

*Research Article***YOLOv10 and RT-DETR for Weld Defect Detection: Experimental Evaluation of Dataset Splits and Explainability****Adem Dilbaz^{a,*}** ^a Sahin Tanker Limited Company, Asagipinarbasi Organized Industrial Zone, 521 Street 7, 42250, Konya, Türkiye

ARTICLE INFO

Article history:

Received 30 July 2026

Accepted 14 September 2026

*Keywords:*CAM,
Deep learning,
Object detection,
RT-DETR,
Weld defect detection,
YOLOv10

ABSTRACT

Automated weld defect inspection is essential for improving product quality, reducing inspection costs, and minimizing human error in industrial manufacturing. Although deep learning-based object detectors have demonstrated promising performance, comprehensive comparisons under identical experimental conditions remain limited. This study presents a systematic evaluation of YOLOv10 and RT-DETR architectures for weld defect detection using the publicly available Weld Defects (WD) dataset. Five object detection models, namely YOLOv10-N, YOLOv10-L, YOLOv10-X, RT-DETR-L, and RT-DETR-X, were trained and evaluated under identical hyperparameter settings. In addition, the influence of different training-validation split ratios on YOLOv10 performance was investigated. Detection performance was assessed using Precision, Recall, F1-score, mAP@50, and mAP@50–95, while qualitative analyses were conducted through representative detection results and class activation mapping (CAM) visualizations. Experimental results showed that RT-DETR-X achieved the highest localization performance with a mAP@50–95 of 0.458 and the best F1-score, whereas YOLOv10-X obtained the highest precision (0.768) while maintaining competitive overall performance. Explainability analysis further demonstrated that RT-DETR-X generated more spatially coherent attention maps, whereas YOLOv10-X produced more localized but occasionally fragmented activations. The findings provide a comprehensive comparison of CNN- and Transformer-based detectors and offer practical guidance for selecting suitable deep learning models for reliable automated weld defect inspection in industrial applications.

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

1. INTRODUCTION

In industrial manufacturing, welding is widely adopted as a primary joining process owing to its high joint strength, economic efficiency, and capability to produce durable structures [1]. It plays a vital role in various industries, including automotive, aerospace, shipbuilding, construction, and pipeline engineering, where the quality of welded joints directly affects the structural integrity, reliability, and safety of the final products [2]. However, welding processes are susceptible to various defects that may compromise joint performance and lead to costly failures if not detected at an early stage. Conventional weld inspection methods rely heavily on manual visual examination and expert interpretation, making the inspection process time-consuming, labor-intensive, and

prone to human error [3]. Recent advances in deep learning have enabled the development of automated inspection systems that provide reliable and efficient weld defect detection from visual data. Consequently, numerous studies have investigated convolutional neural network (CNN) and transformer-based architectures to improve the accuracy and robustness of automated weld inspection. Some of the studies related to this research are presented below.

Zhao et al. used RT-DETR, an end-to-end object detector that eliminates the need for Non-Maximum Suppression (NMS). On the MS COCO dataset, RT-DETR-R50 achieved 53.1 AP with an inference latency of approximately 9.2 ms, outperforming several YOLO-based detectors in the accuracy-speed trade-off. The

* Corresponding author. E-mail address: ademdilbaz38@gmail.com
DOI: 10.58190/ijamec.2026.180

Transformer-based architecture provides strong global feature modeling capabilities compared with conventional CNN-based detectors. However, YOLO remains more suitable for resource-constrained and embedded applications due to its lower computational complexity [4].

In one study, Sivakumar et al. developed a weld defect classification framework based on Swin Transformer and DeiT using the SWRD dataset. Image denoising and enhancement techniques were applied to improve input quality before training. The proposed method achieved a 99.38% F1-score in binary classification and 98.30% F1-score in multi-class classification, demonstrating excellent performance in weld defect identification. Overall, the study showed that Transformer-based models can provide highly accurate and reliable weld defect classification [5].

Xu et al. used DBP-YOLO, a lightweight weld seam defect detection model based on YOLOv8n that enhances feature representation and multi-scale feature fusion through the proposed C2f-BiFusion and DFFN modules. The model also incorporates the Focal PIoU loss function to improve bounding box regression accuracy. Experimental results showed that DBP-YOLO achieved 83.5% mAP (mean Average Precision) at an Intersection over Union (IoU) threshold of 0.50, representing a 3.7% improvement over the baseline YOLOv8n while reducing the number of parameters by 1.21 million. These results demonstrate that the proposed model provides accurate and computationally efficient weld seam defect detection [6].

Parmar et al. evaluated YOLOv5, YOLOv8, YOLO11, and RT-DETR for weld defect detection using thermographic images acquired during TIG welding. The models were compared in terms of precision, recall, mAP, and inference latency. RT-DETR-X demonstrated superior feature extraction capability but incurred the highest inference latency, whereas YOLOv8 achieved the best overall balance between detection accuracy and computational efficiency. The results confirmed the effectiveness of thermography combined with deep learning for automated weld quality inspection [7].

Zhang et al. developed LMS-RTDETR, a lightweight Transformer-based model for X-ray weld defect detection in resource-constrained environments. The proposed framework integrates multi-scale feature aggregation, efficient attention, and an optimized NWD-Inner GIoU loss to improve detection performance. Experimental results showed a 3.10% improvement in mAP@50-95, while reducing the number of parameters by 38.34% and GFLOPs by 29.30%. These findings demonstrate that LMS-RTDETR achieves high detection accuracy with significantly lower computational complexity [8].

Peng et al. used YOLO-PDR, a lightweight defect detection network designed to improve the detection of small and low-contrast defects in complex textured environments. The proposed model integrates parallel

texture fusion, dilated block reparameterization, and a lightweight detail enhancement head to enhance feature representation while reducing computational complexity. Experimental results demonstrated a 97.4% mAP@50, with only 2.11 M parameters and 5.9 GFLOPs, while reducing the number of parameters and computational cost by 18.6% and 6.35%, respectively, compared with the baseline YOLOv11 model. These findings indicate that YOLO-PDR achieves an excellent balance between detection accuracy and computational efficiency [9].

In another study, Zhang et al. developed YOLO-Air, a lightweight YOLOv11-based model for weld surface defect detection. The proposed model enhances feature extraction while reducing computational complexity through optimized backbone, neck, and detection head designs. Experimental results showed a 1.3% improvement in mAP@50, while reducing the number of parameters by 17.3% and computational complexity by 31.7% compared with the baseline YOLOv11 model. These results demonstrate that YOLO-Air provides an efficient and accurate solution for weld surface defect detection [10].

Mafuwe et al. experimentally compared YOLO and RT-DETR for pest detection on yellow sticky traps by investigating the effects of learning rate, optimizer, and batch size. Their results showed that RT-DETR achieved its best performance using the SGD optimizer with a batch size of 32 and a learning rate of 0.01, reaching a mAP@50 of 0.963. In contrast, YOLO demonstrated greater robustness to higher learning rates and achieved perfect classification performance on the test set, with no observed misclassifications. Overall, both models achieved comparable detection performance after hyperparameter optimization [11].

