2026). Online pantograph wear detection under busbar occlusion using line laser scanning with region-based registration. Measurement, 288, Article 122553. https://doi.org/10.1016/j.measurement.2026.122553

[6] Hou, X., Chen, P., & Gu, H. (2024). LM-DeeplabV3+: A lightweight image segmentation algorithm based on multi-scale feature interaction. Applied Sciences, 14(4), Article 1558. https://doi.org/10.3390/app14041558

[7] Sun, J., & Zhang, P. (2026). MSCA-Deeplabv3+: A fine-grained optical image segmentation model based on multi-scale channel attention. Alexandria Engineering Journal, 145, 430–445. https://doi.org/10.1016/j.aej.2026.04.026

[8] Qiu, S., Chen, J., Zhang, H., Wan, R., Xue, X., & Pu, J. (2024). Guided contrastive boundary learning for semantic segmentation. Pattern Recognition, 155, Article 110723. https://doi.org/10.1016/j.patcog.2024.110723

[9] Zhang, P., Cheng, Y., & Tamura, S. (2024). Shape prior-constrained deep learning network for medical image segmentation. Computers in Biology and Medicine, 180, Article 108932. https://doi.org/10.1016/j.compbiomed.2024.108932

[10] Zhang, S., Ni, P., Han, Q., Wen, J., Du, X., & Li, J. (2025). Measurement uncertainty quantification for computer vision-based structural dynamic displacement monitoring. Measurement, 253, Article 117835. https://doi.org/10.1016/j.measurement.2025.117835

[11] Hendrickx, K., Perini, L., Van der Plas, D., Meert, W., & Davis, J. (2024). Machine learning with a reject option: A survey. Machine Learning, 113(5), 3073–3110. https://doi.org/10.1007/s10994-024-06534-x

[12] Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S., & Scheinost, D. (2024). Data leakage inflates prediction performance in connectome-based machine learning models. Nature Communications, 15(1), Article 1829. https://doi.org/10.1038/s41467-024-46150-w

[13] Zschiesche, K., & Reiterer, A. (2024). Optical measurement system for monitoring railway infrastructure—A review. Applied Sciences, 14(19), Article 8801. https://doi.org/10.3390/app14198801

[14] Duan, Q., Wang, Z., Huang, J., Xing, C., Li, Z., Qi, M., Gao, J., & Ai, S. (2024). A deep-learning based high-accuracy camera calibration method for large-scale scene. Precision Engineering, 88, 464–474. https://doi.org/10.1016/j.precisioneng.2024.02.019

[15] Yang, H., Zhu, K., Zhao, J., Lan, T., & Pan, B. (2025). Metrological performance quantification of a 3D-DIC system. Measurement, 249, Article 117008. https://doi.org/10.1016/j.measurement.2025.117008

[16] Liu, H., Chen, Y., Wang, R., Li, M., & Li, Z. (2025). MFA-Deeplabv3+: An improved lightweight semantic segmentation algorithm based on Deeplabv3+. Complex & Intelligent Systems, 11(10), Article 424. https://doi.org/10.1007/s40747-025-02028-y

[17] Lin, Q., Chen, X., Chen, C., & Garibaldi, J. M. (2024). Boundary-wise loss for medical image segmentation based on fuzzy rough sets. Information Sciences, 661, Article 120183. https://doi.org/10.1016/j.ins.2024.120183

[18] Vareto, R. H., Szczerbacki, R., Lima, L. A., Vaz-de-Melo, P. O. S., & Schwartz, W. R. (2026). Distance Transform Loss: Boundary-aware segmentation of seismic data. Computers & Geosciences, 207, Article 106061. https://doi.org/10.1016/j.cageo.2025.106061

[19] Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M. A., Wiesenfarth, M., Kavur, A. E., Sudre, C. H., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Rädsch, T., … Jäger, P. F. (2024). Metrics Reloaded: Recommendations for image analysis validation. Nature Methods, 21(2), 195–212. https://doi.org/10.1038/s41592-023-02151-z

[20] Zeng, L., Li, X., Yang, X., Chen, W., Liu, J., Shen, L., & Wu, S. (2024). SCPMan: Shape context and prior constrained multi-scale attention network for pancreatic segmentation. Expert Systems with Applications, 252, Article 124070. https://doi.org/10.1016/j.eswa.2024.124070

