STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **14**, December 2025, pp 3833–3856.
Published online in International Academic Press (www.IAPress.org)

Advancing Structural Health Monitoring with Lightweight Real-Time Deep Learning-Based Corrosion Detection

Safa Abid[1,*], Mohamed Amroune[2], Issam Bendib[1], Chams Eddine Fathoun[1]

[1]*Laboratory of Mathematics, Informatics and Systems (LAMIS), Echahid Cheikh Larbi Tebessi University, Tebessa, Algeria*
[2]*National School of Nanoscience and Nanotechnology, Sidi Abdallah, Algeria*

Abstract Structural Health Monitoring (SHM) is essential for preserving the safety and service life of industrial infrastructure. Corrosion, in particular, remains one of the most critical degradation phenomena, demanding timely and accurate detection to prevent structural failures and costly downtime. This study proposes a lightweight, real-time corrosion detection framework tailored for SHM applications. The framework integrates design elements inspired by the latest YOLOv11 and YOLOv12 architectures while incorporating task-specific optimizations for detecting small, irregular corrosion patterns under diverse environmental conditions. Two curated datasets, augmented with domain-specific transformations, are used to enhance model robustness and generalization. Comprehensive benchmarking against previous YOLO versions (YOLOv3, YOLOv5, YOLOv7, YOLOv8) demonstrates that our optimized YOLOv11m configuration achieves up to 7.7% improvement in mAP@50 and 12.1% in mAP@50–95 over YOLOv8m, while the YOLOv12s variant offers a competitive accuracy–speed trade-off. These findings highlight the potential of the proposed approach for deployment in edge-based SHM systems for real-time industrial monitoring.

Keywords Automated Corrosion Detection , YOLOv11 , YOLOv12 ,Structural Health Monitoring , Deep Learning , Lightweight

DOI: 10.19139/soic-2310-5070-2919

1. Introduction

Corrosion is a wide-reaching issue with deleterious effects on the structural integrity and service life of large-scale industrial infrastructures such as pipelines, bridges, and offshore platforms. When left unidentified, corrosion can finally lead to disastrous failures involving excessive financial losses, grave ecological concerns, and soul injuries. The traditional methods of detecting corrosion, such as visual inspections and manual testing, at occasions demand much manpower, take time, and rely greatly on human subjectivity. As the audiences in choices of safety and efficiency have become large, the current demand for some automated, reliable, and real-time products for observing corrosion detection is increasing. Deep learning algorithms have become indispensable for automating infrastructure inspections, significantly improving accuracy and efficiency in structural assessments. These technologies are now widely utilized across multiple sectors, including tunnels, roads, bridges, dams, and railways [20,21,22 ,23,24,25,26,27]. Substantial advancements have been made in integrating deep learning into this field, particularly in addressing challenges related to object detection. Among these, the You Only Look Once (YOLO) family of models is one of the leading frameworks for real-time object detection, possessing the advantages of high speed and accuracy. The latest version of these models, YOLOv12 and YOLOv11, further improves upon their older counterparts by introducing certain architectural innovations that allow them to maintain

*Correspondence to: Safa Abid (Email: safa.abid@univ-tebessa.dz). Laboratory of Mathematics, Informatics and Systems (LAMIS), Echahid Cheikh Larbi Tebessi University, Tebessa, Algeria.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

high detection performance while still keeping computation efficient. This makes YOLOv12 and YOLOv11 especially suited to the kind of application that structural health monitoring method requires: real-time processing and high accuracy. In addition to YOLOv11 and YOLOv12, previous versions such as YOLOv8[1] and YOLOv5[2] have also demonstrated strong capabilities in object detection tasks. YOLOv8, being one of the more recent iterations, brings enhanced architectural refinements that improve accuracy and efficiency. Meanwhile, YOLOv5 has been widely adopted in industry due to its balance of speed and performance. Comparing these three YOLO models allows for a comprehensive evaluation of how advancements in the YOLO family contribute to corrosion detection performance. This study examines and contrasts the performance of various YOLO models YOLOv12, YOLOv11, YOLOv8, YOLOv7, YOLOv5, and YOLOv3 in detecting corrosion within industrial infrastructures. Two datasets are utilized: the first, introduced in [12], facilitates the comparison between YOLOv8 and YOLOv5, while the second, from [17], supports an evaluation of YOLOv7, YOLOv5, and YOLOv3. Each model undergoes independent training and assessment on its respective dataset, with performance measured in terms of detection accuracy, inference speed, and resilience to different corrosion levels. The goal is to identify the most effective model for real-time corrosion detection, providing valuable insights into their respective advantages and trade-offs. Our contribution is not merely applying YOLOv11/YOLOv12 but tailoring their architectures, data handling, and optimization strategies to address the specific challenges of corrosion detection in Structural Health Monitoring (SHM), ensuring robustness across diverse surface textures, lighting variations, and degradation patterns. Our main contributions are summarized as follows:

Performance Analysis: We demonstrate the improved performance of both detection and segmentation using YOLOv11 and YOLOv12, where architecture refinements, dataset preprocessing, and targeted training strategies enhance robustness under varied corrosion conditions compared to earlier YOLO models

Comprehensive Benchmarking: A detailed comparative study is conducted across several state-of-the-art object detection models, including YOLOv12, YOLOv11, YOLOv8, YOLOv5, YOLOv7, and YOLOv3, focusing on their strengths and limitations in corrosion detection.

Dual-Dataset Evaluation: Two different datasets are utilized to ensure a robust and fair evaluation of detection and segmentation performance across all models.

Latest YOLO Variants Integration: The latest YOLO iterations: YOLOv11 and YOLOv12 are thoroughly evaluated to assess their real-time performance and effectiveness in detecting and segmenting corrosion.

The remainder of this paper is organized as follows: The State of the Art section reviews recent developments in object detection and corrosion detection methods. The Methodology section details advances in real-time corrosion detection, including YOLOv11 and YOLOv12 architectures, dataset preparation, and training procedures. The Results and Comparative Analysis section presents evaluation metrics, quantitative and qualitative results, and performance comparisons with existing YOLO models. The Inference Test and Variant Analysis section explores YOLOv11 and YOLOv12 variant performance and the impact of dataset differences. Finally, the Conclusion summarizes the findings and outlines directions for future research.

2. Overview of AI-Based Approaches for Corrosion Monitoring and Detection

2.1. Advancements in object detection models

2.1.1. ***Overview of object detection models*** The field of object detection has undergone significant evolution over the past two decades, transitioning from traditional feature-based methods to modern deep learning architectures. Early approaches, such as Haar cascades and Histogram of Oriented Gradients (HOG) [3] combined with Support Vector Machines (SVM) [4], relied on handcrafted features and shallow classifiers. While these methods achieved moderate success in constrained environments (e.g., face detection with Haar cascades), their performance was limited by their inability to generalize to complex, real-world scenarios with diverse object appearances and backgrounds. The advent of deep learning revolutionized the detection of objects with the R-CNN [5] being the tipping point. R-CNN applied region proposal algorithms for detecting candidate regions of the object, with the implementation of CNN-based classification. But its computational inefficiency gave birth to successors like the Fast R-CNN [6] (which shared the convolutional features across proposals) and the Faster R-CNN [7] (which

utilized region proposal networks for end-to-end training). Though these models improved accuracy, their multi-stage pipelines came at the expense of being non-real-time. To address the problem of limited speed, single-shot detectors emerged. Single Shot MultiBox Detector (SSD) [8] and You Only Look Once (YOLO) eliminated the use of region proposals through single-pass detection of the bounding boxes as well as the probabilities of the classes. YOLO revolutionized the use of real-time detection specifically through the use of grids to split the image and process them as wholes. Later versions (e.g., YOLOv3 [9], YOLOv5 [2], YOLOv10 [10], YOLOv11 [11], YOLOv12 [28]) also included innovations like the use of anchor-free detection, feature pyramid networks, attention mechanisms as well as light architecture, optimizing the accuracy-speed compromise.

Accuracy vs. Speed Trade-offs

The choice of object detection models often hinges on balancing accuracy and inference speed:

Two-stage detectors like Faster R-CNN have higher accuracy but incur expensive computational overhead, making them unsuitable for real-time applications.

Single-stage detectors like YOLO sacrifice marginal accuracy in exchange for significant speed improvements, making them ideal for time-sensitive scenarios such as video processing and industrial monitoring.

This trade-off is critical in structural health monitoring, where real-time corrosion detection must process high-resolution images or video streams without latency. YOLO's architecture, particularly its latest iterations (YOLOv11 and YOLOv12), addresses this challenge by optimizing both detection precision and computational efficiency, making it a compelling choice for industrial applications.

2.1.2. ***The Shift Toward Real-Time Detection Systems*** The growing demand for real-time processing in industrial settings has led to the adoption of lightweight, high-speed models like YOLOv11. Unlike traditional methods that process data offline, real-time systems offer several advantages:

• Immediate decision-making: Defects or faults can be quickly identified and corrected.

• Scalability: Efficient models can handle multiple sensors simultaneously, processing high-resolution images or video streams.

• Integration with edge devices: Running YOLOv11 or YOLOv12 on edge devices like drones and IoT cameras reduces latency and bandwidth costs.

For infrastructure monitoring, timely detection is particularly crucial. The Corrosion, for instance, develops gradually but can compromise structural integrity if left unnoticed. YOLOv11 and YOLOv12's balance between speed and accuracy makes it an ideal tool for continuous monitoring of aging infrastructure, where delays in detection could result in costly repairs or even catastrophic failures.

2.2. Corrosion detection methodologies

Recent advancements in deep learning have significantly improved corrosion detection capabilities across various domains.

[12] A comparative analysis of YOLOv5 and YOLOv8 for corrosion segmentation was conducted using three diverse datasets, evaluated through precision, recall, F1-score, and mean average precision (mAP). The study demonstrated that YOLOv8 consistently outperformed YOLOv5 in both segmentation accuracy and computational efficiency. Visual assessments further underscored YOLOv8's superior handling of complex corroded surfaces, though challenges persisted with overlapping bounding boxes in larger datasets. These findings position YOLOv8 as a more robust solution for real-world corrosion detection applications.

[14] An evaluation of DeepLabv3 and DeepLabv3+ for corrosion detection highlighted the critical role of model configuration and dataset augmentation in enhancing segmentation performance. The study revealed that DeepLabv3+ achieved superior results, particularly in addressing class imbalances and capturing contextual information, leading to a higher F1-score and improved detection accuracy.

[15] A fine-tuned SegFormer model was assessed for corrosion detection, specifically targeting challenges such as class imbalance and limited annotations. A preprocessing step for binary corrosion segmentation was incorporated to enhance dataset quality. The model yielded promising results, with a test loss of 0.2621, a mean

accuracy of 0.8139, and a mean IoU of 0.7116, demonstrating its effectiveness in corrosion detection and its potential for advancing semantic segmentation in critical applications.