Guo et al. proposed MCTNet, a CNN-Transformer hybrid network for aluminum alloy die-casting defect detection. The proposed model integrates multilevel feature interaction, super-resolution feature fusion, and an adaptive feature weighting detection head to improve the detection of multiscale and irregular defects. Experimental results showed 4.8% and 5.7% improvements in mAP@50 and mAP@50:90, respectively, compared with the baseline model. These findings demonstrate that MCTNet achieves superior detection accuracy over both CNN-based and Transformer-based approaches [12].

Zhang et al. proposed Weld-YOLO, a YOLO11-based model for weld X-ray defect detection. The model incorporates a dynamic gating pinwheel convolution (DGPCnv) and a dual attention block enhancement (DABE) module to improve feature extraction and multi-scale representation. These enhancements increase robustness against defects with varying sizes and orientations. Experimental results showed that Weld-YOLO achieved 97.5% mAP@0.5 and 93.4% recall. The proposed approach improved detection accuracy while

maintaining a lightweight architecture with only a slight increase in computational cost [13].

In a study, Yang et al. used a multi-image feature sequence fusion method based on YOLOv8 for radiographic casting defect detection. The approach employs gamma transformation to generate dynamic image sequences and a multi-image feature fusion network to enhance spatiotemporal feature representation, followed by confidence similarity-based localization refinement. The proposed method was experimentally compared with state-of-the-art detectors, including RT-DETR, YOLOv10, and YOLOv11, demonstrating superior detection performance. It achieved a 96.71% recall and 77.09% mAP, significantly reducing missed detections and improving localization accuracy for low-semantic defects [14].

Zhang et al. proposed FC-YOLO, an enhanced YOLOv7-based model for aircraft skin defect detection. The model integrates the Coordinate & Channel Attention (ELAN-C) module, an Adaptive-Path Aggregation Network (APAN) with ASFF, and the SCYLLA-IoU loss function to improve multi-scale feature fusion and localization accuracy. Experimental results on the ASD-DET and NEU-DET datasets showed that FC-YOLO outperformed both YOLOv8 and RT-DETR, achieving higher detection accuracy while improving the robustness of aircraft skin defect inspection [15].

Huan et al. proposed EF-YOLO, an enhanced lightweight defect detection network for quality inspection of *Paeoniae Radix Alba* slices. The proposed method integrates a self-developed detection head (LSD-Head) with a Feature Fusion Attention (FCA) mechanism to improve defect feature representation. Experimental comparisons were conducted with several state-of-the-art detectors, including YOLOv5, YOLOv8, YOLOv10, YOLOv8 variants, and RT-DETR. The proposed EF-YOLO achieved 74.1% mAP@0.5 with only 5.4 GFLOPs and 2.62 M parameters, providing competitive accuracy with significantly lower computational complexity [16].

Dilbaz and Ozkan previously investigated the YOLOv3-S and YOLOv5-M models for automated spot-weld defect detection within a simulation-based inspection framework [17]. In contrast, the present study provides a broader comparative evaluation of the CNN-based YOLOv10 and Transformer-based RT-DETR architectures for weld defect detection using a different and larger dataset of real-world images, while also considering different dataset split configurations and incorporating explainability analyses.

Dilbaz previously conducted a comparative evaluation of YOLO12 and YOLO13 architectures for steel surface defect detection using a fixed test sample and different validation ratios on a dataset of 1,800 images [18]. In contrast, the present study uses a different dataset consisting of approximately 5,200 images, incorporates

XAI analyses, and introduces a comparative assessment of CNN-based YOLOv10 and Transformer-based RT-DETR models. Moreover, the inclusion of X-variants enables a more comprehensive investigation of detection accuracy and computational efficiency under industrial inspection scenarios.

Among recent YOLO-based detectors, YOLOv10, originally developed by researchers from Tsinghua University and later adopted by the Ultralytics framework, introduces an end-to-end object detection architecture with consistent dual label assignment and NMS-free inference [19]. These improvements enhance detection efficiency and accuracy, making YOLOv10 suitable for industrial visual inspection applications.

Despite the considerable progress achieved by recent YOLO variants and Transformer-based detectors such as RT-DETR, most existing studies focus on improving individual architectures or evaluating limited model versions. Comparative analyses between advanced YOLO-based detectors and Transformer-based approaches under comparable experimental conditions remain limited, particularly for weld surface defect detection. To address this research gap, this study presents a comprehensive evaluation of the Ultralytics YOLOv10 and RT-DETR models using the publicly available Weld Defects (WD) dataset under identical experimental settings and hyperparameters. Furthermore, the effect of different training-validation split ratios on YOLOv10 performance is systematically investigated. The evaluated models are compared using standard object detection metrics, including precision, recall, F1-score, mAP@50, and mAP@50–95, to analyze their accuracy, robustness, and suitability for industrial weld inspection applications.

The main contributions of this study can be summarized as follows:

- A comprehensive comparative evaluation of YOLOv10 and RT-DETR architectures is conducted for weld defect detection using the same dataset and identical experimental conditions.
- The effect of different training-validation split ratios on YOLOv10 performance is systematically investigated to analyze their effect on evaluation stability and detection performance.
- The evaluated models are compared using widely adopted object detection metrics, including precision, recall, F1-score, mAP@50, and mAP@50–95, providing a comprehensive analysis of detection capability.
- The trade-off between detection accuracy and computational efficiency is investigated by considering model complexity and inference performance for practical industrial inspection scenarios.
- The study provides practical insights into the selection of suitable deep learning-based detectors and data partition strategies for reliable weld inspection systems. Furthermore, the trained RT-DETR model weights are

made available to support reproducibility and future research.

- XAI analysis is performed using Grad-CAM visualizations to examine the spatial attention patterns of the evaluated models and provide complementary insights into their decision-making behavior in challenging weld defect scenarios.

The remainder of this paper is organized as follows. Section 2 describes the materials and methods, including the dataset, model architectures, experimental settings, and evaluation metrics. Section 3 presents the experimental results, followed by a detailed discussion of the findings in Section 4. Finally, the conclusions and future research directions are presented in Section 5.

2. Materials and Methods

2.1. Weld Defects Dataset

The WD dataset, a publicly available benchmark dataset for weld defect detection, was used in this study. The dataset contains a total of 5,198 annotated images obtained from the Roboflow Universe platform, including 4,797 training images, 279 validation images, and 122 test images in its original configuration [20]. In this study, for all evaluated train–validation split ratios (46:3:1, 46:6:1, and 46:9:1), the same independent test set of 122 images was kept unchanged. It consists of three classes: Good Weld, Bad Weld, and Defect, representing different weld conditions and defect categories.

In the original WD dataset configuration, approximately 92.3%, 5.4%, and 2.3% of the images were allocated to the training, validation, and testing sets, respectively, corresponding to a 46:3:1 split ratio. However, to systematically investigate the effect of training–validation partitioning on detection performance, the dataset was reorganized using different training–validation split ratios while maintaining the same experimental conditions. These alternative train–validation configurations were used to analyze their influence on model evaluation reliability, generalization capability, and detection performance.

The dataset includes weld images containing both defective and non-defective regions, enabling the evaluation of deep learning-based object detection models for automated weld inspection. Representative examples from the dataset are presented in Figure 1.



