# Dataset review of exposed reinforcement in concrete bridges and challenges for automated damage detection in UAS-assisted bridge inspections

Erkki Tobias Bartczak*, Maarten Bassier, and Maarten Vergauwen

Department of Civil Engineering, Faculty of Engineering Technology,
Geomatics Research Group, KU Leuven, Gebroeders De Smetstraat 1, B-9000 Gent, Belgium
erkkitobias.bartczak@kuleuven.be

**Keywords:** UAS, bridge inspection, damage detection, concrete damage, exposed rebars.

## ABSTRACT

Corroding reinforcement leads to cross section loss and reduced structural capacity of concrete bridges. Detecting exposed rebars (ER) is crucial during bridge inspection to plan countermeasures early and prevent further corrosion. With advancements in deep learning, several public datasets derived from inspection imagery have been released to identify ER and other concrete damage automatically. At the same time, Uncrewed Aerial Systems (UAS) have become more capable of navigating even underneath the bridge deck. This combination holds promise to improve efficiency of bridge inspection methods, but obtained imagery differs from available datasets, featuring very small damages and complex backgrounds. To address this mismatch, this work reviews publicly available ER datasets, presents a UAS-based bridge inspection dataset for evaluating ER damage (UBID-ER-val), and quantifies similarities and differences between them. We train several YOLOv8 models on conventional inspection documentation images and benchmark the reviewed datasets, scoring **F2 = 0.229** at **S2DS**, **F2 = 0.430** at **CODEBRIM**, **F2 = 0.584** at **Dacl10k**, compared to **F2 = 0.505** at **UBID-ER-val**. We analyse factors influencing performance and find that tiled inference raises Recall (+0.166) but drastically reduces Precision (−0.309), while matching training and validation image resolution underperforms across all datasets (−0.061 to −0.129). The differences in best-performing combinations underscore the underlying domain shift that complicates practical deployment. As a practical outcome of this work, UBID-ER-val is made publicly available to enable objective benchmarking of ER detection models and to assess their reliability under field conditions.

a) columns b) girders c) abutments d) façades e) UAS imagery

**Figure 1: Variety of ER appearance and visual differences between documentation and UAS inspection images**. ER appears in various forms and on various structural elements but are typically central and quasi-orthogonal (a-d) while UAS imagery contains damages in the background, typically smaller and in various lighting conditions (e).

## 1. Introduction

Exposed reinforcement is a visible symptom of concrete cover loss that arises from diverse deterioration mechanisms, such as corrosion-induced cracking and spalling, construction imperfections, drainage issues inside the structure and impact or fatigue-related stresses. Once the protective alkaline environment is compromised and steel is exposed, corrosion can accelerate, section loss may follow, and bond between steel and concrete can degrade, with consequences for serviceability and ultimate state capacity. Early detection and maintenance of these damages is mandatory to prevent further degradation.

During bridge inspections, inspectors document ER alongside other common concrete defects such as cracking, spalling, delamination, efflorescence, corrosion staining, joint deterioration, leakage and seepage, often using close-range photography, notes, and sketches. Achieving full coverage of

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-767-2026 | © Author(s) 2026. CC BY 4.0 License.

the underdeck, pier caps, and other hard-to-reach locations typically requires lifts, scaffolding, or rope access, which is time-consuming, costly, and includes risks for inspectors. Over the last decade, UASs have been researched to overcome these challenges as they provide safe, rapid access to complex geometries, enable repeated capture under controlled flight plans (Bartczak et al., 2025) , supply imagery suitable for photogrammetric 3D reconstruction (Chen et al., 2019) and conduct subsequent bridge assessments (Seo et al., 2018; Morgenthal et al., 2019; Lin et al., 2021). However, manual data revision is time consuming and automated, robust concrete damage detection in UAS imagery would allow agencies to accelerate inspections and keep traceable, spatially indexed damage maps over time.

In the past decade, several general-purpose object-detection architectures have been developed, such as one-stage detectors i.e., **YOLOv8** (Jocher et al., 2023), **RetinaNet** (Lin et al., 2017), **EfficientDet** (Tan et al., 2020), two-stage detectors i.e., **Faster R-CNN** (Ren et al., 2015), **Cascade R-CNN** (Cai and Vasconcelos, 2018), **Mask R-CNN** (He et al., 2017) and transformer-based detectors such as **DETR** (Carion et al., 2020), but adaptation for concrete damage detection on bridges remains challenging (Li et al., 2023; Wu et al., 2023). While several concrete damage datasets have been established, they are commonly based on decades of documentation data of conventional bridge inspections, typically containing centrally framed, orthogonal photographs captured with various cameras (**Figure 1, a - d**). By contrast, in UAS inspection imagery ER is often small relative to the wide field of view, typically recorded under oblique viewing angles, and is embedded in cluttered backgrounds where shadows, stains, and surface roughness can mimic visual cues (**Figure 1, e**). This variability is critical for automated detection, as it defines what is visually available in an image, and thus what models can plausibly learn. This domain gap creates a major challenge and training dataset selection, data preparation, model training and inference parameter must be chosen carefully to obtain high performing detection models.

This paper addresses these limitations by reviewing ER datasets against UBID-ER-val, benchmarking YOLOv8 models across them, and analysing the influence of training and inference settings.

## 2. Related work

Automated bridge-damage detection has advanced rapidly with deep learning and growing image datasets. Although visual inspection datasets have been reviewed extensively (Bianchi and Hebdon, 2022), ER-specific datasets have not yet been compared at the level of image and annotation characteristics. **Table 1** summarizes the main public datasets containing ER. Their acquisition conditions, dataset splits, and evaluation protocols differ substantially, limiting direct comparison and transfer to cluttered UAS imagery. Brief definitions of common performance metrics are given in **Section 3**.