[21] Hasan, M. M., Abdar, M., Khosravi, A., Aickelin, U., Lio, P., Hossain, I., Rahman, A., & Nahavandi, S. (2025). Survey on leveraging uncertainty estimation towards trustworthy deep neural networks: The case of reject option and post-training processing. ACM Computing Surveys, 57(9), Article 236. https://doi.org/10.1145/3727633

[22] Gao, S., Felgner, A., & Brand, U. (2025). Metrological-characteristics-based calibration of optical areal surface measuring instruments and evaluation of measurement uncertainty for surface texture measurements. Instruments, 9(2), Article 11. https://doi.org/10.3390/instruments9020011

[23] Brookshire, G., Kasper, J., Blauch, N. M., Wu, Y. C., Glatt, R., Merrill, D. A., Gerrol, S., Yoder, K. J., Quirk, C., & Lucero, C. (2024). Data leakage in deep learning studies of translational EEG. Frontiers in Neuroscience, 18, Article 1373515. https://doi.org/10.3389/fnins.2024.1373515

[24] Trombini, M., Maculotti, G., Maisano, D. A., Pagani, A., & Franceschini, F. (2025). Scaling photogrammetry: A comparative evaluation and metrological assessment across small- and meso-scale domains. Precision Engineering, 92, 124–140. https://doi.org/10.1016/j.precisioneng.2024.12.002

[25] Dias, L. C., Faria, A. D., Costa, P. B., & Amorim, I. L. R. (2025). Traceability of line scales using image processing. Acta IMEKO, 14(2), 1–5. https://doi.org/10.21014/actaimeko.v14i2.1972

[26] Sasse, L., Nicolaisen-Sobesky, E., Dukart, J., Eickhoff, S. B., Götz, M., Hamdan, S., Komeyer, V., Kulkarni, A., Lahnakoski, J. M., Love, B. C., Raimondo, F., & Patil, K. R. (2025). Overview of leakage scenarios in supervised machine learning. Journal of Big Data, 12(1), Article 135. https://doi.org/10.1186/s40537-025-01193-8

[27] Hallberg Szabadváry, J., Löfström, T., Johansson, U., Sönströd, C., Ahlberg, E., & Carlsson, L. (2025). Classification with reject option: Distribution-free error guarantees via conformal prediction. Machine Learning with Applications, 20, Article 100664. https://doi.org/10.1016/j.mlwa.2025.100664

[28] Reinke, A., Tizabi, M. D., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Kavur, A. E., Rädsch, T., Sudre, C. H., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A., Buettner, F., Cardoso, M. J., Cheplygina, V., Chen, J., Christodoulou, E., Cimini, B. A., … Maier-Hein, L. (2024). Understanding metric-related pitfalls in image analysis validation. Nature Methods, 21(2), 182–194. https://doi.org/10.1038/s41592-023-02150-0

[29] Silveira, P. S. P., Vieira, J. E., & Siqueira, J. d. O. (2024). Is the Bland-Altman plot method useful without inferences for accuracy, precision, and agreement? Revista de Saúde Pública, 58, Article 01. https://doi.org/10.11606/s1518-8787.2024058005430

[30] Zhao, Y., Chen, W., Xu, Z., Patwardhan, M., Wang, C., Liu, Y., Vig, L., & Cohan, A. (2025). AbGen: Evaluating large language models in ablation study design and evaluation for scientific research. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 12479–12491). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.611

[31] Ni, J., Zhao, H., Yang, Y., & Guo, D. (2026). Deep neural network calibration by reducing classifier shift with stochastic masking. Pattern Recognition, 177, Article 113217. https://doi.org/10.1016/j.patcog.2026.113217

[32] Cabezas, L. M. C., Otto, M. P., Izbicki, R., & Stern, R. B. (2025). Regression trees for fast and adaptive prediction intervals. Information Sciences, 686, Article 121369. https://doi.org/10.1016/j.ins.2024.121369

[33] Li, Z., Paolieri, M., & Golubchik, L. (2024). Inference latency prediction for CNNs on heterogeneous mobile devices and ML frameworks. Performance Evaluation, 165, Article 102429. https://doi.org/10.1016/j.peva.2024.102429