[16] Ameli et al. utilized a publicly available dataset comprising 514 corrosion images, employing pixel-wise annotations based on BIRM and AASHTO guidelines to classify corrosion severity into "Fair," "Poor," and "Severe." YOLOv8 and Mask R-CNN were trained on the annotated dataset, achieving mAP50 scores of 0.726 and 0.674, respectively. The study demonstrated the models' effectiveness in corrosion segmentation and structural condition assessment.

[17] Nabizadeh and Parghi proposed an automated corrosion detection system leveraging deep learning and computer vision techniques to address the challenges of inspecting aging civil structures. Their study compared the performance of YOLOv3, YOLOv5s, and YOLOv7 in detecting concrete corrosion from real-world images, employing evaluation metrics such as accuracy, F1-score, recall, and mAP. The results indicated that YOLOv5s achieved the highest mAP@0.5 of 0.88, outperforming other models and underscoring its suitability for corrosion detection in structural health monitoring.

[19] proposed a serial architecture-based corrosion detection method for hydraulic metal structures using an improved YOLOv10 model. Their approach involves two stages: YOLOv10-vit for corrosion localization and YOLOv10-vit-cls for corrosion severity classification. By integrating MobileViTv3 into the YOLOv10n backbone, YOLOv10-vit achieves superior precision, while YOLOv10-vit-cls leverages transfer learning to enhance classification accuracy. This method simplifies annotation by eliminating the need for corrosion severity labeling, improving efficiency.

[33] addressed the specific challenges of high salinity and humidity environments by exploring various improved YOLOv5 models with modified IoU loss functions. Their work on the Zhoushan seawater station dataset demonstrated that YOLOv5-NWD achieved a 7.2% precision improvement, highlighting the importance of specialized loss functions for small-object corrosion detection in marine settings.

[34] introduced MCD-Net, a convolution and sequence encoding combined network that employs a visual Transformer sequence encoder within a convolutional encoder-decoder framework. Their method achieved an F1 score of 84.53% on public corrosion data by enhancing global information processing and establishing long-range feature dependencies through attention-based feature fusion.

While their approachs demonstrates strong performance on general corrosion segmentation, Proposing a new approach is essential to advance the field by addressing existing limitations and exploring innovative solutions because real-time corrosion detection requires both high precision and fast processing to ensure timely and accurate identification. To meet these demanding requirements, continuous efforts toward improvement and innovation are essential, driving the development of more effective and efficient detection approaches.

3. Methodology

This section details the development and evaluation of the proposed corrosion detection framework, whose architecture is presented in Figure 1. The design capitalizes on the proven advantages of real-time object detection paradigms, integrating key concepts inspired by models such as YOLOv11 and YOLOv12 while adapting them to the specific requirements of corrosion detection. The proposed pipeline comprises three main stages: data acquisition and preprocessing, customized model architecture and training, and an extensive performance evaluation.

Figure 1 provides an overview of the system's components and their interconnections, illustrating the end-to-end detection process

3.1. Advances in Real-Time Corrosion Detection

The evolution of real-time object detection has profoundly influenced a wide range of industrial applications, including corrosion monitoring and inspection. In this study, we harness the capabilities of two state-of-the-art object detectors YOLOv11 and YOLOv12 each introducing novel advancements in detection speed, accuracy, and robustness. These models are designed to operate efficiently in real-world environments, making them highly

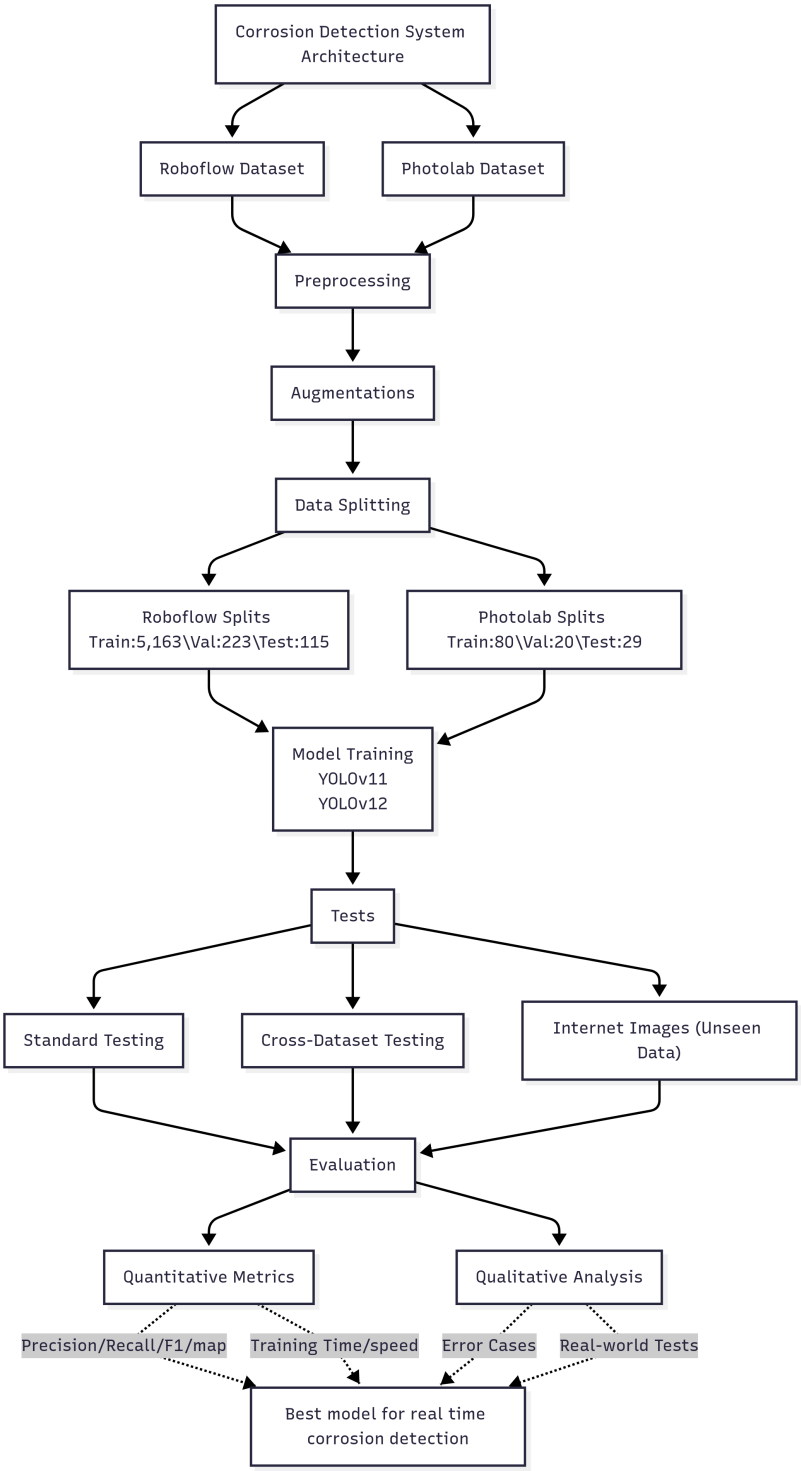


Figure 1. Corrosion Detection System Architecture Overview

suitable for the rapid and precise identification of corrosion across various material surfaces and structural conditions.

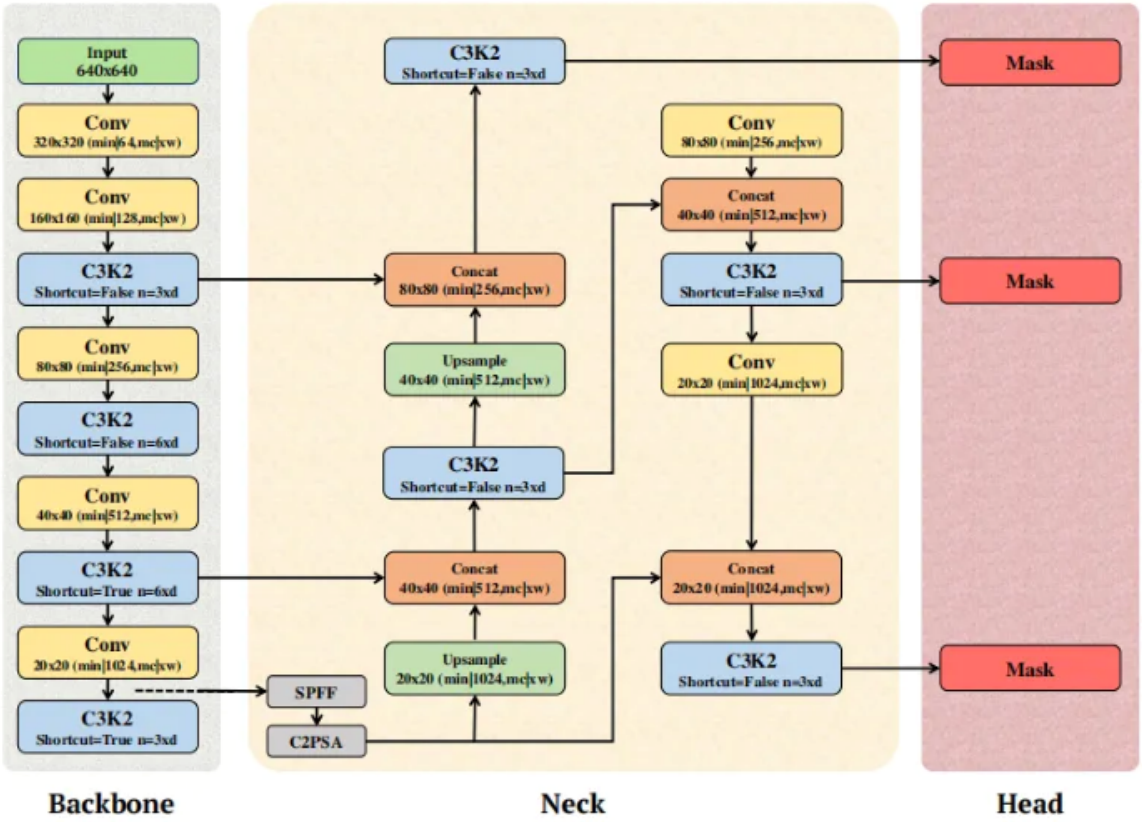


Figure 2. YOLOv11 Architecture Overview [11]]

3.1.1. ***YOLOv11: A New Era in Real-Time Object Detection*** Introduced by Ultralytics in late 2024[11], YOLOv11 marks a major step forward in deep learning-based object detection. Designed to enhance both speed and accuracy, it refines key architectural components while improving generalization across diverse datasets.