Among the public datasets, **CODEBRIM** (Mundt et al., 2019) remains one of the most widely used resources for concrete bridge defects, including i.e., crack, spallation, efflorescence, exposed bars and corrosion stain annotations. These images were captured at high resolution during conventional inspections and with UAS. While the dataset was annotated with bounding boxes, it is mainly recognised for the cropped classification variant. The dataset is constructed as a multi-label dataset, reflecting the frequent co-occurrence of defects e.g., exposed bars within spalled regions. Similarly, **BiNet** offers multi-label classification curated from bridge inspections, with classes including exposed reinforcing bars, cracks, spalling, and corrosion stains (Bukhsh et al., 2022). Hüthwohl et al. (2019) further expand this concept by systematically dividing classification tasks in multiple stages. Once an instance is classified as spalling, a subsequent stage determines whether ER is present and, if so, whether rust staining accompanies the exposure. The corresponding dataset **MCDS** is constructed from various sources containing documentation images from conventional inspections. However, these datasets do not contain any bounding box annotations and since they are derived from conventional inspection documentation images, contain viewpoint bias that complicates downstream detection tasks. To test the capabilities of classification models, Flotzinger et al. (2024) created a multi-label dataset **dacl1k** which is designed to stress diversity and realism but only contains image level annotations. Although defects are nominally centred, the dataset is markedly more challenging because targets occupy a small fraction of the frame and are embedded in visually complex backgrounds, increasing scale sensitivity and visual clutter. Building on their work, **dacl10k** introduces pixel annotations for semantic segmentation of bridge damages and structural components (Flotzinger et al., 2023). This dataset is not only much larger than previous datasets but also suitable for damage localization tasks. Similarly, **S2DS** also targets segmentation tasks but only provides 1024×1024 patches (Andres et al., 2022). It does not explicitly annotate ER as a separate class but contains many examples in the overall spalling label and as a patch-based

**Table 1: Public concrete damage datasets including ER**

| # | Dataset | Task | Classes | Resolution [px] | Images |
|---|---|---|---|---|---|
| 1 | Dacl1k: Real-World Bridge Damage Dataset (2024) | Classification | Crack, Spalling, Efflorescence, Rust/Corrosion, **Exposed reinforcement**, No damage | 300 to 6000 | 1,474 |
| 2 | Dacl10k: Semantic Bridge Defect Segmentation Benchmark (2024) | Segmentation | Crack, Spalling, Efflorescence, Corrosion, Seepage, Honeycomb, Voids, Wet areas, **Exposed reinforcement** and structural components (18 total classes) | 336 to 6000 | 9,920 |
| 3 | BiNet: Bridge Visual Inspection Dataset (2021) | Classification | Cracks, Spalling, Corrosion stains, **Exposed reinforcing bars** | 37 - 3729 | 3,588 |
| 4 | MCDS: Multi-Classifier for RC Bridge Defects (2019) | Classification | Crack, Efflorescence, Spalling, Surface scaling, Rust staining, **Exposed reinforcement** | 37 - 4145 | 3,607 |
| 5 | S2DS: Structural Defects Dataset (2022) | Segmentation | Background, Crack, **Spalling**, Corrosion, Efflorescence, Vegetation, Control point | 1024x1024 | 743 |
| 6 | CODEBRIM: Concrete Bridge Defect Image Dataset (2019) | Object detection & classification | Crack, Spalling, Corrosion, Efflorescence, **Exposed reinforcement**, No defect | 1200 - 6000 | 1,590 |
| 7 | GYU-DET: Multi-defect Beam Bridge Detection Dataset (2025) | Object detection | Crack, Spalling, Seepage, Honeycomb, **Exposed rebar**, Holes | 123 - 6000 | 11,123 |

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-767-2026 | © Author(s) 2026. CC BY 4.0 License.

dataset, it inherits a centrality bias. Lastly, **GYU-DET** focuses on object detection in high resolution images with six exclusive defect classes under varied conditions, including exposed rebars (Li et al., 2025).

Classification models have been trained and validated on these datasets and the results show that deciding which class a given crop out belongs to can achieve high accuracy. On the **CODEBRIM** classification dataset, Mundt et al. (2019) used meta-learned CNNs to reach multi-target test accuracies around $\mathbf{A_{multi} \approx 66 - 68\ \%}$, with per-class validation accuracy $\mathbf{A_{ER} \approx 94\ \%}$ for exposed bars. Similarly, the three-stage hierarchical classifier trained on **MCDS** achieved an average **F1-score ≈ 83.5 %** for multi-class patches, and a dedicated binary head achieved **F1-score ≈ 88.6 %** for ER damage (Hüthwohl et al., 2019). Comparable results are achieved by Bukhsh et al. (2022) on the **BiNet** dataset with a top-1 accuracy $\mathbf{A_{top\text{-}1} \approx 88\ \%}$ using a transfer-learned VGG16 architecture for multi-class classification including exposed bars. By contrast, Flotzinger et al. (2024) report a much lower multi-match accuracy dubbed Exact Match Ratio of **EMR ≈ 32.42%** on the much more difficult **dacl1k** dataset. However, Recall increases for the ER subset from $\mathbf{R_{ER} \approx 56.6\ \%}$ to $\mathbf{R_{ER} \approx 62.6\ \%}$ when training on **dacl1k** combined with meta datasets that include e.g. **CODEBRIM**, indicating modest gains from adding diverse training data.

Considering the object detection task, Hebbache et al. (2023) report **mAP@0.5 = 0.991** on **CODEBRIM** with their detection model **SMDD-Net**, which leverages saliency to focus on the most dominant parts in the image, achieving an $\mathbf{AP_{ER} \approx 0.90}$ for exposed bars. Next to their own models, they show strong baselines such as **YOLOX**-l Ge et al., 2021 and **YOLOR**-P6 (Wang et al., 2021) reach **mAP@0.5 ≈ 0.92** and **mAP@0.5 ≈ 0.89** compared to **mAP@0.5 ≈ 0.596** of **a YOLOv8** variant. However, they do not report clear dataset split, which complicates comparison and reproducibility. Similar results are achieved on a self-collected dataset by Xu et al. (2024) with **BD-YOLOv8s**, showing **mAP@0.5 = 0.862** and **F1 = 0.89**. Their modifications allow the network to focus on likely defect regions, adapt filters to image content, and preserve small details during up-sampling, improving performance (**ΔmAP@0.5 = +0.053** and **ΔF1 = 0.09**) over a baseline YOLOv8s model (**mAP@0.5 = 0.809**, **F1 = 0.80**). On the **GYU-DET** dataset, Li et al. (2025) trained standard pre-trained YOLOv11n model weights from Ultralytics with an overall **mAP@0.5 ≈ 0.57** and $\mathbf{AP@0.5_{ER} \approx 0.69}$ for ER.