Figure 1. Sample Images from the WD Dataset: Defect, Bad Weld, and Good Weld, Respectively

The images in the WD dataset are provided in .jpg format and represented in the RGB color space. Each image contains annotated weld regions belonging to one or more of three classes: Class 0 (Bad Weld), Class 1 (Defect), and Class 2 (Good Weld). A total of 3,424 images contain Class 0, 2,456 images contain Class 1, and 3,208 images contain Class 2. Since a single image may contain instances from multiple classes, these image counts are not mutually exclusive. The dataset exhibits substantial variations in weld appearance, shape, illumination, background conditions, and defect characteristics, providing a challenging benchmark for the training and evaluation of deep learning-based weld quality inspection models.

2.2. YOLOv10 Architecture

YOLOv10 is a state-of-the-art one-stage object detection model designed to achieve high detection accuracy while minimizing computational cost and inference latency [21]. The architecture introduces an end-to-end detection framework that eliminates redundant post-processing operations, enabling efficient object detection. By combining effective feature extraction, multi-scale feature fusion, and an optimized prediction mechanism, YOLOv10 provides robust performance across a wide range of object detection tasks, including industrial visual inspection and automated weld defect detection. The network consists of three main components: a backbone that extracts hierarchical feature representations from the input image, a neck that fuses multi-scale features to improve the detection of objects with different sizes, and a detection head that directly predicts object classes, confidence scores, and bounding box coordinates in an end-to-end manner [22].

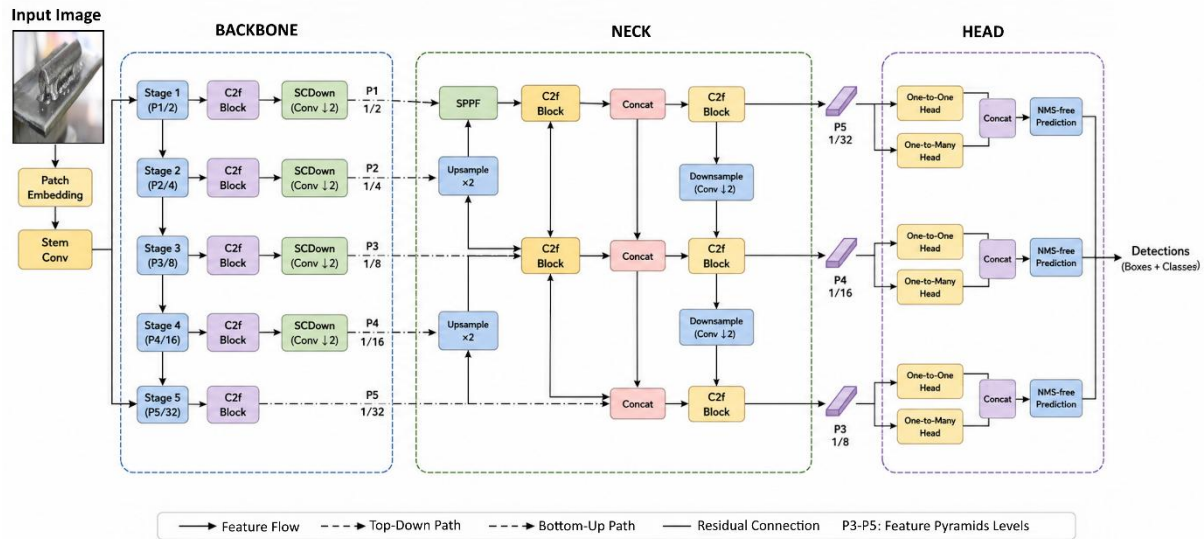


Figure 2. Layers and Implementation Stages of the YOLOv10 Architecture

Figure 2 illustrates the overall architecture of the YOLOv10 model. The backbone extracts hierarchical feature representations using convolutional layers, C2f blocks, and SCDown modules for efficient spatial downsampling [23]. The extracted features are then fused in the neck through an FPN-PAN structure, where the SPPF module enlarges the receptive field and enhances multi-scale feature representation with low computational cost [24]. Finally, the end-to-end NMS-free detection head directly predicts object classes, confidence scores, and bounding box coordinates without redundant post-processing. The decoupled prediction branches improve classification and localization performance, while the SiLU activation function facilitates efficient feature learning and CIoU loss contributes to accurate bounding box regression during training [25].

2.3. RT-DETR with HGNetv2 (High-Performance GPU Network v2) Architecture

RT-DETR is a Transformer-based object detector that combines convolutional feature extraction with Transformer-based object decoding in an end-to-end detection framework. Unlike conventional CNN-based detectors, RT-DETR eliminates heuristic post-processing steps and directly predicts object classes and bounding boxes through a Transformer decoder, achieving an effective balance between detection accuracy and inference speed [26].

The backbone of RT-DETR is based on HGNetv2, a lightweight convolutional architecture specifically designed for efficient GPU inference. Compared with the widely used ResNet backbone, HGNetv2 adopts hierarchical aggregation blocks that improve feature reuse while reducing computational redundancy. Instead of relying on residual bottleneck structures with numerous skip connections, HGNetv2 employs efficient convolutional operations and optimized feature

aggregation, enabling stronger multi-scale feature extraction with lower computational complexity. As a result, HGNetv2 achieves a better trade-off between detection accuracy and inference time, making it particularly suitable for industrial vision applications.

Following the backbone, RT-DETR utilizes a Hybrid Encoder to process multi-scale feature maps. The Hybrid Encoder integrates Attention-based Intra-scale Feature Interaction (AIFI) and Cross-scale Cross Feature Fusion Module (CCFM) to effectively capture both local details and global contextual information while maintaining computational efficiency. Finally, a Transformer decoder performs end-to-end object prediction by directly estimating object classes and bounding box coordinates without requiring NMS, thereby reducing inference latency and improving deployment efficiency [27].

Table 1. Comparison of the Architectural Characteristics of YOLOv10 and RT-DETR

Feature	YOLOv10	RT-DETR with HGNetv2
Core Design	CNN-based one-stage detector with end-to-end NMS-free detection	Transformer-based end-to-end object detector
Backbone	Conv, C2f, and SCDown modules	High-Performance GPU Network v2
Neck	FPN-PAN with SPPF for multi-scale feature fusion	Hybrid Encoder with AIFI and CCFM modules
Detection Head	Decoupled NMS-free detection head	Transformer Decoder with object queries
Detection Strategy	Dense prediction with dual assignment	Query-based object prediction
Feature Representation	Multi-scale convolutional feature extraction and fusion	Hybrid CNN–Transformer representation with global contextual modeling

Table 1 summarizes the key architectural differences between YOLOv10 and RT-DETR with the HGNetv2 backbone. YOLOv10 employs a CNN-based architecture

with an end-to-end NMS-free detection head, whereas RT-DETR combines the HGNetv2 backbone, a Hybrid Encoder, and a Transformer decoder for end-to-end object detection. To quantitatively evaluate the performance of these architectures, several standard object detection metrics were employed. In addition, confusion matrices were analyzed to provide a detailed assessment of class-wise prediction performance.

2.4. Normalized Confusion Matrix

In multi-class object detection, evaluating the prediction performance of individual classes is essential for identifying classification errors and assessing the overall reliability of the detection model. To provide a detailed class-wise analysis, the confusion matrix is commonly employed as a complementary evaluation tool. By examining the relationship between actual and predicted class labels, it helps reveal both correctly classified samples and misclassification patterns among different classes.