Key Innovations in YOLOv11

- Advanced Feature Representation: Efficient backbone and neck structures for better feature extraction.
- Faster and More Efficient Processing: High-speed inference with robust accuracy, ideal for real-time applications [11].
- Improved Accuracy with Fewer Computational Resources: Optimized parameter efficiency.
- Versatile Deployment: Seamless integration across cloud, edge, and GPU platforms.
- Expanded Capabilities: Extended to segmentation, classification, and oriented bounding boxes.

YOLOv11 Architecture Highlights:

- **Backbone:** C3K2 blocks for efficient feature extraction.
- **Neck:** Spatial Pyramid Pooling Fast (SPPF) for multi-scale feature aggregation.
- **Attention Mechanisms:** C2PSA block improves focus on important spatial regions.
- **Detection Head:** Multi-scale predictions for objects of varying sizes.

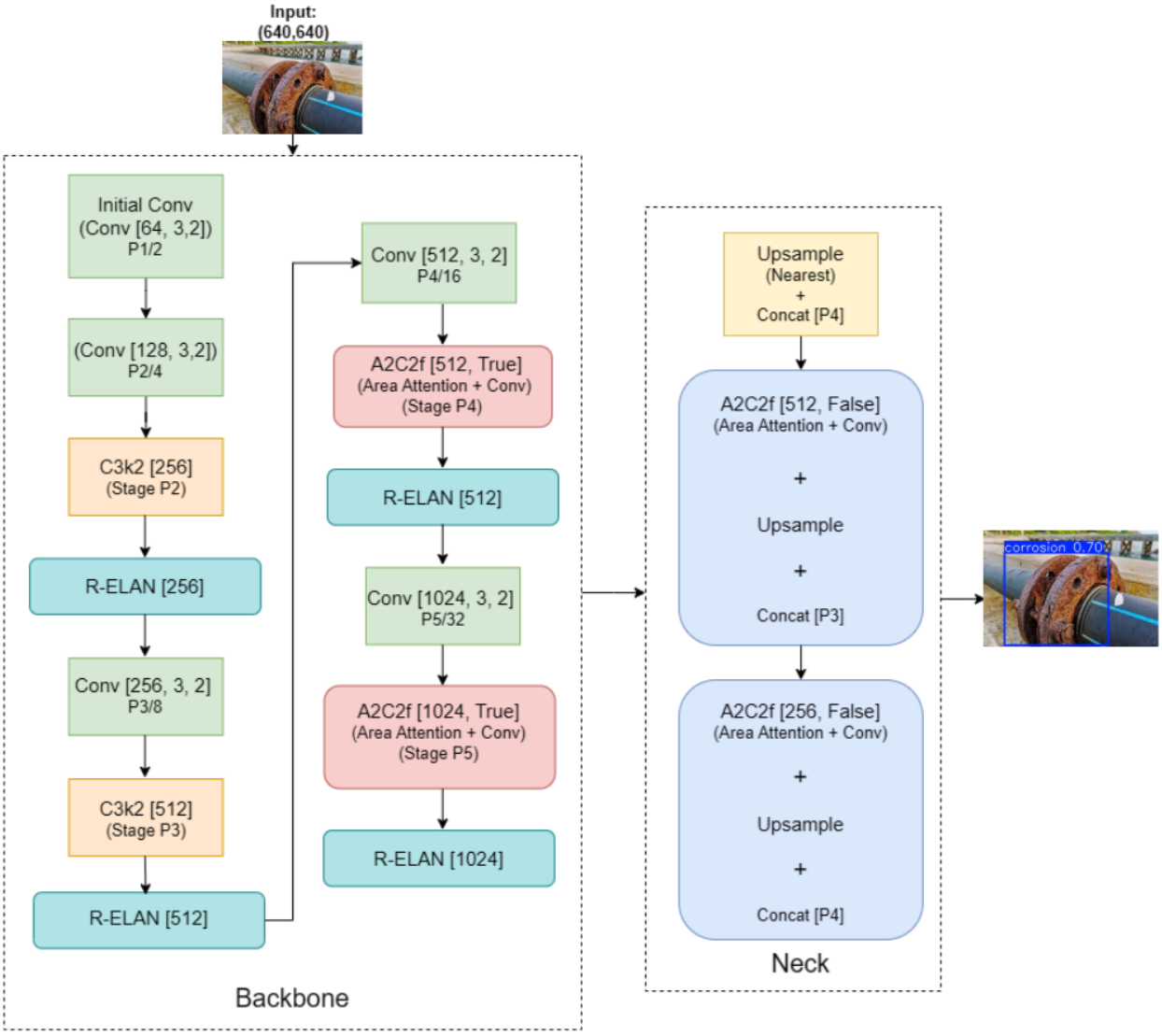


Figure 3. YOLOv12 Architecture Overview. [28]

3.1.2. ***YOLOv12: Pushing the Boundaries of Real-Time Detection*** Building upon the advancements made with YOLOv11, YOLOv12, introduced in early 2025 [28], continues to refine real-time object detection. YOLOv12 offers superior accuracy and faster processing, with several innovations aimed at optimizing performance, especially in complex scenarios like corrosion detection.

Key Innovations in YOLOv12

- Feature Reassembly Module (FRM): This novel module enhances small object detection, especially in crowded or overlapping scenarios. It improves detection precision and robustness by reassembling and refining features from different scales.
- Dynamic Label Assignment: This innovation optimizes the assignment of labels during training, improving detection accuracy, particularly for objects that are close together or occluded, which is often the case in corrosion detection scenarios.

- Area Attention (A^2) Module: The A^2 module introduces local attention within segmented areas, significantly improving computational efficiency and detection accuracy for small, irregularly shaped objects like corrosion patterns.
- Residual Efficient Layer Aggregation Networks (R-ELAN): R-ELAN integrates residual connections at the block level, improving feature aggregation. This innovation enhances gradient flow, training stability, and learning efficiency, especially in larger attention-based models.
- FlashAttention Integration: This integration optimizes memory usage and speeds up attention operations, reducing computation time while maintaining high accuracy.
- Position Perceiver: Rather than relying on traditional positional encoding, YOLOv12 uses separable convolution for capturing positional information, resulting in improved speed and accuracy.

YOLOv12 Strenghs:

- Combines fast inference speeds with enhanced performance through attention mechanisms.
- Achieves state-of-the-art object detection accuracy while maintaining real-time performance.

To analyze the segmentation and detection performance of the proposed models, we evaluated different versions of YOLOv11 and YOLOv12. For YOLOv11, three scaled variants were tested: YOLOv11n, YOLOv11s, and YOLOv11m. Similarly, for YOLOv12, we evaluated YOLOv12n, YOLOv12s, and YOLOv12m. Each variant differs in model size, number of parameters, and computational complexity, influencing both detection accuracy and inference speed. Table 1 summarizes the key characteristics of these YOLOv11 and YOLOv12 models, including the number of parameters (in millions) and floating point operations (FLOPs in gigaflops). Compared to YOLOv11, the YOLOv12 variants introduce approximately a 10% increase in the number of parameters and a slight rise in FLOPs, due to enhancements in the backbone and neck structures. Despite this growth, YOLOv12 models maintain real-time performance, offering improved feature representation and detection accuracy without significant computational overhead.

Table 1. Segmentation and Detection: YOLOv11 and YOLOv12 Versions Evaluated

Model	**Input Size**	**Params (M)**	**FLOPs (G)**
YOLOv11n	640×640	3.9	14.2
YOLOv11s	640×640	13.4	48.3
YOLOv11m	640×640	30.5	124.8
YOLOv12n	640×640	4.5	16.3
YOLOv12s	640×640	14.8	52.5
YOLOv12m	640×640	33.2	135.7

These variations allow us to systematically assess the trade-offs between model complexity, computational efficiency, and detection accuracy in real-time corrosion detection tasks.

To provide a comprehensive comparison of recent object detection architectures for corrosion detection, we analyzed the structural and functional differences among YOLOv11 and YOLOv12. Table 2 summarizes key architectural innovations, including backbone designs, detection head configurations, feature aggregation strategies, and attention mechanisms.

These advancements collectively contribute to faster inference speeds, improved handling of small or partially occluded corrosion areas, and enhanced real-time deployment capabilities, positioning YOLOv11, YOLOv12 as highly promising solutions for industrial corrosion monitoring applications.

Table 2. Comparison of YOLOv11 and YOLOv12 Architectures

Feature	**YOLOv11**	**YOLOv12**
Building Block	C2f+ (Enhanced)	C2f++ with Attention
Detection Head	Anchor-free (Improved)	Unified Detection and Segmentation Head
Kernel Size	Adaptive Kernel	Dynamic Kernel with Spatial Awareness
Backbone Output Channels	128, 256, 512, 768	128, 256, 512, 1024
Neck Architecture	Efficient BiFPN	Multi-Scale BiFPN++
Feature Aggregation	CSP with Transformer Enhancements	Enhanced Transformer-CSP Fusion
Training Efficiency	Optimized for Low Compute	Fast Convergence with Self-Distillation
Small Object Detection	Enhanced	Superior with Contextual Attention
Real-time Performance	Ultra-fast, Low Latency	High-Speed with Lightweight Optimization

3.2. Datasets and Preprocessing

To ensure a robust evaluation of the corrosion detection models, this study utilizes two distinct datasets with different characteristics. A summary of the key properties of both datasets is provided in Table 3 to ensure clarity and reproducibility.

Primary Dataset The primary dataset for this study, titled ”Corrosion Instance Segmentation,” was sourced from Roboflow Universe [13]. This dataset is a compilation of images from various sources, including internet searches and manual captures, providing a diverse representation of corrosion in real-world settings (e.g., bridges, ship hulls, pipelines) and laboratory environments. The images feature corrosion on a variety of surfaces, primarily structural steel and painted metals. The original image resolutions vary widely, from approximately 640x427 pixels to 4032x3024 pixels.

The annotations were provided in an instance segmentation format (Polygon) by the dataset creators [12], with detailed masks precisely outlining the boundaries of corroded regions. We selected the second version of this dataset for training due to its optimal balance between size and augmentation diversity. The dataset consists of 5,501 images, split into 5,163 images for training, 223 for validation, and 115 for testing (approximately a 94%/4%/2% ratio). The applied augmentations, which were pre-applied by the Roboflow platform, include horizontal and vertical flipping, rotations from -15° to +15°, exposure adjustments from -25% to +25%, Gaussian blur up to 2.5 pixels, and salt-and-pepper noise affecting up to 10% of pixels. These augmentations enhance the model’s robustness to real-world variations. Examples of images from this dataset are shown in . 4.