The reported results come from validation on datasets dominated by close-range, centrally framed, high-resolution images where defects cover a large portion of the frame and the search space is limited. The performance results are not directly comparable, due to the lack of official fixed train/val/test splits, and although the strong results show general feasibility of damage detection model training, they do not imply robustness to UAS imagery. Taken together, literature shows that detectors can perform well in these constrained settings, but a public benchmark that tests under real-world UAS bridge inspections conditions is missing. To address this gap, we introduce **UBID-ER-val**, a UAS-based evaluation set for exposed reinforcement, and quantify how it differs from prior datasets.

## 3. Methodology

In this work, we introduce UBID-ER-val, a UAS-based evaluation dataset for ER damage, and comparatively validate baseline YOLOv8n models across curated datasets to analyse domain shift under realistic UAS inspection conditions (**Figure 2**).

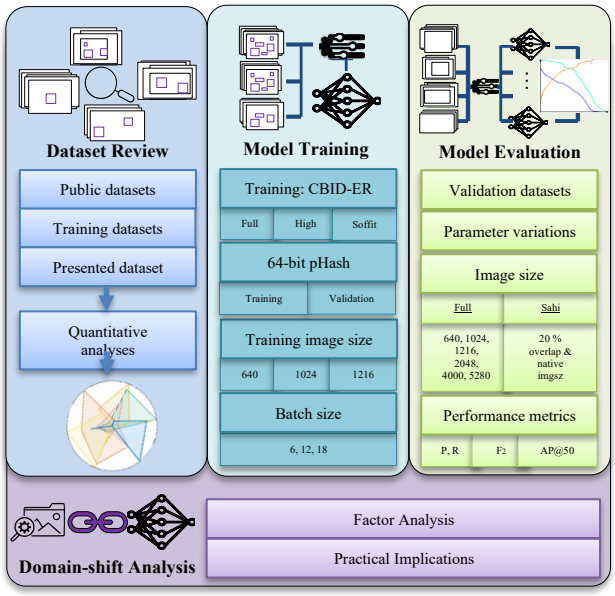


**Figure 2: Overview of the study workflow.** Reviewed datasets are curated and characterised, YOLOv8n models are trained on CBID-ER variants, validation settings are swept across inference modes and image sizes, and the resulting scores are used for domain-shift analysis.

First, we review and characterise the datasets used in this study (**Table 1**), including the training dataset, selected public ER datasets, and the validation dataset. Original images and YOLO annotations are filtered to ER samples and manually quality-controlled to remove low-quality or erroneous entries. We then compute per-annotation statistics: object size S_O (shorter bounding-box side in pixels), annotation area A (bounding-box area fraction of the full image), centrality C, and sharpness S_H. Centrality is implemented as a visibility score that combines (i) how close the box centre is to the image centre and (ii) how much the box overlaps the central image region, so large boxes near corners can still receive high centrality if they cover the centre. Sharpness is computed using Tenengrad gradient energy (Scharr-based) with higher values indicating stronger edge-content and therefore sharper appearance.

Second, we use the Ultralytics YOLOv8 framework with pretrained nano weights to train multiple baseline models on a conventional bridge inspection-based documentation image dataset, CBID-ER, provided by the Flemish Department of Transportation (DoT). To assess the relative importance of dataset size, target resolution, and domain similarity for transfer to UAS imagery, we train the models on three training variants, namely CBID-ER-Full representing the complete dataset, CBID-ER-High retaining only high-resolution annotations with SO > 150 px and CBID-ER-Soffit containing only underdeck views. Together, these variants test whether performance is driven more strongly by training set volume, target resolution, or similarity to the target inspection domain. To avoid leakage and ensure representative training and validation splits, we first group near duplicate images using perceptual hashing with 64-bit pHash and Hamming distance. These groups are then assigned to the splits while preserving the overall distribution of key characteristics, including image size, object scale, centrality, and sharpness and aiming for a 70/30 distribution between training and validation images. All models were trained for 200 epochs at image sizes of 640,

**Figure 3**: Excluded annotations of the CBID-ER-FULL training dataset with $S_O < 80$ px.


**Figure 4**: Representative samples of ER excluded datasets (Dacl1k, BiNet and MCDS) and included S2DS.

1024, and 1216, with batch sizes of 6, 12, and 18. The architecture was otherwise unchanged to provide a data-centric baseline. All training was conducted on three NVIDIA GeForce RTX 2080 Ti GPUs.

Third, the resulting models are validated on UBID-ER-val and the reviewed datasets comparatively to investigate performance differences. Since model performance highly depends on training settings as well as inference hyperparameters, we conduct a large grid search sweep validation, varying not only the trained models but also image size during validation from 640 px up to 5280 px with a confidence threshold of 0.1. Additionally, we use a slicing aided hyper inference (SAHI) module (Akyon et al., 2022) to perform lossless inference on tiled images, using 20 % overlap and the native training image size. We report the results in a parallel coordinates diagram to account for the high dimensionality of the experiment, visualizing the performance differences caused by training and validation parameter variations (**Figure 9**).
A predicted bounding box is counted as a true positive (TP) if it matches a ground-truth annotation with an Intersection-over-Union (IoU) $\geq 0.5$, while unmatched predictions are false positives (FP) and missed ground-truth boxes are false negatives (FN). Precision is $P = \frac{TP}{TP+FP}$ and recall is $R = \frac{TP}{TP+FN}$. Since a high recall may be achieved by wild guessing and therefore very low precision, commonly the average precision (AP) is used at an IoU of 50 % and reported as the median $mAP@0.5$ in case of multiple classes. Instead of treating precision and recall as equally important, we use the F2 score to give priority to higher recall $F_\beta = \frac{(1+\beta^2)PR}{\beta^2 P+R}$ with $\beta = 2$.