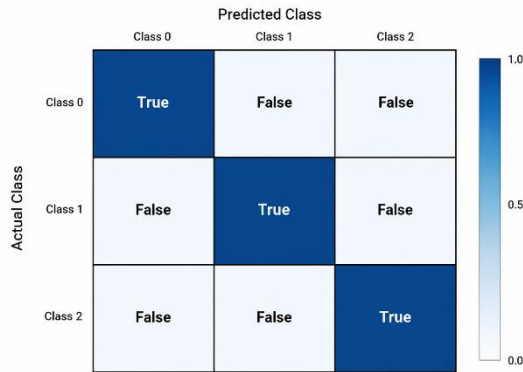


Figure 3. Normalized Confusion Matrix of Classification Results

The confusion matrix summarizes the prediction results for each class by comparing the predicted labels with the corresponding ground-truth labels. It can be presented either as raw counts or in a normalized form. As shown in Figure 3, the normalized confusion matrix illustrates the relationship between ground-truth and predicted classes. The values are expressed as proportions ranging from 0 to 1, allowing a more intuitive comparison of classification performance across classes regardless of differences in sample size. Additionally, off-diagonal entries represent instances in which samples are incorrectly assigned to other classes, providing insight into class-specific confusion patterns.

2.5. Performance Metrics

To quantitatively assess the detection performance of the proposed models, standard evaluation metrics derived from the confusion matrix were employed. The confusion matrix compares the predicted labels with the corresponding ground-truth labels and summarizes the numbers of true positives (TP), false positives (FP), and

false negatives (FN). Based on these values, four widely adopted object detection metrics, namely Precision, Recall, F1-score, and mAP, were used to comprehensively evaluate the classification and localization performance of the YOLOv10 and RT-DETR models. The mathematical formulations and corresponding descriptions of these evaluation metrics are summarized in Table 2.

Table 2. Performance Metrics, Formulas, and Explanation

Performance Metrics	Mathematical Formulas	Explanation
Precision (P)	$\frac{TP}{TP + FP}$	Measures the accuracy of positive predictions by indicating how many detected objects are correctly classified
Recall (R)	$\frac{TP}{TP + FN}$	Measures the ability of the model to identify all relevant objects in the dataset
F1-score	$\frac{2 \times P \times R}{P + R}$	Represents the harmonic mean of precision and recall, balancing both metrics in a single score
mAP	$\frac{1}{N_c} \times \sum_{i=1}^{N_c} AP_i$	Represents the mean detection performance by averaging the average precision across all object classes.

Average Precision (AP) measures the detection performance for a single object class by calculating the area under the precision–recall (P–R) curve [28]. The mAP is obtained by averaging the AP values across all object classes. Depending on the evaluation protocol, mAP can be calculated at a single IoU threshold, such as mAP@50 (IoU = 0.50), or averaged over multiple IoU thresholds ranging from 0.50 to 0.95 with a step size of 0.05 (mAP@50–95), providing a more comprehensive assessment of detection performance.

2.6. YOLO and RT-DETR Variants

YOLOv10 is provided in multiple model variants, including nano (N), small (S), medium (M), large (L), and extra-large (X). Although these variants share the same network architecture, they differ in terms of network depth, width, number of parameters, and computational complexity. Smaller variants are computationally efficient and well suited for resource-constrained environments, whereas larger variants offer greater feature representation capacity and generally achieve higher detection accuracy on more challenging datasets. The RT-DETR family includes large and extra large variants, offering different trade-offs between detection accuracy and computational efficiency. Unlike the YOLO family, RT-DETR does not offer smaller model scales in its official model family, as its architecture primarily targets high-performance object detection.

3. Experimental Results

All experiments were conducted in the Google Colab environment. The experimental configurations adopted in this study are presented in Table 3. All models were optimized using the AdamW optimizer with an initial

learning rate of 0.001 and a momentum value of 0.937. Weight decay was applied through parameter groups, with zero decay assigned to bias and normalization-related parameters and a decay factor of 0.0001 applied to weight parameters. The batch size was fixed at 16 throughout all experiments. The models were trained for 50 and 100 epochs using the same optimization strategy to ensure a fair and unbiased comparison among the evaluated models.

Table 3. Experimental setup configuration for all models

Component	Specification
Operating System	Google Colab (Linux)
GPU	NVIDIA Tesla T4 (15 GB VRAM)
Python	3.12.13
PyTorch	2.11.0 + CUDA 12.8 (torch-2.11.0+cu128)
CUDA	CUDA 12.8
Ultralytics	V8.4.105

To ensure a fair and consistent comparison, all baseline models were initialized using the official pretrained weights provided by the Ultralytics framework (v8.4.105). Each model retained its original network architecture without any architectural modifications. All models were trained and evaluated under identical experimental conditions.

The comparative analysis of the evaluated models in terms of the number of parameters and FLOPs is presented in Table 4. The evaluated models include the YOLOv10-X, YOLOv10-L, YOLOv10-N, RT-DETR-L, and RT-DETR-X architectures, representing different levels of model capacity and computational complexity. All models were initialized using official pretrained weights and trained on the same weld defect dataset under identical experimental conditions and hyperparameter settings. The architectural characteristics provide insights into the trade-offs between model complexity and computational efficiency. The comparative evaluation results based on standard object detection metrics are presented in the following section.

For clarity and to facilitate reproducibility, the exact numbers of images and corresponding approximate percentage allocations for the training, validation, and test sets under each split configuration are summarized in Table 5, providing a detailed overview of the dataset

distributions used across all evaluations.

Table 4. Architectural Characteristics of the YOLOv10 and RT-DETR

Model	Pre-trained	Layers	Params (M)	Flops (G)
YOLOv10-N	yolov10n	272	2.70	8.4
YOLOv10-L	yolov10l	365	25.77	127.2
YOLOv10-X	yolov10x	401	31.66	171.0
RT-DETR-L	rt detr-l	470	32.81	108.0
RT-DETR-X	rt detr-x	592	67.3	232.3

Table 5. WD Dataset Configurations and Split Ratios Used Across All Evaluations

	Configuration 1	Configuration 2	Configuration 3
Train–Validation–Test Ratio	46:3:1	46:6:1	46:9:1
Training Images	4,797	4,490	4,246
Validation Images	279	586	830
Test Images	122	122	122
Training (%)	92.3	86.4	81.7
Validation (%)	5.4	11.3	16.0
Test (%)	2.3	2.3	2.3

Following training under identical experimental conditions, the performance of all models was evaluated using standard object detection metrics. The comparative quantitative results, detailed performance analysis, and qualitative evaluations are presented in the following section.

The YOLOv10-N model was trained and evaluated under identical experimental settings using an input image size of 640×640 pixels. The dataset was divided into training, validation and test sets using three different split ratios: 46:3:1, 46:6:1, and 46:9:1. Model performance was quantitatively assessed using standard object detection metrics, while qualitative comparisons were conducted using representative validation images. The confusion matrices obtained from all YOLOv10-N experiments are presented in Figure 4.