Secondary Dataset The second dataset, titled ”photolab” [18], was sourced from Roboflow Universe and comprises 129 images. This dataset is highly specific, consisting of close-up, high-resolution photographs of bolt groups and connections on a galvanized steel structure. It is annotated with bounding boxes for object detection tasks, focusing on localized corrosion on the bolts and nuts. Given the very small size and highly uniform nature of this dataset, we acknowledge a significant risk of overfitting when used for training deep learning models. Therefore, its primary use in this study is for a preliminary assessment of model generalization to a very specific, targeted scenario. The dataset was split by the source provider into 90 images for training, 19 for validation, and 20 for testing (approximately a 70%/15%/15% ratio). The original images are of a uniform high resolution, typically 4032x3024 pixels. Examples of images from this dataset are shown in 5

Table 3. Detailed summary of the datasets used for model training and evaluation.

Property	**Roboflow-Corrosion (Instance Segmentation)**	**Photolab-Corrosion**
Original Source & Content	Mixed sources (internet, manual capture). Real-world infrastructure & lab settings.	Close-up photos of bolt groups on a steel structure.
Materials Depicted	Structural steel, painted surfaces (bridges, ships).	Galvanized steel bolts and connections.
Total Images	5,501	129
Train/Val/Test Split	5,163 / 223 / 115	90 / 19 / 20
Image Resolution Range	$\sim$640×427 – 4032×3024 pixels	$\sim$4032×3024 pixels (uniform)
Annotation Type	Instance Segmentation (Polygon)	Object Detection (Bounding Box)
Annotation Process	Annotated by the dataset creator [12].	Annotated by the dataset creator[18]].
Primary Use in Study	Main training and quantitative evaluation.	Preliminary generalization assessment to a specific use case.
Key Limitations	Heterogeneous sources; test set is relatively small.	Extremely small size, highly uniform, high risk of overfitting.

Figure 4. Examples of 1st dataset images highlighting corrosion on different metal structures and surfaces [12].

3.3. Training Procedure

3.4. Training Settings

In this research work, there is a rigorous evaluation of YOLOv11 and YOLOv12 with their three versions being compared using the selected datasets. Each variant is fine-tuned with a standard base setup for a level playing field. The input images have been resized to 640×640 pixels for the optimal use of computational resources with the quality of feature extraction. We have trained the models for 300 epochs, allowing them to capture complex corrosion patterns effectively.

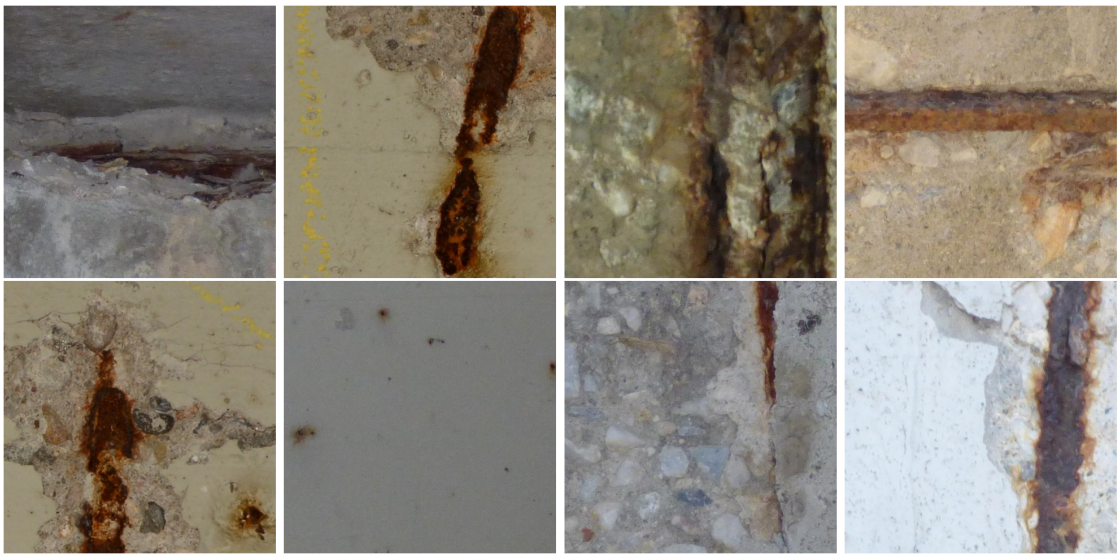


Figure 5. Examples of 2nd dataset images showing corrosion on Close-up photos of bolt groups on a steel structure .

The learning rate of 0.01 with the SGD optimizer and a momentum of 0.937 was initially adopted from the Ultralytics repository, where these values serve as strong baselines. To ensure their suitability for corrosion detection, we conducted a small ablation study by testing alternative learning rates (0.001 and 0.005), a smaller batch size (8), and the AdamW optimizer. Results showed that the SGD configuration (lr=0.01, momentum=0.937) provided the most stable convergence and highest mean average precision (mAP). AdamW produced slower convergence and slightly lower mAP, while smaller batch sizes offered no significant benefit. A batch size of 16 was therefore retained for its balance between stability and efficiency.

We also assessed the effect of data augmentations (blur, noise, and exposure adjustments). Removing these augmentations led to weaker generalization on external validation images (mAP drop of approximately 2%), confirming their contribution to model robustness.

We systematically assess the train and validation performance across the core metrics of assessment in terms of detecting underfitting or overfitting, with a clear perspective of the generalization power of each model. We also assess the train and validation times of the three YOLOv11 and YOLOv12 models with regard to their efficiency as well as their feasibility for use in real-time. To further guarantee the validity as well as the generalizability of our models, a second stage of evaluation is applied. Other online resources, aside from the initial dataset, were used for qualitative assessment. Visualization using graphs reveals each model's ability for adjustment when compared with actual-world data conditions.

By following this methodical evaluation process, we obtain a comprehensive overview of the performance of YOLO models in corrosion detection, with particular emphasis on their resilience and flexibility under varying input conditions. The results highlight each model's ability to maintain high detection accuracy across diverse environmental settings and corrosion severity levels.

4. Results and Discussions

This section presents a comprehensive evaluation of the YOLOv11 and YOLOv12 based systems for real-time corrosion detection. Their performance is assessed in comparison with other YOLO variants, including YOLOv3, YOLOv5, YOLOv7, and YOLOv8, focusing on accuracy, inference speed, and qualitative robustness. The analysis is structured as follows: first, we report the quantitative performance metrics; second, we assess the models' real-time capabilities; third, we present qualitative visual results; and finally, we provide a detailed comparative analysis.

4.1. Evaluation Metrics

To assess the performance of the proposed corrosion detection model, we employ standard object detection and segmentation evaluation metrics. These include **Precision, Recall, F1-score, mean Average Precision (mAP) at different IoU thresholds, Training Time, and Inference Speed**. These metrics provide a comprehensive analysis of the model's accuracy, robustness, and efficiency.

4.1.1. Precision Precision measures the proportion of correctly detected corrosion regions among all detected instances. It is defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

where **TP (True Positives)** are correctly predicted corrosion instances, and **FP (False Positives)** are incorrectly predicted instances. High precision indicates a lower false detection rate.

4.1.2. Recall Recall measures the proportion of correctly detected corrosion regions among all actual corrosion instances. It is given by:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

where **FN (False Negatives)** represents actual corrosion regions that were not detected. High recall indicates that the model is detecting most of the corrosion instances.

4.1.3. F1-Score The **F1-score** is the harmonic mean of **Precision** and **Recall**, balancing both metrics. It is defined as:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

A high F1-score indicates a good balance between precision and recall.

4.1.4. Mean Average Precision (mAP) **Mean Average Precision (mAP)** is a standard metric for evaluating object detection and segmentation models. It measures the area under the precision-recall curve at different **Intersection over Union (IoU)** thresholds.

- **mAP@50 (mAP50):** This evaluates the model's performance at **IoU = 0.50**, meaning a detection is correct if it overlaps at least **50%** with the ground truth.
- **mAP@[50:75] (mAP50-75):** This represents the mean of **mAP** scores calculated at multiple IoU thresholds ranging from **0.50 to 0.75** in increments of 0.05. This metric ensures that the model not only detects corrosion but also accurately localizes it.

4.1.5. Training Time is the total time required to train the model over a defined number of epochs. This metric is crucial for evaluating the computational efficiency of the model. It depends on **dataset size, model architecture, and hardware configuration**.

4.1.6. Inference Speed Inference speed measures how fast the model can process an image during real-time detection. It is typically measured in:

- **Frames Per Second (FPS):** Number of images the model processes per second.
- **Latency (ms per image):** The time taken to process a single image.

A higher **FPS** and lower **latency** indicate better real-time performance, which is essential for industrial corrosion detection applications.

These evaluation metrics provide a **comprehensive** assessment of the corrosion detection model, ensuring that it is both accurate and efficient for real-world applications.

4.2. Quantitative Performance and Comparative Analysis

4.2.1. Primary Dataset

The detection and instance segmentation performance of multiple YOLO model generations, from YOLOv3 to the latest YOLOv12, were rigorously evaluated on the primary dataset, which consists of images representing various corrosion severities and environmental conditions. Table 4 presents a comparative analysis of these models on both training and validation sets, including key metrics such as precision, recall, mAP, and computational efficiency.

Overall Performance Trends and Evolution The results clearly illustrate the performance evolution across YOLO generations. The older architectures, **YOLOv3** and **YOLOv7**, establish a baseline but are generally outperformed by newer models in both accuracy and speed. The **YOLOv5** family shows a significant leap, with its medium variant (YOLOv5m) achieving a respectable validation mAP@50 of 0.568. However, the most recent generations, **YOLOv8**, **YOLOv11**, and **YOLOv12**, demonstrate the pinnacle of performance, consistently pushing the boundaries of precision and mean average precision.

Top-Tier Models in Accuracy Among all models tested, **YOLOv11m** achieved the highest overall accuracy metrics. It recorded the best validation precision (0.690), shared the highest validation mAP@50 (0.620) with YOLOv12s, and achieved the top validation mAP@50:95 (0.420). This makes it the most accurate model for this specific corrosion detection task, albeit with a higher computational footprint during inference (29.5 ms).

The **YOLOv8n** model is noteworthy for achieving the highest training precision (0.703) and the fastest training speed (1.1 ms per iteration), indicating excellent learning efficiency from the dataset. Meanwhile, the **YOLOv12m** model achieved the highest validation recall (0.587), suggesting a superior ability to identify all true corrosion instances, which is critical for inspection tasks where missing a defect is costly.