Finally, we aggregate the validation metrics (precision, recall, AP@0.5, F2) and relate them to the dataset characteristics (image size, relative box area, centrality, sharpness). By comparing score distributions, top-performing runs, and required inference settings across datasets against the measured characteristic distributions, we identify which dataset properties most plausibly explain why identical model weights behave differently, thereby quantifying the domain shift.

## 4. Dataset review

In the following we describe the reviewed datasets qualitatively and present their main characteristics. We start chronologically with our training dataset CBID-ER, the reviewed public datasets, present our validation dataset UBID-ER-val and quantify key dataset characteristics comparatively.

**CBID-ER-Full** consists of conventional bridge inspection documentation photographs supplied by the Flemish department of transportation (DoT), spanning decades of inspections with diverse cameras and practices. The raw archive contains 12,624 images with inspector overlays (red boxes/circles/arrows) from which 6068 images passed general image quality checks. After approx. 500 manually annotated images, a preliminary YOLOv8 model assisted in auto-labelling the dataset. This process leads to frequent false annotations and especially difficult to notice few pixels small labels. These residual artefacts were mitigated by reviewing context-free crops ($S_O < 80$ px) and thus removing 453 unsuitable boxes (**Figure 3**). Finally, the pre-labelled dataset was reviewed over multiple rounds and by multiple inspectors to decide which images were retained in the final dataset, moving from initially liberal inclusion to stricter curation in later passes.
The prepared training dataset comprises 3553 images with 8808 ER instances. Since the dataset is derived from documentation images, the framing is relatively central compared to other datasets with a median centrality of $C \approx 0.24$ (**Figure 1**). In addition to CBID-ER-Full, two derived training variants are considered in later experiments. CBID-ER-High retains only annotations with $SO > 150$ px and contains 2208 images with 3427 annotations. CBID-ER-Soffit retains only underdeck view images and contains 207 images with 659 targets. The annotations in the dataset vary widely. Some images focus tightly on the steel bar while others include the surrounding spalled region. In multi-bar scenes, annotations sometimes delineate individual bars and sometimes group several bars into a single region, reflecting the inherent ambiguity of the task. The imagery itself is extremely heterogeneous and ranges from very high-resolution images with crisp detail to blurred captures. It covers a broad set of structural components, including insides of box girders with flashlights and underground structures as well as walls, pillars, abutments and the underdeck views (**Figure 1, a-d**).

**Dacl1k** is a multi-target classification dataset with 195 ER-positive images from conventional inspections. Although many images contain sufficient scene context for localisation, only 154 were retained after revision, leaving the dataset too small for object-detection experiments.

**Dacl10k** is originally a semantic segmentation benchmark. Although 705 images contain ER classes, the masks are very tight around the bars, so we query for instances that are surrounded by a spalling to increase visual cues and convert the labels to bounding boxes. We manually review the dataset while correcting occasional class errors, removing image sized bounding boxes, very small or ambiguous instances and occasionally reducing the bounding box sizes, retrieving in total 481 images with 783 ER instances.
The dataset is highly diverse in respect of lighting conditions, structural components and image and target resolution, reaching from extremely detailed high-resolution close-up captures, to small, blurry ER instances in the background. In majority, the views are rather orthogonal, contain little

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-767-2026 | © Author(s) 2026. CC BY 4.0 License.

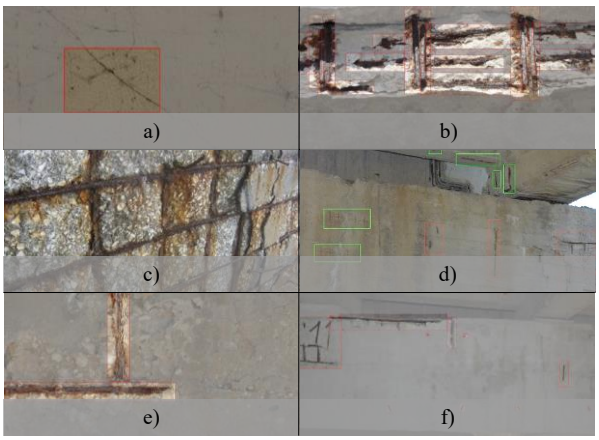


**Figure 5:** CODEBRIM samples of ER annotations including a) Crack annotated as ER, b) high annotation overlap, c) full sized annotation, d) missed ER and e-f) included data.

background and high centrality. Overall, Dacl10k is a relatively large and challenging source for detection experiments with various scenes and various ER appearances and frequent small targets.

**BiNet** and **MCDS** are multi-label classification datasets of tightly cropped defect images. Despite their variety, they are excluded because they lack scene context and are therefore unsuitable for detection-oriented evaluation. BiNet additionally contains up to 21 near-duplicate views of the same damage (**Figure 4**).

**S2DS** is a segmentation dataset of 1024 by 1024 patches. We use only the ER subset, convert the masks to bounding boxes and proceed with manual review. Near duplicates are removed and cases that show only corrosion staining without visible bars are excluded, leaving 196 images with 696 ER instances. Annotations follow the bars rather than the surrounding spalled concrete, which yields narrow boxes and many very small targets. Because this dataset comprises extreme close-up views derived from segmentation maps, frames show high instance density (up to 19 per image). Box centres are widely distributed rather than central, with mostly orthogonal views against simple backgrounds. However, the derived fine-grained ER bounding boxes are still small in the overall image which makes this dataset suitable for object detection with emphasis on usually small ER appearances in high-resolution and clean, low background noise scenes.