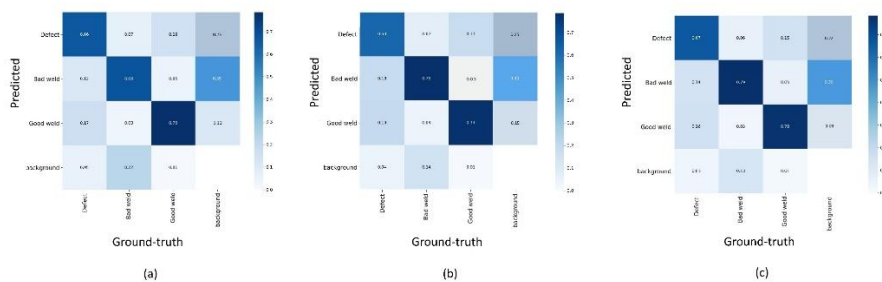


Figure 4. Normalized Confusion Matrices of the YOLOv10-N under different dataset split ratios: (a) 46:3:1, (b) 46:6:1, and (c) 46:9:1

Among the evaluated dataset split ratios, the 46:9:1 configuration achieved the best overall classification

performance. It produced the highest diagonal values for all three defect classes, indicating improved class discrimination and reduced inter-class confusion. In contrast, the 46:6:1 split showed a noticeable decrease in the Defect class accuracy, while the 46:3:1 split yielded comparatively lower overall performance. These results demonstrate that the 46:9:1 split provides the most balanced and reliable classification performance for YOLOv10-N. Following this confusion matrix analysis, a comprehensive quantitative evaluation was performed using the Precision, Recall, and mAP@50–95 metrics. The test results of the five evaluated object detection models, namely YOLOv10-N, YOLOv10-L, YOLOv10-X, RT-DETR-L, and RT-DETR-X, under identical experimental conditions are summarized below.

Table 6. Comparative Precision, Recall, and mAP@50–95 Performance of YOLO- and RT-DETR-Based Models

Model	Epoch	Precision	Recall	mAP@50-95
YOLOv10-N	50	0.691	0.574	0.369
YOLOv10-N	100	0.706	0.644	0.401
YOLOv10-L	50	0.722	0.614	0.416
YOLOv10-L	100	0.759	0.699	0.448
YOLOv10-X	50	0.719	0.601	0.410
YOLOv10-X	100	0.768	0.702	0.453
RT-DETR-L	50	0.681	0.695	0.405
RT-DETR-L	100	0.739	0.715	0.450
RT-DETR-X	50	0.690	0.679	0.432
RT-DETR-X	100	0.751	0.709	0.458

Table 6 summarizes the quantitative performance of the five evaluated models under both 50- and 100-epoch training settings. Among these models, YOLOv10-N consistently exhibited the lowest overall detection performance across the evaluated metrics. In contrast, YOLOv10-X and RT-DETR-X achieved the most competitive overall results. YOLOv10-X attained the highest precision (0.768), whereas RT-DETR-X achieved

the highest mAP@50–95 (0.458). Notably, RT-DETR-L produced the highest recall (0.715), indicating its superior ability to identify defect instances. Although YOLOv10-L and RT-DETR-L also demonstrated strong detection performance, their overall results were slightly inferior to those of YOLOv10-X and RT-DETR-X. Therefore, the subsequent analyses primarily focus on the comparison among YOLOv10-N, YOLOv10-X, and RT-DETR-X, including the F1-score–confidence curves, Precision–Recall curves, qualitative detection results, and explainability evaluations. YOLOv10-N was included as the baseline model, whereas YOLOv10-X and RT-DETR-X were selected because they represent the highest-performing YOLO-based and Transformer-based object detection architectures, respectively.

Figure 5 presents the F1-score–confidence curves for the selected models under identical training and hyperparameter settings after 50 epochs. Among the three models, RT-DETR-X achieved the highest overall F1-score, reaching 0.68 at a confidence threshold of 0.572, indicating superior balance between precision and recall. YOLOv10-X ranked second with a maximum F1-score of 0.65 at a confidence threshold of 0.374. YOLOv10-N achieved peak F1-scores of 0.63 at confidence thresholds of 0.266. Class-wise analysis further shows that the Good Weld category consistently attained the highest F1-score across all models, whereas the Defect class remained the most challenging, exhibiting substantially lower F1 values throughout the confidence range. In addition, RT-DETR-X maintained comparatively higher and more stable F1 performance over a broader confidence interval, suggesting greater robustness to confidence threshold selection than the YOLOv10 variants.

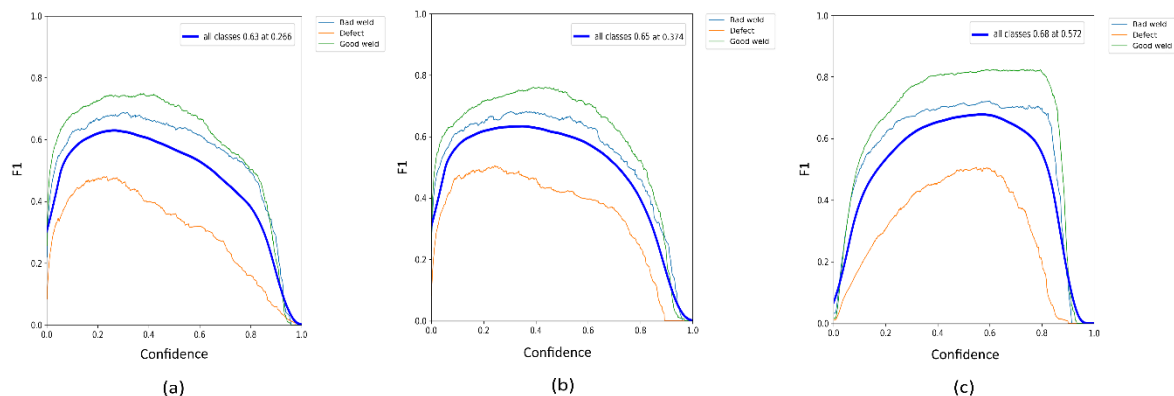


Figure 5. F1-score–confidence curves for the selected models under identical hyperparameter settings: (a) YOLOv10-N, (b) YOLOv10-X, and (c) RT-DETR-X

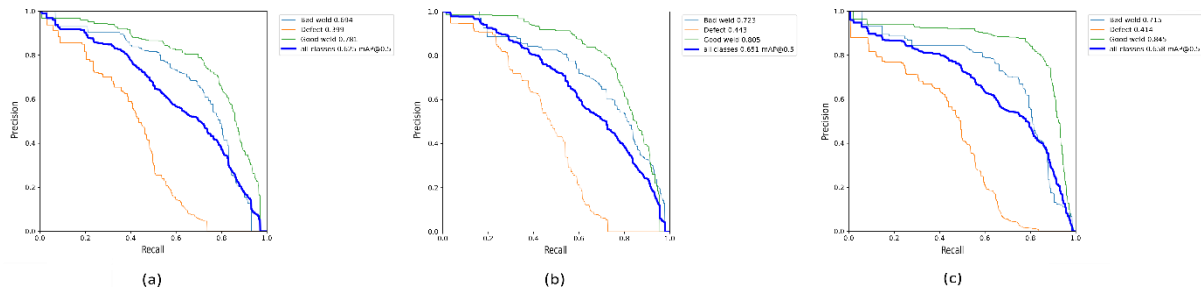


Figure 6. Precision-Recall curves for the selected models under identical hyperparameter settings: (a) YOLOv10-N, (b) YOLOv10-X, and (c) RT-DETR-X

Figure 6 presents the Precision–Recall (PR) curves for the selected object detection models under identical training and hyperparameter settings after 50 epochs. Overall, RT-DETR-X achieved the highest detection performance, obtaining an overall mAP@0.50 of 0.658, followed by YOLOv10-X (0.651) and YOLOv10-N (0.625). Class-wise analysis indicates that the Good Weld category consistently produced the highest average precision across all models, increasing from 0.781 for YOLOv10-N to 0.805 for YOLOv10-X and reaching 0.845 with RT-DETR-X. Similarly, the Bad Weld class demonstrated gradual improvement, with AP values of 0.694, 0.723, and 0.715, respectively. In contrast, the Defect category remained the most challenging class, although its AP increased from 0.399 (YOLOv10-N) to 0.443 (YOLOv10-X) and 0.414 (RT-DETR-X). Furthermore, the PR curve of RT-DETR-X remained closer to the upper-right corner over a wider recall range, indicating a more favorable balance between precision and recall. These results demonstrate that the Transformer-based RT-DETR-X provides superior overall localization and classification performance on the weld defect dataset while maintaining more robust detection characteristics across varying recall levels.