The YOLOv12 Family: A Focus on Robustness and Modern Efficiency The **YOLOv12** family demonstrates itself as a highly competitive modern architecture. While **YOLOv12m** matches YOLOv11m closely in key metrics like mAP@50 (0.618) and achieves the highest recall (0.587), its standout feature is the efficient inference speed of the 's' and 'n' variants. **YOLOv12s** provides a particularly strong trade-off, matching the best-in-class validation mAP@50 (0.620) while maintaining a rapid inference speed (6.0 ms). **YOLOv12n** also produces solid results for a nano-sized model, outperforming its predecessor YOLOv11n in recall and mAP@50 while being faster. It is important to note that this performance comes at the cost of significantly longer training times compared to YOLOv11 and earlier versions, indicating a more complex architecture or training regimen.

Balance of Speed and Accuracy For practical deployment where a balance of speed and accuracy is essential, several models stand out. **YOLOv11s** delivers outstanding performance, with validation metrics nearly matching the top-performing YOLOv11m (mAP@50 of 0.617 vs. 0.620) while boasting the fastest inference speed (4.3 ms) among all medium and small-sized models. This makes it an exceptionally strong candidate for real-time applications.

The **YOLOv8** family also offers compelling speed-accuracy trade-offs, with YOLOv8n having the second-fastest inference time (9.0 ms) and YOLOv8s providing robust all-around performance.

Conclusion on Model Selection

In conclusion, the choice of the optimal model depends heavily on the deployment priorities:

For **maximum accuracy** without strict speed constraints, **YOLOv11m** is the definitive choice.

For a **strong trade-off between modern architecture performance and speed**, the **YOLOv12s** model is an excellent option, offering top-tier mAP with fast inference.

For **the fastest high-accuracy deployment**, **YOLOv11s** provides an unbeaten combination of speed and accuracy.

For scenarios demanding **extremely low latency** with good enough accuracy, **YOLOv8n** and **YOLOv12n** are excellent lightweight options.

These findings demonstrate that both YOLOv11 and YOLOv12 families offer viable and leading-edge solutions for corrosion detection tasks. YOLOv11m excels in peak accuracy, while YOLOv12s provides a very strong

balance of modern performance and speed, making the two families complementary depending on the specific application requirements.

Table 4. Comparative Analysis of YOLOv3, YOLOv5, YOLOv7, YOLOv8, YOLOv11, and YOLOv12 Models on the primary dataset

Model	Precision		Recall		F1-score		mAP@50		mAP@50:95		Speed (ms)		Training Time
	Train	Val	Train	Val	Train	Val	Train	Val	Train	Val	Train	Inference	(hours)
YOLOv3	0.600	0.580	0.580	0.560	0.590	0.570	0.580	0.560	0.350	0.320	12.0	50.0	12.0
YOLOv5n	0.643	0.661	0.549	0.474	0.592	0.552	0.553	0.461	0.301	0.234	5.8	50.9	0.327
YOLOv5s	0.659	0.618	0.545	0.593	0.597	0.605	0.568	0.552	0.334	0.305	7.9	24.7	0.357
YOLOv5m	0.627	0.636	0.593	0.579	0.592	0.593	0.606	0.568	0.531	0.343	9.8	28.9	0.452
YOLOv7	0.640	0.615	0.600	0.576	0.620	0.595	0.610	0.561	0.390	0.337	8.0	20.0	4.0
YOLOv8n	**0.703**	0.644	0.531	0.564	0.605	0.601	0.577	0.587	0.345	0.315	**1.1**	**9.0**	0.336
YOLOv8s	0.632	0.624	0.574	0.598	0.602	0.611	**0.589**	0.578	**0.357**	0.319	2.0	7.6	0.361
YOLOv8m	0.630	0.621	**0.588**	**0.604**	**0.608**	**0.613**	0.590	0.543	0.370	0.299	2.4	9.7	0.477
YOLOv11n	0.643	0.647	0.534	0.536	0.574	0.575	0.588	0.588	0.389	0.389	4.3	6.8	**0.349**
YOLOv11s	0.662	0.663	0.574	0.573	0.617	0.616	**0.619**	0.617	0.417	0.417	7.4	**4.3**	0.473
YOLOv11m	**0.687**	**0.690**	0.570	0.570	**0.619**	**0.620**	**0.619**	**0.620**	**0.420**	**0.420**	14.8	29.5	0.517
YOLOv12n	0.645	0.646	0.553	0.555	0.594	0.594	0.592	0.594	0.386	0.387	1.2	5.5	1.481
YOLOv12s	0.655	0.659	0.584	0.585	0.615	0.621	0.619	0.620	0.419	0.419	2.7	6.0	1.924
YOLOv12m	0.646	0.647	**0.587**	**0.587**	0.615	0.618	0.618	0.618	0.418	0.418	17.2	4.8	9.239

4.2.2. Secondary Dataset

To assess model generalizability across different data environments, performance was evaluated on a second, more focused dataset. This dataset differs from the primary one in three key aspects: context, diversity, and scale.

Context Diversity: The primary dataset is a large, mixed-source collection from real-world infrastructure and lab settings, presenting a wide variety of corrosion types, backgrounds, and scales. In contrast, the secondary dataset consists of close-up photos of a single, specific structural element: bolt groups on a steel structure. This results in a much more homogeneous set of images with less variation in scene composition.

Scale: The primary dataset is significantly larger, providing more data for models to learn generalizable features. The smaller size of the secondary dataset presents a different challenge, testing a model's ability to learn effectively from limited, specific examples.

These fundamental differences directly explain the shifts in model performance observed in Table 5.

Performance Shift in a Homogeneous, Limited-Data Context

The performance hierarchy changes notably on the secondary dataset. YOLOv5s emerges as a top performer, achieving the highest validation mAP@50 (0.893). This suggests that its architecture is particularly efficient at learning and recognizing patterns from a smaller, more focused dataset without overfitting.

Conversely, YOLOv8s demonstrates its robustness by delivering the most well-rounded performance. It achieves the highest validation recall (0.852) and, critically, the highest validation mAP@50:95 (0.481). Its ability to maintain high precision and the best localization accuracy on this dataset underscores its adaptability to both diverse and specific visual tasks.

Analysis of Modern Architectures on a Narrower Task

The YOLOv11 family, particularly the YOLOv11m variant, maintained its signature strength of high precision (0.903), proving highly reliable for minimizing false positives even in this specific context. However, the more complex YOLOv12 models did not show a clear advantage here. This indicates that their advanced architectures, which may excel on large and varied datasets, do not necessarily provide a performance boost on smaller, more homogeneous datasets where the learning problem is less complex. The larger models might be prone to overfitting or simply not have enough data to leverage their full capacity.

Computational Implications

The smaller dataset size also contributes to the dramatically shorter training times across all models compared to the primary dataset. In this context, YOLOv8s again presents a compelling package by combining top-tier accuracy with the fastest inference speed (11.6 ms), making it ideal for a potential dedicated inspection system for this specific component.

The comparative analysis across the two datasets leads to a critical conclusion: the optimal model is not universal but is heavily influenced by the nature of the training data.

For large, diverse datasets mimicking real-world variability (like the primary dataset), the latest models like YOLOv11m and YOLOv12s excel by leveraging their complex learning capabilities.

For smaller, focused datasets of a specific component (like the secondary dataset), older, well-established architectures like YOLOv5s can achieve peak mAP@50, while robust modern models like YOLOv8s provide the best all-around accuracy and speed.

This underscores the importance of matching the model architecture to the data environment. A larger, more complex model is not inherently better; for targeted applications, a simpler model trained on precise, context-specific data can be the most effective and efficient solution.

4.3. Qualitative Results

To further assess the performance of the YOLOv11 and YOLOv12 variants, we present qualitative results comparing their detection and segmentation capabilities on various corrosion images. Tables 6 and 7 illustrates the results obtained alongside the original images. Each row in the table corresponds to a different test image, showcasing the effectiveness of each model in detecting and segmenting corrosion regions.

The qualitative results in Table 3 highlight detection and segmentation differences among the YOLOv11 variants. In Row 1, all models accurately detect widespread surface rust, but YOLOv11m achieves slightly sharper mask boundaries. Row 2 (bridge girder) shows YOLOv11m producing tighter segmentation around the corrosion patch, whereas YOLOv11s and YOLOv11n slightly extend into unaffected areas. In Row 3 (fence corrosion), all variants detect small rust spots, though YOLOv11s and YOLOv11n exhibit minor over-segmentation of the surrounding metal. Row 4 (bolts) demonstrates that YOLOv11m maintains more consistent mask alignment with the true corroded region. These visual trends align with the quantitative evaluation in Section 4.2, where YOLOv11m achieves the highest mAP and IoU scores, confirming its superior balance between detection precision and segmentation quality for real-time SHM corrosion monitoring.

The qualitative comparison in Table 4 illustrates the detection behavior of YOLOv12 variants across various corrosion conditions. In Row 1, all models successfully detect multiple corrosion spots on the beam, though YOLOv12m and YOLOv12n show slightly cleaner bounding box placement than YOLOv12s. Row 2 (rusted bolt) reveals that all variants capture the corroded region accurately, with YOLOv12m offering the tightest fit around the circular rust area. In Row 3 (heavily corroded nuts and bolts), YOLOv12s tends to produce more bounding boxes, occasionally overlapping, whereas YOLOv12n and YOLOv12m maintain more concise detections. Row 4 (metal sample) shows minimal difference between the three, with all variants producing consistent localization.