**CODEBRIM** was primarily released as a multi-label multi-target classification, while we use the original full-sized images after curating 446 ER containing images and bounding box annotations. During review, we found frequent issues in the ER subset as depicted in **Figure 5**, including (a) occasional class inaccuracies, (b) overlapping boxes that describe the same region, (c) boxes that span almost the full frame which are therefore suited for classification rather than object detection and (d) missed annotations, leaving 311 images with 1001 annotations. A key characteristic is that most of the images are orthogonal with targets typically being dominant in the foreground (e), while few images contain complex scenes (f). Damages are recorded in very high quality and size in the images, and the background often lacks detail, which simplifies a single object detection task. In scenes with several exposed rebars the annotation practice is inconsistent, sometimes annotating large ER regions as one bounding box and sometimes as multiple instances and often even both in the same image, which is common challenge in dataset preparation. In sum, CODEBRIM provides high-detail images

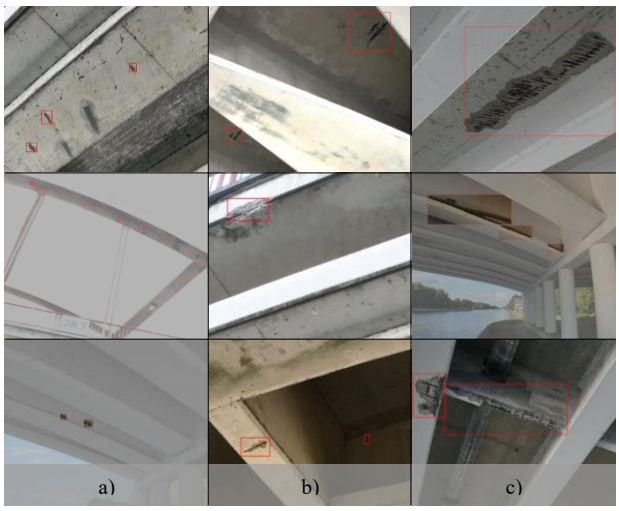


**Figure 6**: UBID-ER-val samples of column a) small ER often in the background, b) medium sized, c) large and complex instances.

of ER and might be useful for training and validation but needs careful revision before use.

**GYU-DET** is a recent full-frame object detection dataset, but we exclude it from experiments because data quality is insufficient for reliable training or evaluation. During revision, we observed heavy blur, rotation-related box misalignment, dense duplicate annotations, and clearly false ER labels suggesting semi-automatic labelling without adequate review. In addition, many ER-labelled images depict construction activities or loose wire and tie steel rather than genuine exposed reinforcement on concrete, and the dataset also includes CODEBRIM images. Despite these issues, parts of the imagery still contain valuable high-resolution views of complex scenes, but the dataset would require comprehensive reannotation and stricter class definitions before use.

**UBID-ER-val** comprises 250 selected images obtained from UAS-assisted inspections on 18 canal concrete bridges, carried out on 6 days in 2025 using a DJI Mavic 3 Enterprise. The data acquisition was aided by our flight planning tool ORBIT, which generates waypoint flight routes specifically designed for bridge inspections, that cross underneath the main spans, while assuring safe and reproducible flight mission (Bartczak et al., 2025). The UAS automatically follows the planned flight routes, while the pilot remains in control of the gimbal and viewing direction. The dataset contains a uniform image size of 5280 by 3956 pixels with lens distortion and vignette. Due to challenging lighting conditions, various exposure settings were used to assure bright imagery during underdeck views, leading to frequent over exposure. The quality of the images and size of the annotations is mainly influenced by the flight offset and viewing angle to the bridge and maximum upward tilt of 35°. Damage positions were not known a priori, and the camera was controlled manually during flight with the intent to maximize bridge coverage rather than detailed damage documentation.
Given the data acquisition scenario, the scenery contains highly complex backgrounds due to perspective angle and highly cluttered texture on concrete surfaces i.e., spider webs and generally aged surfaces. Additionally, the images may also include other damages, such as corrosion staining and efflorescence (not annotated in this work) and already repaired damage patches which show darker concrete surfaces, further adding complex details to the scenes (**Figure 6**). The ER annotations are predominantly small, highly off-centre, in the

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-767-2026 | © Author(s) 2026. CC BY 4.0 License.

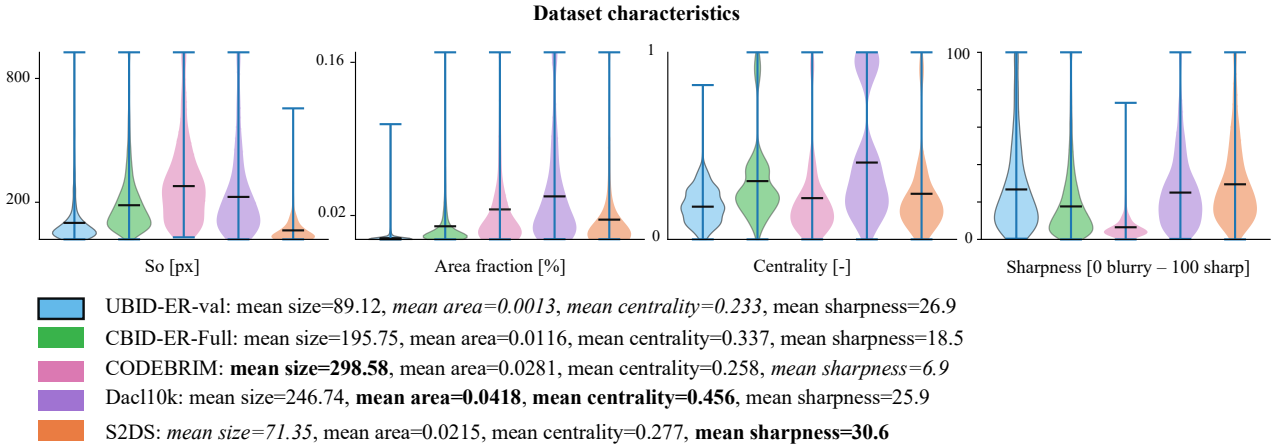


**Figure 7: Comparison of characteristics across datasets.** The size of the annotations in UBID-ER-val differs drastically from the public and training datasets with lowest mean area and centrality values.

background and on surfaces that are not perpendicular to the camera view, making this an especially challenging small-object detection dataset. We annotated damages even in the far background to analyse the limits of the detection models. Other ER appearances include medium sized individual instances as well as large, complex patches and damage regions that were cleaned and prepared for repair measures. In contrast to previous datasets, UBID-ER-val reflects realistic UAS bridge inspection conditions and challenges object detection models with real-world background complexity, off-centre targets, and small object scales.