Figure 7 presents a qualitative comparison of the detection results produced by YOLOv10-X and RT-DETR-X on representative weld defect samples. Overall, both models successfully identify the primary defect regions; however, several notable differences are observed. In sample (a), YOLOv10-X produces duplicate detections for the same defect, whereas RT-DETR-X generates a single, well-localized bounding box. In sample

(c), RT-DETR-X introduces an additional false-positive detection by identifying a non-existent defect in the lower image region, while YOLOv10-X correctly localizes only the actual defect. In samples (b) and (d), both models accurately detect the weld defects with comparable localization performance. In sample (e), both models identify the defective weld, although YOLOv10-X predicts an additional overlapping bounding box, resulting in redundant detections. These observations indicate that YOLOv10-X is more susceptible to duplicate detections, whereas RT-DETR-X exhibits a tendency to generate occasional false-positive predictions in challenging background regions. Despite these limitations, both models demonstrate reliable qualitative detection performance across representative welding scenarios.

As shown in Figure 8, YOLOv10-X generally produces more localized activation maps, whereas RT-DETR-X captures defect regions more comprehensively. In particular, samples (d) and (e) demonstrate that RT-DETR-X focuses more consistently on the weld defect, while YOLOv10-X exhibits either scattered background attention (d) or limited attention confined to a small portion of the weld joint (e). These results suggest that RT-DETR-X provides more coherent explanations in challenging welding scenarios. However, Grad-CAM visualizations may not always accurately reflect the actual decision-making process of the model, and some activation patterns may not correspond precisely to the true defect regions. Therefore, these visualizations should be interpreted as complementary evidence rather than definitive representations of model reasoning.

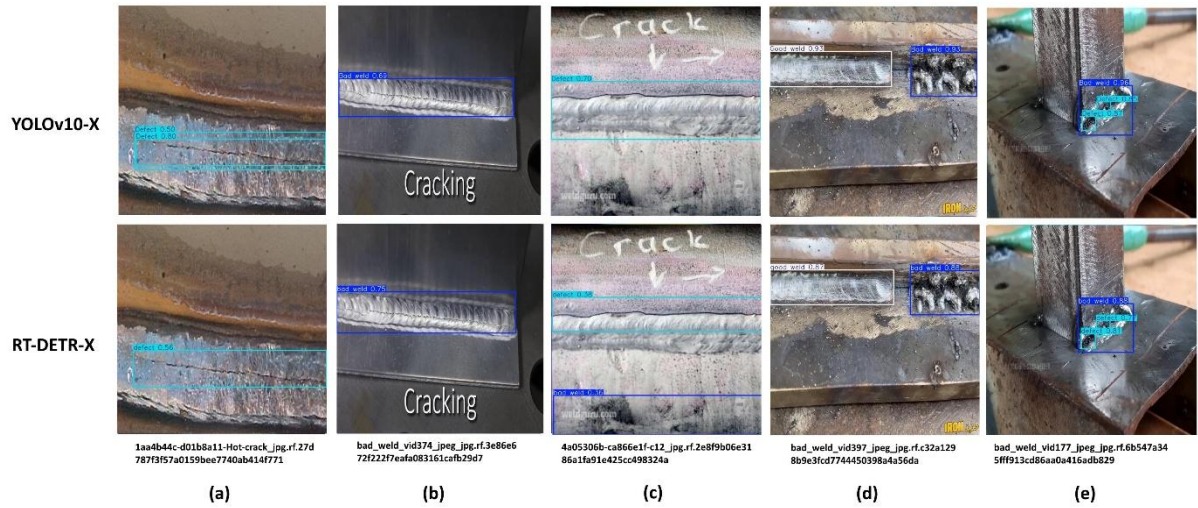


Figure 7. Qualitative comparison of object detection results obtained by YOLOv10-X and RT-DETR-X on representative weld defect images.

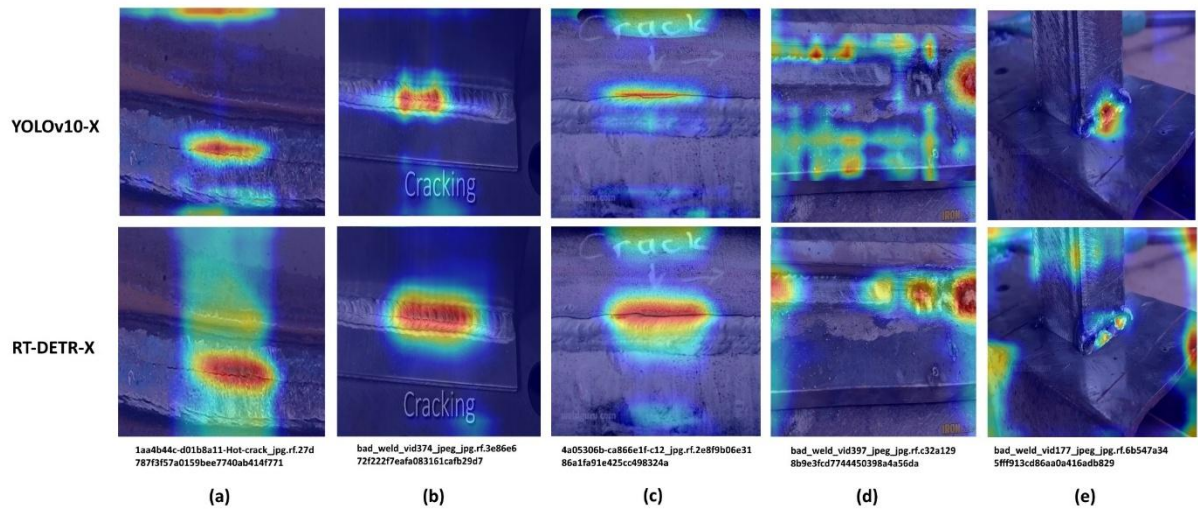


Figure 8. Comparative Grad-CAM visualizations of YOLOv10-X and RT-DETR-X on representative weld defect images.