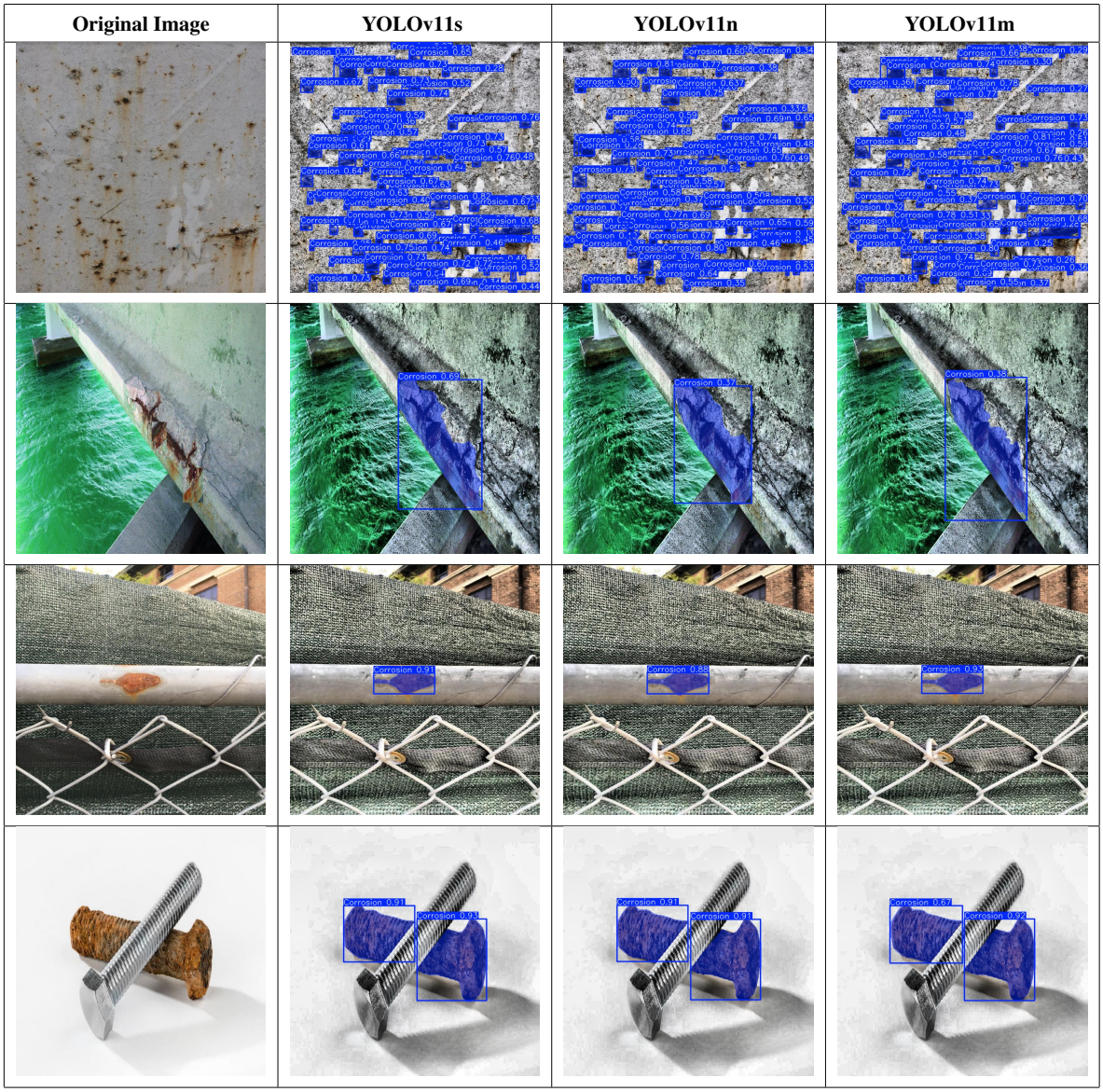
Table 5. Comparative Analysis of YOLOv3, YOLOv5, YOLOv7, YOLOv8, YOLOv11, and YOLOv12 Models on the secondary dataset

Model	Precision		Recall		F1-score		mAP@50		mAP@50:95		Speed (ms)		Training Time
	Train	Val	Train	Val	Train	Val	Train	Val	Train	Val	Train	Inference	(hours)
YOLOv3	0.780	0.720	0.700	0.680	0.738	0.699	0.750	0.720	0.400	0.380	1.0	25.0	0.250
YOLOv5n	0.836	0.771	0.722	0.750	0.775	0.760	0.794	0.777	0.451	0.434	0.2	18.1	0.194
YOLOv5s	0.844	0.849	**0.902**	0.806	0.871	0.827	**0.925**	**0.893**	0.416	0.419	0.2	20.8	0.167
YOLOv5m	0.810	0.780	0.820	0.770	0.815	0.775	0.835	0.780	0.440	0.420	0.2	22.0	0.353
YOLOv7	0.820	0.790	0.800	0.770	0.810	0.780	0.830	0.780	0.450	0.420	0.2	19.0	0.350
YOLOv8n	0.843	0.770	0.778	0.806	0.810	0.788	0.825	0.784	0.444	0.411	0.2	16.8	0.145
YOLOv8s	0.891	0.814	0.861	**0.852**	0.876	0.833	0.889	0.884	**0.561**	**0.481**	0.2	**11.6**	0.598
YOLOv8m	0.858	**0.904**	0.861	0.786	0.860	**0.841**	0.905	0.881	0.470	0.460	0.2	25.0	0.409
YOLOv11n	0.882	0.822	0.722	0.722	0.794	0.769	0.822	0.797	0.446	0.441	3.2	26.5	**0.130**
YOLOv11s	0.885	0.894	0.858	0.703	0.871	0.787	0.886	0.881	0.470	0.459	6.3	13.7	0.256
YOLOv11m	**0.912**	0.903	0.778	0.777	**0.839**	0.835	0.879	0.857	0.472	0.455	10.5	30.9	0.424
YOLOv12n	0.868	0.839	0.833	0.806	0.850	0.822	0.861	0.844	0.464	0.450	6.0	13.6	0.219
YOLOv12s	0.875	0.833	0.777	0.832	0.823	0.832	0.870	0.784	0.452	0.465	8.6	23.8	0.379
YOLOv12m	0.798	0.867	0.878	0.724	0.835	0.790	0.871	0.834	0.447	0.441	18.0	60.7	0.598

4.4. Limitations and Trade-offs

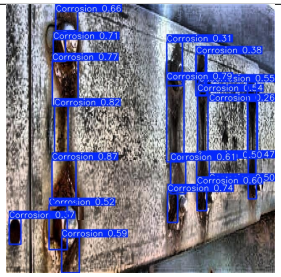
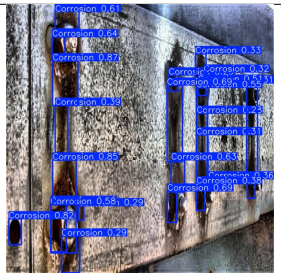
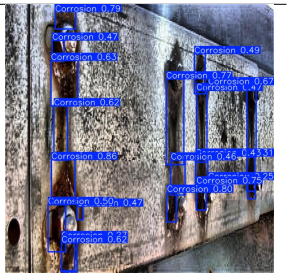
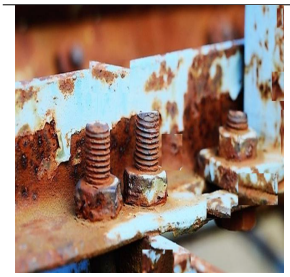
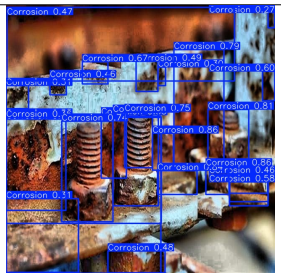
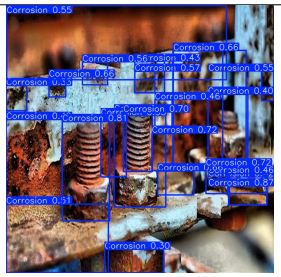
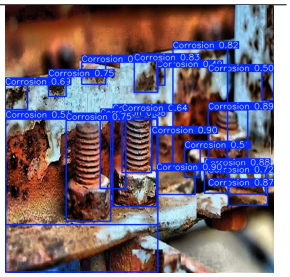
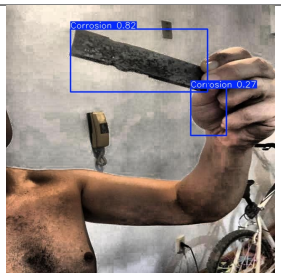
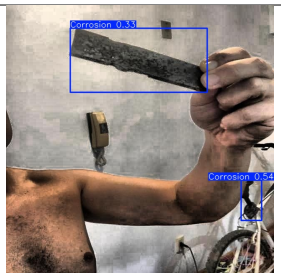
This study has some limitations. The primary dataset is sufficiently large and diverse, but the secondary dataset is relatively small (129 images) and highly domain-specific. This may reduce the generalizability of the findings to other specialized contexts. In terms of training settings, while we conducted a focused ablation study (testing alternative learning rates, batch sizes, optimizers, and augmentations), we did not perform an exhaustive hyperparameter search. More advanced schedules, augmentation policies, or regularization strategies may yield additional improvements and remain an avenue for future work. Finally, computational resources posed practical constraints: larger models required substantially longer training times and greater hardware capacity, which may limit their accessibility for some practitioners. Another limitation is that the datasets were obtained from Roboflow with pre-existing annotations. While this facilitated rapid experimentation, we did not independently validate annotation quality or compute inter-annotator agreement, which may affect label reliability in cases of ambiguous corrosion boundaries. Future work should incorporate independent annotation checks and formal agreement metrics to strengthen dataset robustness. An important practical consideration is the trade-off between accuracy, speed, and model size. As shown in Table 4, smaller models such as YOLOv11n (3.9M params, 14.2 GFLOPs) and YOLOv12n (4.5M params, 16.3 GFLOPs) achieve competitive validation mAP (0.588 and 0.594, respectively) while maintaining very fast inference times (6.8 ms and 5.5 ms). These lightweight models are attractive for deployment on edge devices with limited GPU capacity. In contrast, medium-sized models such as YOLOv11m (30.5M params, 124.8 GFLOPs) and YOLOv12m (33.2M params, 135.7 GFLOPs) provide the highest accuracy

Table 6. Qualitative Comparison of YOLOv11 Variants for Detection and Segmentation : a sample images from the first dataset



(mAP@50 up to 0.620) and F1-scores (0.620) but at the cost of slower inference (up to 29.5 ms) and higher training times. This illustrates the classical trade-off: accuracy improves with model size, but speed and hardware requirements worsen. Therefore, model selection should be guided by application constraints: edge deployment favors compact “n/s” versions, while server-side or offline batch processing can exploit the more accurate “m” versions.

Table 7. Qualitative Comparison of YOLOv12 Variants for Detection : a sample images from the first dataset

Original Image	YOLOv12s	YOLOv12n	YOLOv12m
			
			
			
			

5. Inference Test and Comparative Analysis of YOLOv11 and YOLOV12 Variants

To evaluate the performance of different YOLOv11 and YOLOv12 variants for corrosion detection, we conducted a series of inference tests using multiple model architectures: YOLOv11s, YOLOv11n, YOLOv11m (trained on two distinct datasets), as well as YOLOv12s (trained on two distinct datasets), YOLOv12n, and YOLOv12m. The qualitative results, presented in Figures 6, 7, 8, and 9, demonstrate each model's detection accuracy and confidence levels. These comparisons highlight the strengths, limitations, and generalization capabilities of the YOLOv11 and YOLOv12 variants when applied to real-world corroded metal surfaces.

5.1. YOLOv11s vs. YOLOv11n: Model Variants Comparison

The results on Figures 6 and 7 indicate that YOLOv11n achieves a higher confidence score compared to YOLOv11s, demonstrating its improved detection capability. Specifically:

- YOLOv11s provides a detection confidence of 0.72 and 0.76 across different samples. While it correctly identifies corrosion regions, the segmentation mask suggests that it might miss finer details, leading to partial coverage.
- YOLOv11n, in contrast, achieves a confidence score of 0.87 and 0.79, indicating more reliable detection with better coverage of corrosion regions. The bounding boxes are more consistent, and the segmentation areas closely align with the corroded regions, suggesting improved generalization.