The review shows that only few public datasets are suitable for ER detection benchmarking and that key characteristics differ across datasets (**Figure 7**). While S2DS contains the smallest annotations in terms of absolute box sizes, the strongest difference between the datasets was found in the area fraction in UBID-ER-val. In contrast to our perception, the UBID-ER-val annotation centrality and sharpness metrics are comparable between the datasets.

## 5. Results

**Figure 8** shows that input resolution was the dominant training hyperparameter within the tested range. Models trained at *imgsz = 640* did not learn effectively, while performance improved substantially at higher resolutions. The best training result was obtained at *imgsz = 1024*. A further increase did not provide a consistent additional benefit. The effect of higher batch sizes was small and inconsistent across the few tested combinations.

The validation sweep shows strong dataset dependence. The best performances were obtained on Dacl10k with $\mathbf{F2_{top1} = 0.584}$ and on UBID-ER-val with $\mathbf{F2_{top1} = 0.505}$, while lower values were obtained on CODEBRIM with $\mathbf{F2_{top1} = 0.430}$ and on S2DS with $\mathbf{F2_{top1} = 0.229}$, as shown in **Figure 9**. Within each dataset, the spread among the top five runs was modest, with $\mathbf{\Delta F2_{top5} = 0.053}$ for Dacl10k, **0.045** for UBID-ER-val, **0.016** for CODEBRIM, and **0.015** for S2DS.

Inference settings strongly affected model performance. UBID-ER-val achieved its best result with full frame inference at ***imgsz* = 4000** px, whereas Dacl10k performed best at ***imgsz* = 640** px. SAHI did not improve Dacl10k, where the best SAHI run remained **ΔF2 = 0.128** below the top result. In UBID-ER-val, the best SAHI configuration ranked second and increased recall by **ΔR = 0.166** relative to the top run, but reduced precision by **ΔP = 0.309** and remained **ΔF2 = 0.025** below the best full frame result. In the top 5 of S2DS, 3 runs use SAHI and the other various *imgsz*, without showing any trend. Higher training ***imgsz* = 1216** was beneficial in most cases, except for Dacl10k, where imgsz = 1024 scored **ΔF2 = 0.034** higher in the top configuration. Matching training and validation image size was not generally optimal. It ranked fourth in Dacl10k with **ΔF2top1 = -0.051** and twelfth in UBID-ER-val with ($\mathbf{\Delta F2_{top1} = -0.129}$).

The **training dataset variations** (High, Soffit) did not show any positive impact on the UBID-ER-val, decreasing F2-score drastically by $\mathbf{\Delta F2_{top1} = -0.076}$ and $\mathbf{\Delta F2_{top1} = -0.146}$, respectively. In contrast, for Dacl10k, CBID-ER-High represents 4 of the 5 top-performing runs and increases **ΔF2 = +0.133** compared with the full dataset variant under otherwise identical top parameters.

## 6. Discussion

The dataset review highlights recurring challenges in ER dataset preparation, including (i) undefined or unrepresentative validation splits, (ii) repeated views of the same physical damage, (iii) leakage risks during train/validation partitioning, (iv) annotation errors and inconsistencies, (v) semi-automatic labelling artefacts, and (vi) ambiguity in whether very small or far-background instances should be retained. These issues are evident in Dacl1k and Dacl10k, whose original validation splits contain only 24 and 67 ER images, and in BiNet, where repeated views of the same damage inflate nominal dataset size without equivalent growth in effective variance. Additional recurring problems include missed and false labels, rotation-