4. Results and Discussion

The present study conducted a comprehensive comparative evaluation of five object detection models, namely YOLOv10-N, YOLOv10-L, YOLOv10-X, RT-DETR-L, and RT-DETR-X, under identical experimental conditions using the WD dataset. Among the evaluated models, RT-DETR-X achieved the strongest overall detection performance, outperforming YOLOv10-X by approximately 1.1% in relative terms $mAP@50-95$ while also attaining the highest F1-score. As illustrated by the Precision–Recall curves (Figure 6), scaling the YOLOv10 architecture from the N to the L variant resulted in a 4.2% improvement in $mAP@50$, followed by an additional 1.1% improvement with the X variant. Similarly, the F1–confidence analysis (Figure 5) showed that the optimal F1-score increased by 3.2% from YOLOv10-N to YOLOv10-L and by a further 4.6% from YOLOv10-L to YOLOv10-X, yielding an overall improvement of approximately 7.9% across the YOLOv10 family. These findings demonstrate that increasing the model capacity

consistently improved the balance between precision and recall while enhancing detection robustness.

The analysis of different dataset partitioning strategies demonstrated that the 46:9:1 training–validation split provided the most reliable classification performance for YOLOv10-N. Specifically, the classification accuracy of the Bad Weld class improved by 10.3% when the training–validation split was changed from 46:3:1 to 46:6:1, followed by a further 5.3% improvement under the 46:9:1 configuration. Overall, the classification accuracy of the Bad Weld class increased by 16.2% compared with the initial configuration. Among the evaluated split configurations, the 46:9:1 split achieved the highest observed classification accuracy while maintaining comparable overall detection performance.

The comparative evaluation further revealed different characteristics of the YOLOv10 and RT-DETR architectures. RT-DETR-X consistently achieved superior localization performance, as reflected by the highest $mAP@50-95$ and the most stable F1-score across different confidence thresholds. In contrast, YOLOv10-X

demonstrated the highest precision, indicating fewer false-positive predictions overall. These findings suggest that the Transformer-based architecture benefits from stronger global contextual modeling, whereas the CNN-based YOLO architecture provides highly reliable object classification with efficient feature extraction.

The quantitative results reveal distinct characteristics of the YOLOv10 and RT-DETR architectures. RT-DETR-X consistently achieved superior localization performance, as reflected by the highest mAP@50–95 and the most stable F1-score across different confidence thresholds. In contrast, YOLOv10-X achieved the highest precision, indicating fewer false-positive predictions while maintaining competitive overall accuracy. These findings suggest that the Transformer-based architecture benefits from stronger global feature modeling, enabling more accurate localization of complex and elongated weld defects, whereas the YOLO-based architecture provides reliable classification through efficient feature extraction.

The Grad-CAM visualizations (Figure 8) provide additional insight into the decision-making behavior of the evaluated models. YOLOv10-X generally produced localized activation maps concentrated around the predicted defect regions. In contrast, RT-DETR-X generated more spatially coherent attention maps covering a larger portion of the weld defect, particularly in challenging samples with irregular defect boundaries. For example, RT-DETR-X consistently focused on the complete defective weld region in representative samples, whereas YOLOv10-X occasionally produced fragmented activations or concentrated on only a limited portion of the defect. These observations indicate that the Transformer-based architecture captures global contextual information more effectively, resulting in more comprehensive and spatially consistent visual explanations. Nevertheless, a more comprehensive assessment of model interpretability could be achieved by comparing Grad-CAM with additional CAM methods, such as Grad-CAM++, Eigen-CAM, or related XAI approaches. Such a multi-method comparison would provide a more robust evaluation of the consistency and reliability of the highlighted defect regions and strengthen the interpretability analysis of the evaluated models.

Overall, the experimental results demonstrate that both YOLOv10 and RT-DETR provide reliable solutions for automated weld defect detection. RT-DETR-X achieved the highest overall detection and localization performance, whereas YOLOv10-X offered competitive accuracy together with the highest precision. The complementary quantitative, qualitative, and explainability analyses consistently support these findings, providing practical guidance for selecting appropriate object detection architectures according to the accuracy requirements and deployment constraints of industrial weld inspection systems. Future work will focus on validating the proposed

framework on additional weld inspection datasets and improving its suitability for practical industrial deployment through inference latency analysis, model compression, and hardware-specific optimization techniques.

5. Conclusions

This study presented a comprehensive experimental evaluation of YOLOv10 and RT-DETR architectures for automated weld defect detection using the publicly available Weld Defects (WD) dataset. Five models, namely YOLOv10-N, YOLOv10-L, YOLOv10-X, RT-DETR-L, and RT-DETR-X, were evaluated under identical training and evaluation conditions. In addition, the influence of different training–validation split ratios on YOLOv10 performance was investigated to assess their effect on model reliability.

The experimental results demonstrated that RT-DETR-X achieved the strongest overall detection performance, obtaining the highest mAP@50–95 (0.458) and F1-score (0.68), while YOLOv10-X achieved the highest precision (0.768). These findings indicate that RT-DETR-X provides stronger localization and a more favorable precision–recall balance, whereas YOLOv10-X offers an effective alternative when precision and false-positive reduction are prioritized.

The investigation of dataset partitioning showed that the 46:9:1 training–validation–test configuration provided the most reliable classification performance for YOLOv10-N. Compared with the original 46:3:1 configuration, the overall classification accuracy increased by 16.2%. These results suggest that the choice of training–validation split can influence the observed classification performance of the model.

The qualitative and explainability analyses further supported the quantitative results. RT-DETR-X generally produced more spatially coherent Grad-CAM activation maps and captured larger portions of defective weld regions, whereas YOLOv10-X generated more localized activations. Qualitative results also indicated that YOLOv10-X was more prone to duplicate detections, while RT-DETR-X occasionally produced false-positive predictions in challenging background regions.

Overall, both YOLOv10 and RT-DETR demonstrated their suitability for automated weld defect inspection, while exhibiting different strengths. RT-DETR-X provided the best overall localization and F1 performance, whereas YOLOv10-X achieved the highest precision. Therefore, model selection should be based on the specific requirements of the target industrial inspection system.

As future work, the evaluated models will be tested on additional weld-defect datasets and real-world industrial inspection scenarios to further assess their generalization and robustness. Future studies will also investigate

inference efficiency, model compression, and hardware-specific optimization to improve the suitability of the models for real-time industrial deployment. In addition, advanced explainable artificial intelligence techniques will be explored to provide deeper insights into model decision-making and to quantitatively assess the correspondence between model attention and actual weld-defect regions.

Declaration of Ethical Standards

The article does not contain any studies with human or animal subjects.

Credit Authorship Contribution Statement

Adem Dilbaz: Conceptualization, Methodology, Software, Investigation, Data Curation, Formal Analysis, Visualization, Writing – Original Draft, Writing – Review & Editing.

Declaration of Competing Interest

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding / Acknowledgements

This research received no external funding.