These results suggest that YOLOv11n provides a better trade-off between accuracy and efficiency, making it a preferred option for lightweight deployment scenarios.

5.2. YOLOv11m: Impact of Training Dataset on Detection Performance

To analyze the effect of dataset variation, we trained YOLOv11m on two different datasets and evaluated their performance on corrosion detection. refer to Figures 6 and 7The results show:

- YOLOv11m trained on the first dataset achieves a confidence score of 0.83, indicating a strong ability to detect corrosion regions with high precision. The bounding box is well-aligned, and the segmentation mask effectively captures the rusted area.
- YOLOv11m trained on the second dataset, however, shows a lower confidence score of 0.70, suggesting a decline in detection reliability. Similarly, in other test cases, the confidence score drops to 0.49, reinforcing the hypothesis that the dataset may influence generalization capability.

The performance variation observed in models trained on the second dataset can be attributed to several key factors:

- **Dataset Specificity vs. Generalizability**: The second dataset's focused nature on close-up bolt group images creates a highly specialized domain. Models trained on this data may struggle with the diverse corrosion types and environmental variations present in the more generalized first dataset, leading to reduced cross-dataset performance.
- **Scale and Diversity Limitations**: compared to the larger primary dataset, the second dataset provides limited variation in corrosion patterns, viewing angles, and environmental conditions. This restricted diversity can limit the model's ability to learn robust, generalizable features.
- **Annotation Consistency Challenges**: The homogeneous nature of bolt group images, often featuring repetitive structural elements and similar corrosion patterns, may introduce annotation ambiguities that affect the model's learning of distinct corrosion boundaries and features.

5.3. Key Observations and Implications

- **Model architecture influences detection reliability**: YOLOv11n outperforms YOLOv11s, showing better localization and segmentation accuracy.
- **Training dataset quality plays a crucial role**: YOLOv11m trained on the first dataset generalizes better than the second dataset-trained model.
- **Confidence score variations indicate dataset challenges**: The second dataset-trained model demonstrates lower confidence, suggesting the need for enhanced preprocessing or data augmentation techniques.

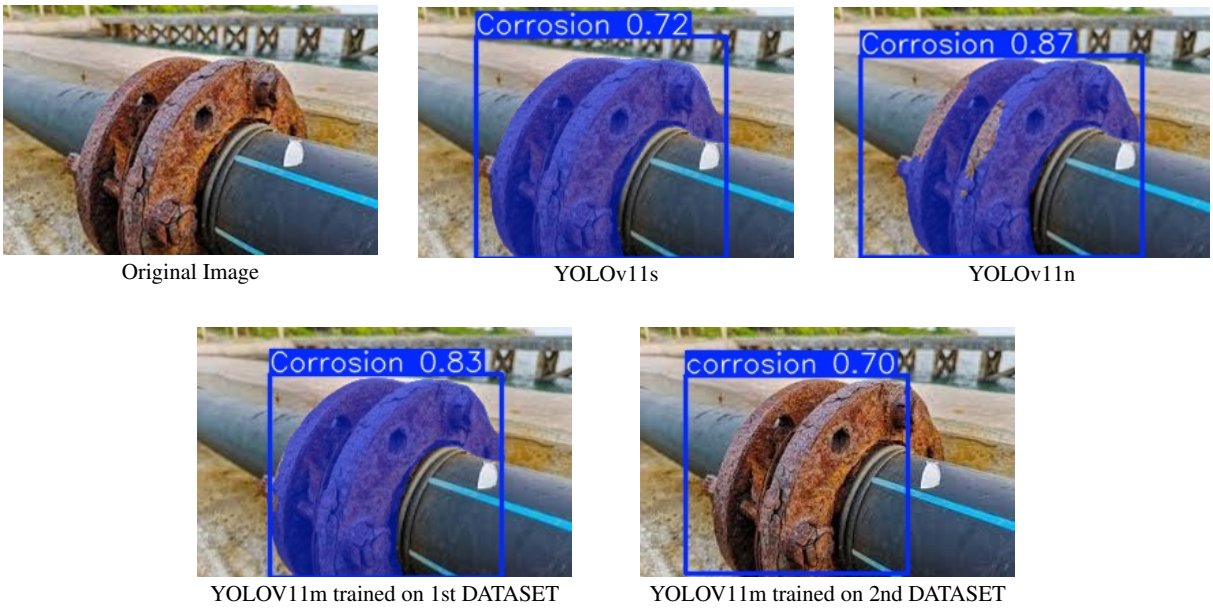
5.4. YOLOv12 Variants: Visual Inference and Detection Insights

Figures 8 and 9 provide a qualitative evaluation of YOLOv12s, YOLOv12n, and YOLOv12m on two corroded surfaces. The results show:

- **YOLOv12s** consistently produced accurate detections with high confidence scores (up to 0.83), demonstrating precise localization of corroded areas. Its performance remained stable across both simple and complex surfaces, making it well-suited for general-purpose corrosion inspection tasks.
- **YOLOv12n**, despite its lightweight architecture, underperformed compared to the other variants. It failed to detect corrosion in one instance and showed reduced confidence scores (e.g., 0.70) in the presence of noisy textures, suggesting limitations in capturing subtle corrosion features.
- **YOLOv12m** identified multiple corrosion regions, even in highly degraded surfaces. However, the model exhibited a wide range of confidence scores (0.80 to as low as 0.33), and some bounding boxes extended into non-corroded regions. This suggests a trade-off between sensitivity and precision, with potential over-segmentation in some cases.

These findings emphasize that while model selection is important, dataset quality and diversity are equally crucial in ensuring robust and accurate corrosion detection. Future work will focus on enhancing dataset annotations, applying advanced augmentation techniques, and testing on real-world corrosion scenarios to validate these models further.

Figure 6. Inference test results of YOLOv11 models on a corroded metal surface, showing variations in detection confidence. Higher confidence scores (above 0.7) generally indicate strong and reliable detections in object detection tasks.



Original Image | YOLOv11s | YOLOv11n

YOLOV11m trained on 1st DATASET | YOLOV11m trained on 2nd DATASET

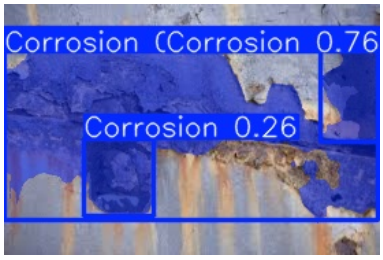
5.5. Deployment Scenarios for Structural Health Monitoring

The proposed corrosion detection framework demonstrates strong applicability for deployment in real-world Structural Health Monitoring (SHM) environments, where continuous and accurate assessment of structural integrity is essential . In operational contexts, the system can be integrated into various sensing and inspection platforms to enable automated, scalable, and cost-effective monitoring.

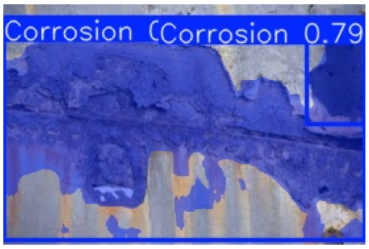
Aerial Inspections via UAVs : Unmanned Aerial Vehicles (UAVs) equipped with high-resolution optical or multispectral imaging systems can be employed to survey large-scale structures such as bridges, transmission towers, and offshore platforms [29]. By embedding the corrosion detection model into onboard computing modules, preliminary defect assessments can be generated in near real time, reducing the need for manual post-processing and accelerating maintenance decision-making [32].

Figure 7. Second inference test results of YOLOv11 models on a different corroded surface, highlighting model generalization. Higher confidence scores (above 0.7) generally indicate strong and reliable detections in object detection tasks.

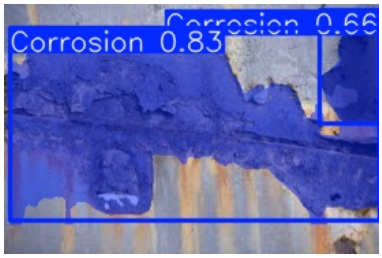

Original Image



YOLOv11s



YOLOv11n



YOLOV11m trained on 1st DATASET

YOLOV11m trained on 2nd DATASET

Figure 8. First inference test results of YOLOv12 models on a corroded metal surface, showing variations in detection confidence. Confidence scores closer to 1.0 represent stronger detections, while lower scores (e.g., below 0.5) suggest weaker or uncertain detections. In this case, the YOLOv12 models produced lower confidence values compared to YOLOv11, indicating less reliable corrosion detection on this surface.


Original Image

YOLOv12s

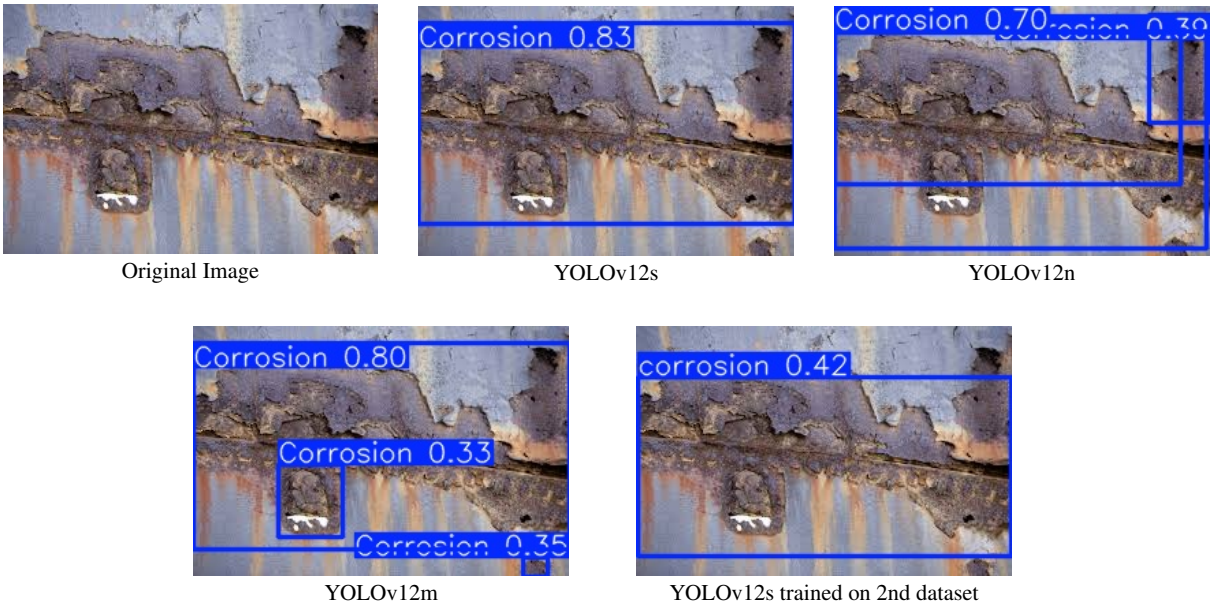

YOLOv12n

YOLOv12m

YOLOv12s trained on 2nd dataset

IoT-Enabled Fixed Monitoring Systems: The model can be deployed on low-power edge devices integrated into stationary IoT sensor networks, enabling continuous monitoring of critical infrastructure components such as pipelines, storage tanks, and wind turbine towers [30].