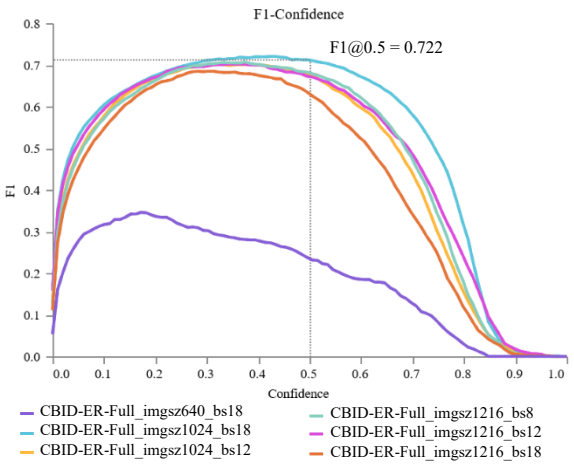


**Figure 8: Model training results.** Validated on CBID-ER-Full.

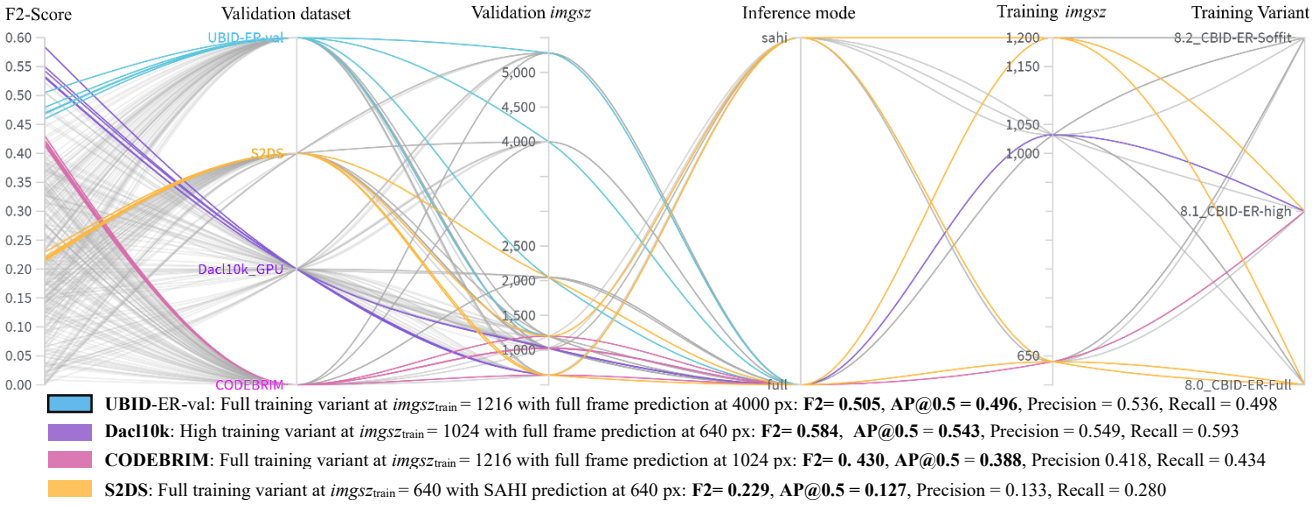


**Figure 9: Results of validation sweep.** The 5 combinations for each dataset are highlighted.

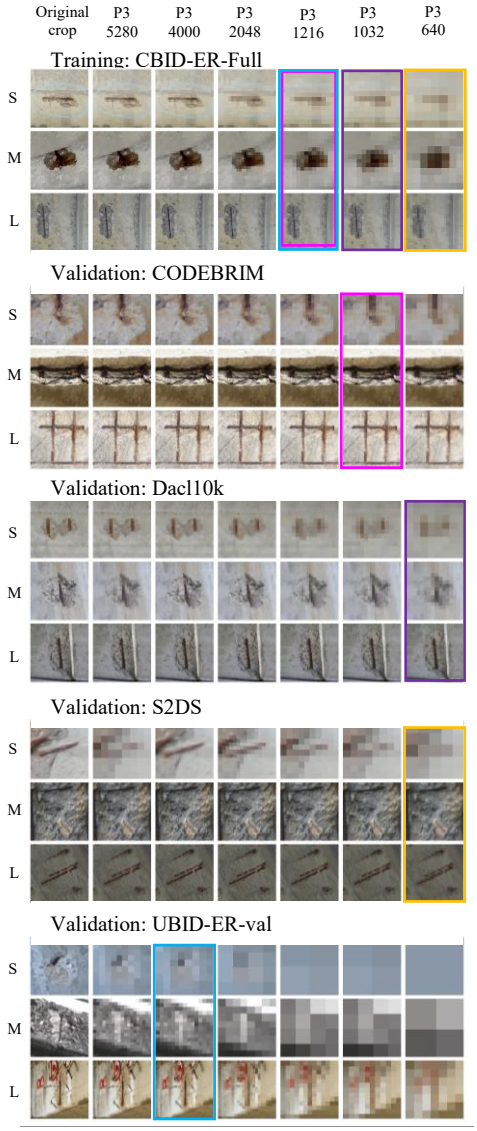


**Figure 10: Visualization of learnable visual features.** Rows with small, to large ER per dataset. Columns: original crop and P3-head appearance at input resolutions. Coloured borders mark the train/validation resolution combination of the best sweep result for each dataset.

related box misalignment in datasets such as CODEBRIM and GYU-DET, and the need for repeated review cycles, crop-out inspection, and explicit split design to avoid leakage and improve annotation quality. Overall, the review shows that large unbiased datasets remain unavailable for reliable assessment of model robustness and deployment readiness in UAS-assisted bridge inspection.

The validation sweep reveals dataset-dependent performance differences, with Dacl10k and UBID-ER-val scoring highest and CODEBRIM and S2DS trailing behind. This pattern is consistent with the reviewed characteristic distributions. Dacl10k aligns most closely with CBID-ER-Full in terms of object size, centrality, and sharpness, compared to the other datasets, which reflects the achieved F2-scores (**Figure 7** , **Figure *9***). At the same time, UBID-ER-val differs most strongly by its much smaller object area fraction within the full image, with ER appearing in visually complex high-resolution UAS scenes and often away from the image centre. During inference, the original image is reduced to the selected imgsz, after which YOLOv8’s P3 head operates at stride 8, so only limited ER detail remains recoverable unless large inference sizes are used. This is reflected in the best UBID-ER-val result at 4000 px, whereas Dacl10k already peaks at 640 px because it is less affected by full-frame scale loss (**Figure 10**).
The results show that model optimization is highly multi-dimensional for detecting ER damage in UAS imagery. While performance optimization strategies focus mainly on hyperparameter tuning of existing models (Chung et al., 2024), architectural extensions (Hebbache et al., 2023; Xu et al., 2024) and model pruning (Guan and Li, 2024), the present sweep shows that performance also depends strongly on inference image size and tiling strategy, even for fixed model weights. Taken together, these findings quantify a scale and context driven domain shift and show that ER localization in UBID-ER-val is fundamentally a small-object detection problem in large field of view imagery.

Visual review of the predicted ER damage highlights a limitation of the present evaluation. In complex ER regions with multiple bars, predictions often are practically correct, while still failing to match the annotated number and extent (i.e., IoU) of boxes, which lowers both precision and recall. In addition, all results are reported at a fixed global confidence threshold of 0.1 without dataset specific threshold optimization or post processing. The reported scores should therefore be interpreted as baseline results rather than fully optimized performance.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-767-2026 | © Author(s) 2026. CC BY 4.0 License.

## 7. Conclusion

This work offers a differentiated review of datasets for automated detection of exposed reinforcement for bridge inspections and highlights challenges in dataset preparation, training and validation. We introduce UBID-ER-val, a UAS-based evaluation set that reflects real, reproducible inspection conditions with small, off-centre targets in visually complex backgrounds. Using baseline YOLOv8n models trained on DoT provided documentation imagery, our cross-dataset validation shows pronounced domain shift with F2-scores ranging from **0.229** (S2DS) and **0.430** (CODEBRIM) to **0.584** (Dacl10k) and **0.505** on UBID-ER-val. The experiments highlight that inference choices matter as much as training settings and are dataset specific, i.e., validation image size choices dominate performance over training hyperparameters, and optimal settings vary from 640 to 4000 px, depending on the validation dataset.

Building on these findings, future work should optimize models for UBID-ER-val and analyse instance-level errors in relation to flight planning and image acquisition. The dataset preparation process also highlighted the need for explicit annotation guidelines, for example for grouped versus individual bars, inclusion of surrounding spalled regions, and treatment of very small or far-background instances. Future multi-view datasets could further benefit from polygon-based annotations, which can be reprojected more faithfully across views and better respect occlusions when 3D models are available and a refined matching logic for practical evaluation. These steps can improve the reliability of automated ER detection models in UAS-assisted bridge inspections and support practical deployment.