Data Availability

<https://universe.roboflow.com/yolov8-daldb/weld-defects-mlopr>

References

- [1] X. Chen, Y. Duan, Z. Chen, and X. Chen, "Multimodal defect detection in CMT welding based on melt pool images and acoustic signals," in 2025 IEEE International Conference on Robotics and Biomimetics (ROBIO), 2025, pp. 330–335.
- [2] G. Chudasama, S. Kashyap, J. Patel, V. Kalyankar, and J. Shewale, "AISI 2205 DSS for body-in-white automotive structure: Effect of welding current and time on the performance of spot welded joint," *Welding International*, vol. 36, no. 12, pp. 769–778, 2022.
- [3] V. Gaidhane and S. Kolhe, "Enhanced weld defect detection through advanced image preprocessing techniques," in 2025 International Conference on Electronics, Computing, Communication and Control Technology (ICECCC), 2025, pp. 1–6.
- [4] F. Zuo, J. Liu, M. Fu, L. Wang, and Z. Zhao, "An efficient anchor-free defect detector with dynamic receptive field and task alignment," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 6, pp. 8536–8547, 2024.
- [5] B. Xiao, Y. Yang, Y. He, and G. Ma, "Multi-type weld defect detection in galvanized sheet MIG welding using an improved YOLOv10 model," *Materials*, vol. 19, no. 6, Art no. 1178, 2026.
- [6] C. Zhang, C. Xu, W. Shan, and Z. Han, "DBP-YOLO: A dynamic bifurcated-path fusion network for weld seam defect detection," *IEEE Access*, vol. 13, pp. 163749–163768, 2025, doi: 10.1109/ACCESS.2025.3609890
- [7] K. Parmar, A. Biber, M. Olesch, R. Sharma, and U. Reisgen, "Leveraging thermal images for automatic surface defect detection in TIG welding process of thin metal sheets," *Welding in the World*, vol. 1, pp. 1–17, 2026.
- [8] X. Zhang, H. Chen, Z. Liu, and S. Anli, "Lightweight DETR algorithm for X-ray weld defect detection," *Scientific Reports*, vol. 15, Art no. 41873, 2025, doi: 10.1038/s41598-025-25916-2.
- [9] S. Peng, H. Bao, J. Li, and J. Gao, "YOLO-PDR: A lightweight defect detection network for hot-pressed light guide plates based on parallel texture fusion and multi-scale feature expression," *Measurement Science and Technology*, vol. 37, no. 12, p. 125902, 2026.
- [10] R. Zhang, C. Pan, Z. Chen, J. Fan, Z. Chen, and B. Yuan, "An improved lightweight YOLOv11 algorithm for weld surface defect detection," *Scientific Reports*, vol. 16, no. 1, Art no. 11440, 2026.
- [11] K. Mafuwe, R. V. S. Dulam, C. Kambhamettu, and M. S. Crossley, "Comparative hyperparameter optimization of object detection models for precision monitoring of cucumber beetles and similar insects on yellow sticky cards," *Scientific Reports*, vol. 16, Art no. 22099, 2026, doi: 10.1038/s41598-026-51483-1.
- [12] C. Guo, Y. Du, Y. Yang, S. Wang, D. Liu, and M. Cong, "MCTNet: Defect detection in aluminum alloy die-casting using a multilevel CNN-transformer interaction network," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–11, 2025, doi: 10.1109/TIM.2025.3589704.
- [13] Y. Zhang, H. Zhang, T. Ma, Z. Xu, J. Jiang, and F. Xue, "Weld-YOLO: A YOLOv11-based model for high accuracy defect detection in weld X-ray images," in *Proc. SPIE*, vol. 14247, 2026, Art. no. 1424714.
- [14] D. Yang, H. Jiang, H. Yang, Y. Wang, Z. Liu, X. Zhang, and H. Cheng, "Casting defect-detection method based on multi-image feature of fusion artificial intelligence model," *Applied Intelligence*, vol. 55, no. 11, p. 774, 2025.
- [15] W. Zhang, J. Liu, Z. Yan, M. Zhao, X. Fu, and H. Zhu, "FC-YOLO: An aircraft skin defect detection algorithm based on multi-scale collaborative feature fusion," *Measurement Science and Technology*, vol. 35, no. 11, p. 115405, 2024.
- [16] M. Huan, T. Wang, Q. Xu, L. Gao, Z. Lu, X. Shi, M. Qian, L. Jia, and C. Yao, "An efficient lightweight object detection algorithm for defect identification in Paeoniae Radix Alba with rapid area and diameter estimation using OpenCV," *Food Science & Nutrition*, vol. 13, no. 9, Art. no. e70947, 2025.
- [17] A. Dilbaz and İ. A. Ozkan, "Simulation-based spot welding inspection on automotive chassis using YOLO-powered image processing," *International Journal of Automotive Engineering and Technologies*, vol. 14, no. 3, pp. 170–180, 2025, doi: 10.18245/ijaet.1729908.
- [18] A. Dilbaz, "YOLOv12 and YOLOv13 for steel surface defect detection: Experimental evaluation," *Intelligent Methods in Engineering Sciences*, vol. 5, no. 2, pp. 55–68, 2026, doi: 10.58190/imiens.2026.175.
- [19] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "YOLOv10: Real-time end-to-end object detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 107984–108011, 2024, doi: 10.52202/079017-3429.
- [20] Roboflow Universe, "Weld defects dataset," 2024. [Online]. Available: <https://universe.roboflow.com/yolov8-daldb/weld-defects-mlopr> [Accessed: Aug. 1, 2026]
- [21] L. Jia, Y. Zhu, and B. Li, "DPCNet: A dual path cross perception network for small object detection in UAV imagery," *PLOS ONE*, vol. 21, no. 3, Art no. e0344091, 2026.
- [22] B. İnan, İ. A. Özkan, and N. Doğan, "Deep learning-based skin disease detection: Comparative performance analysis of YOLOv8 and YOLOv11 models," *International Journal of Applied Methods in Electronics and Computers*, vol. 13, no. 4, pp. 121–130, 2025.
- [23] S. Haitao, Z. Li, X. Xiaoxue, S. Wang, and Y. Zhang, "Rapid identification of mangrove leaves based on improved YOLOv10 model," *Forests*, vol. 16, no. 7, Art no. 1068, 2025.
- [24] Y. Shen and C. Zhang, "Research on the application of YOLOv10n in helmet detection," *IAENG International Journal of Computer Science*, vol. 53, no. 2, pp. 721–729, 2026.
- [25] M. Hussain, "YOLOv5, YOLOv8 and YOLOv10: The go-to detectors for real-time vision," *arXiv preprint arXiv:2407.02988*, 2024.
- [26] L. Liu, R. Yuan, Y. Lv, W. Bo, X. Hu, and E. D. Gedikli, "An efficient high-resolution remote sensing object detection approach using a multilevel optimized RT-DETR model," *IEEE Sensors Journal*, vol. 25, no. 22, pp. 42306–42315, 2025.

- doi: 10.1109/JSEN.2025.3617485.
- [27] Q. Chen, X. Su, X. Zhang, J. Wang, J. Chen, Y. Shen, C. Han, Z. Chen, W. Xu, F. Li, and S. Zhang, "LW-DETR: A transformer replacement to YOLO for real-time detection," arXiv preprint arXiv:2406.03459, 2024.
- [28] Q. Wang, Y. Li, B. Luo, Z. Tu, and C. Li, "Dual-correlation hypergraph network for unaligned RGBT video object detection and a large-scale benchmark," arXiv preprint arXiv:2607.08191, 2026.
- [29] F. Liu, F. Wang, M. Liu, and D. Zhang, "RT-DETR-Crack: A tunnel lining crack detection method via multiscale feature enhancement and spatial attenuation mechanism," *IAENG International Journal of Applied Mathematics*, vol. 56, no. 4, pp. 1593–1606, 2026.