Figure 9. Second inference test results of YOLOv12 models on a different corroded surface, showing model consistency and generalization. Confidence scores above 0.7 (e.g., 0.83 and 0.80) indicate strong and reliable detections, while lower values (e.g., 0.42, 0.33, and 0.15) reflect weaker confidence, suggesting model uncertainty in identifying corrosion across all regions of the surface.



Original Image | YOLOv12s | YOLOv12n

YOLOv12m | YOLOv12s trained on 2nd dataset

Robotic and Autonomous Inspection Systems: In industrial facilities or confined environments, the detection system can be embedded in robotic crawlers, smart pigging devices, or remotely operated vehicles[31,32]. Such platforms allow internal and external surface inspection without human presence, mitigating safety hazards and ensuring consistent inspection quality.

6. Conclusion

This study emphasizes the advancements introduced by YOLOv11 and YOLOv12 in the domain of real-time corrosion detection, demonstrating notable improvements over earlier YOLO versions. The results confirm that both models offer enhanced accuracy and segmentation quality, positioning them as strong candidates for structural health monitoring applications. The comparative analysis with YOLOv5 and YOLOv8 reinforces the effectiveness of YOLOv11 and YOLOv12, particularly in detailed corrosion assessment scenarios where precise localization is essential. While trade-offs remain between accuracy and inference speed, the flexibility of the models variants allows for adaptation to a variety of real-world industrial settings. However, the evaluation of YOLOv11m and YOLOv12s trained on two different datasets revealed a notable drop in confidence scores when applied to the second dataset, indicating potential dataset limitations. Factors such as class imbalance, inconsistent annotation quality, and low variability in corrosion patterns likely contributed to the reduced detection reliability. Although YOLOv12 was not evaluated on the second dataset, its strong performance on the primary dataset suggests similar potential when applied under optimized conditions. Addressing the aforementioned limitations through improved dataset curation and advanced augmentation techniques will be critical for enhancing model generalization. These findings contribute to the progress of deep learning in automated corrosion detection, offering a more reliable and efficient approach for industrial applications. Future research will explore further optimizations and enhancements to improve performance across diverse environmental conditions.

Generalization Across Materials: Expanding datasets to include more diverse materials and environmental conditions for better generalization.

Integration with Edge Devices: Deploying our models on embedded systems and edge devices for real-time, low-power corrosion monitoring.

Multi-Modal Analysis: Combining image data with sensor-based corrosion detection to enhance reliability.

Declaration of interests

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.

Data availability Statement

The data supporting the findings of this study are publicly available on the Roboflow Universe platform (https://universe.roboflow.com). Specifically, the datasets were obtained from the following open-access resources:

- Corrosion Instance Segmentation (https://universe.roboflow.com/cawilai-interns-july-2023/corrosion-instance-segmentation-sfcpc)
- Photolab (https://universe.roboflow.com/photolab/bgc-1).

REFERENCES

1. R. Varghese, S. M., YOLOv8: A novel object detection algorithm with enhanced performance and robustness, in: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), IEEE, 2024, pp. 1–6. https://doi.org/10.1109/ADICS58448.2024.10533619.
2. G. Jocher, YOLOv5, Ultralytics (2020). Available at: https://github.com/ultralytics/yolov5.
3. P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Kauai, HI, USA, 2001, pp. I–I.
4. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Vol. 1, San Diego, CA, USA, 2005, pp. 886–893.
5. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Columbus, OH, USA, 2014, pp. 580–587.
6. R. Girshick, Fast R-CNN, in: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, 2015, pp. 1440–1448.
7. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 28 (2015).
8. W. Liu, et al., SSD: Single shot multibox detector, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer, Cham, 2016, pp. 21–37.
9. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 779–788.
10. J. Redmon, A. Farhadi, YOLOv3: An incremental improvement, arXiv preprint arXiv:1804.02767 (2018).
11. The architecture of YOLOv11: A leap in object detection, Medium (2025). Available at: https://medium.com/thedeephub/the-architecture-of-yolov11-a-leap-in-object-detection-2ce1e946c74a.
12. E. Casas, L. Ramos, C. Romero, F. Rivas-Echeverría, A comparative study of YOLOv5 and YOLOv8 for corrosion segmentation tasks in metal surfaces, Array 22 (2024) 100351. https://doi.org/10.1016/j.array.2024.100351.
13. CawilAI Interns July 2023, Corrosion Instance Segmentation Dataset, Roboflow Universe (2023). Available at: https://universe.roboflow.com/cawilai-interns-july-2023/corrosion-instance-segmentation-sfcpc.
14. S. Abid, M. Amroune, B. Issam, M.Y. Haouam, Analyzing the performance of semantic segmentation techniques in automatic corrosion detection, in: ECTE-Tech Conf., IEEE, 2024. https://doi.org/10.1109/ECTE-Tech62477.2024.10851097.
15. A. Safa, A. Mohamed, B. Issam, M.Y. Haouam, SegFormer: Semantic segmentation based transformers for corrosion detection, in: 2023 Int. Conf. Networking and Advanced Systems (ICNAS), Algiers, Algeria, 2023, pp. 1–6. https://doi.org/10.1109/ICNAS59892.2023.10330461.
16. Z. Ameli, S.J. Nesheli, E.N. Landis, Deep learning-based steel bridge corrosion segmentation and condition rating using Mask RCNN and YOLOv8, Infrastructures 9 (1) (2024) 3. https://doi.org/10.3390/infrastructures9010003.
17. E. Nabizadeh, A. Parghi, Automated corrosion detection using deep learning and computer vision, Asian J. Civ. Eng. 24 (2023) 2911–2923. https://doi.org/10.1007/s42107-023-00684-4.

18. Photolab, BGC-1 dataset, Roboflow Universe (2022). Available at: https://universe.roboflow.com/photolab/bgc-1.
19. H. Cheng, F. Kang, Corrosion detection and grading method for hydraulic metal structures based on an improved YOLOv10 sequential architecture, Appl. Sci. 14 (24) (2024) 12009. https://doi.org/10.3390/app142412009.
20. H. Xu, X. Su, Y. Wang, H. Cai, K. Cui, X. Chen, Automatic bridge crack detection using a convolutional neural network, Appl. Sci. 9 (2019) 2867. https://doi.org/10.3390/app9142867.
21. E. Protopapadakis, I. Katsamenis, A. Doulamis, Multi-label deep learning models for continuous monitoring of road infrastructures, in: Proc. 13th ACM Int. Conf. Pervasive Technologies Related to Assistive Environments, Corfu, Greece, 2020, pp. 1–7. https://doi.org/10.1145/3389189.3393727.
22. Y. Ren, J. Huang, Z. Hong, W. Lu, J. Yin, L. Zou, X. Shen, Image-based concrete crack detection in tunnels using deep fully convolutional networks, Constr. Build. Mater. 234 (2020) 117367. https://doi.org/10.1016/j.conbuildmat.2019.117367.
23. S. Iyer, T. Velmurugan, A.H. Gandomi, V. Noor Mohammed, K. Saravanan, S. Nandakumar, Structural health monitoring of railway tracks using IoT-based multi-robot system, Neural Comput. Appl. 33 (2021) 5897–5915. https://doi.org/10.1007/s00521-020-05444-5.
24. D. Chen, B. Huang, F. Kang, A review of detection technologies for underwater cracks on concrete dam surfaces, Appl. Sci. 13 (2023) 3564. https://doi.org/10.3390/app13063564.
25. X. Huang, Z. Duan, S. Hao, J. Hou, W. Chen, L. Cai, A deep learning framework for corrosion assessment of steel structures using Inception v3 model, Buildings 15 (4) (2025) 512. https://doi.org/10.3390/buildings15040512.
26. D. Chen, F. Kang, J. Li, S. Zhu, Y. Sun, Unsupervised dam crack image segmentation algorithm based on adversarial learning and image fusion, Autom. Constr. 178 (2025) 106423. https://doi.org/10.1016/j.autcon.2025.106423.
27. X. Huang, C. Liang, X. Li, F. Kang, An underwater crack detection system combining new underwater image-processing technology and an improved YOLOv9 network, Sensors 24 (18) (2024) 5981. https://doi.org/10.3390/s24185981.
28. Ultralytics, YOLOv12 Models Documentation, Ultralytics Docs (2025). Available at: https://docs.ultralytics.com/models/yolo12/.
29. T. Omar, M.L. Nehdi, Remote sensing of concrete bridge decks using unmanned aerial vehicle infrared thermography, Autom. Constr. 83 (2017) 360–371. https://doi.org/10.1016/j.autcon.2017.06.024.
30. S. Iyer, T. Velmurugan, A.H. Gandomi, V. Noor Mohammed, K. Saravanan, S. Nandakumar, Structural health monitoring of railway tracks using IoT-based multi-robot system, Neural Comput. Appl. 33 (11) (2021) 5897–5915. https://doi.org/10.1007/s00521-020-05366-9.
31. A. Das, S. Dorafshan, N. Kaabouch, Autonomous image-based corrosion detection in steel structures using deep learning, Sensors 24 (11) (2024) 3630. https://doi.org/10.3390/s24113630.
32. I. Godwin, D.O. Ene, I. Udo, P.E. Awe, E. Barthelomew, D. Oghenefegor, G. Etebenumeh, B. Oghenenyerovwo, G. Ofualagba, O.T.A. Ejofodomi, Pipeline inspection for corrosion using a mobile robotic system, Int. J. Robot. Eng. 1 (1) (2015) 001. https://doi.org/10.35840/2631-5106/4101.
33. Q. Yu, Y. Han, Y. Han, X. Gao, L. Zheng, Enhancing YOLOv5 performance for small-scale corrosion detection in coastal environments using IoU-based loss functions, J. Mar. Sci. Eng. 12 (12) (2024) 2295. https://doi.org/10.3390/jmse12122295.
34. J. Guo, L. Wang, L. Hua, Efficient metal corrosion area detection model combining convolution and transformer, Appl. Sci. 14 (21) (2024) 9900. https://doi.org/10.3390/app14219900.