**Data availability.** The corrected ER subset labels produced during our revision of Dacl10k, S2DS, and CODEBRIM, together with the code used to compute dataset characteristics, are available on GitHub. UBID-ER-val is publicly available as a Kaggle benchmark competition to support further improvements in automated ER detection for UAS-assisted bridge inspection.

## References

Akyon, F.C., Altinuc, S.O., Temizel, A., 2022. Slicing Aided Hyper Inference and Fine-tuning for Small Object Detection. *IEEE International Conference on Image Processing.* [CrossRef]

Andres, B., Bernard, F., Cremers, D., Frintrop, S., Goldlücke, B., Ihrke, I., 2022. Image-Based Detection of Structural Defects Using Hierarchical Multi-scale Attention. *DAGM German Conference on Pattern Recognition*.

Bartczak, E.T., Bassier, M., Vergauwen, M., 2025. ORBIT: Optimized Routing for Bridge Inspection Toolkit. An open-source UAS flight path planning tool for comprehensive bridge inspections under realistic constraints. *ISPRS Archives*. [CrossRef]

Bianchi, E., Hebdon, M., 2022. Visual structural inspection datasets. *Automation in Construction*. [CrossRef]

Bukhsh, Z.A., Anžlin, A., Stipanović, I., 2022. BiNet: Bridge Visual Inspection Dataset and Approach for Damage Detection. *Proceedings of the 1st Conference of the European Association on Quality Control of Bridges and Structures*. Springer International Publishing, pp. 1027–1034. [CrossRef]

Cai, Z., Vasconcelos, N., 2018. Cascade R-CNN: Delving Into High Quality Object Detection. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [CrossRef]

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-End Object Detection with Transformers. [CrossRef]

Chen, S., Laefer, D.F., Mangina, E., Zolanvari, S.M.I., Byrne, J., 2019. UAV Bridge Inspection through Evaluated 3D Reconstructions. *Journal of Bridge Engineering*. [CrossRef]

Chung, S.-W., Hong, S.-S., Kim, B.-K., 2024. Hyperparameter Tuning Technique to Improve the Accuracy of Bridge Damage Identification Model. *Buildings*. [CrossRef]

Flotzinger, J., Rosch, P.J., Braml, T., 2023. dacl10k: Benchmark for Semantic Bridge Damage Segmentation. *CVPR.* [CrossRef]

Flotzinger, J., Rösch, P.J., Oswald, N., Braml, T., 2024. dacl1k: Real-world bridge damage dataset putting open-source data to the test. *Engineering Applications of Artificial Intelligence* [CrossRef]

Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J., 2021. YOLOX: Exceeding YOLO Series in 2021. *CVPR*. [CrossRef]

Guan, B., Li, J., 2024. Lightweight detection network for bridge defects based on model pruning and knowledge distillation. *Structures* [CrossRef]

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. *CVPR.* [CrossRef]

Hebbache, L., Amirkhani, D., Allili, M.S., Hammouche, N., Lapointe, J.-F., 2023. Leveraging Saliency in Single-Stage Multi-Label Concrete Defect Detection Using Unmanned Aerial Vehicle Imagery. *Remote Sensing*. [CrossRef]

Hüthwohl, P., Lu, R., Brilakis, I., 2019. Multi-classifier for reinforced concrete bridge defects. *Automation in Construction*. [CrossRef]

Jocher, G., Chaurasia, A., Qiu, J., 2026. Ultralytics YOLO (Version 8.4.30) [Computer software]. Zenodo. [CrossRef]

Li, R., Yu, J., Li, F., Yang, R., Wang, Y., Peng, Z., 2023. Automatic bridge crack detection using Unmanned aerial vehicle and Faster R-CNN. *Construction and Building Materials*. [CrossRef]

Li, R., Zhao, L., Wei, H., Hu, G., Xu, Y., Ouyang, B., Tan, J., 2025. Multi-defect type beam bridge dataset: GYU-DET. *Scientific data*. [CrossRef]

Lin, J.J., Ibrahim, A., Sarwade, S., Golparvar-Fard, M., 2021. Bridge Inspection with Aerial Robots: Automating the Entire Pipeline of Visual Data Capture, 3D Mapping, Defect Detection, Analysis, and Reporting. *Journal of Computing in Civil Engineering*. [CrossRef]

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal Loss for Dense Object Detection. *CVPR.* [CrossRef]

Morgenthal, G., Hallermann, N., Kersten, J., Taraben, J., Debus, P., Helmrich, M., Rodehorst, V., 2019. Framework for automated UAS-based structural condition assessment of bridges. *Automation in Construction*. [CrossRef]

Mundt, M., Majumder, S., Murali, S., Panetsos, P., Ramesh, V., 2019. Meta-learning Convolutional Neural Architectures for Multi-target Concrete Defect Classification with the COncrete DEfect BRidge IMage Dataset. *CVPR*. [CrossRef]

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *CVPR.* [CrossRef]

Seo, J., Duque, L., Wacker, J., 2018. Drone-enabled bridge inspection methodology and application. *Automation in Construction* [CrossRef]

Tan, M., Pang, R., Le, Q.V., 2020. EfficientDet: Scalable and Efficient Object Detection. *CVPR*. [CrossRef]

Wang, C.-Y., Yeh, I.-H., Liao, H.-Y.M., 2021. You Only Learn One Representation: Unified Network for Multiple Tasks. *CVPR*. [CrossRef]

Wu, Y., Han, Q., Jin, Q., Li, J., Zhang, Y., 2023. LCA-YOLOv8-Seg: An Improved Lightweight YOLOv8-Seg for Real-Time Pixel-Level Crack Detection of Dams and Bridges. *Applied Sciences*. [CrossRef]

Xu, W., Li, X., Ji, Y., Li, S., Cui, C., 2024. BD-YOLOv8s: enhancing bridge defect detection with multidimensional attention and precision reconstruction. *Scientific reports* [CrossRef]