Training a YOLOv12 Model: Predicting Weld Structural Integrity

Aaron Celles
Mapua University – Makati
School of Information Technology (SOIT)
Antipolo, Philippines
akmcelles@mymail.mapua.edu.ph

Lee Leviste
Mapua University – Makati
School of Information Technology (SOIT)
Caloocan, Philippines
lraleviste@mymail.mapua.edu.ph

Kyle Lim
Mapua University – Makati
School of Information Technology (SOIT)
Makati, Philippines
khwlim@mymail.mapua.edu.ph

Enrique Santeco
Mapua University – Makati
School of Information Technology (SOIT)
Makati, Philippines
essanteco@mymail.mapua.edu.ph

***Abstract*— This study employs a deep learning model, YOLOv12, to automatically detect and analyze critical weld features from image data, specifically welding line, porosity, spatters, and cracks. Precision, recall, and mean Average Precision (mAP), among other important detection and segmentation metrics, were used to assess performance using a segmentation-based model. Due to their unique visual features, spatter and porosity detections produced the highest mAP scores, demonstrating the model's strong overall accuracy. Crack detection, on the other hand, performed worse, probably because of small training samples and subtle visual characteristics. With bounding box prediction exhibiting marginally better accuracy than mask segmentation, the model showed balanced detection and segmentation capabilities. Future work should concentrate on enhancing image quality to guarantee better feature learning, masking, and detection reliability to increase model robustness.**

Keywords— Weld feature prediction, deep learning, image segmentation, object detection, computer vision, defect classification, automated inspection, quality control, mAP.

I. Introduction

Ensuring the structural integrity of welded joints is a critical aspect of industrial manufacturing and construction, as even minor flaws in welds can compromise safety, durability, and overall performance. Traditional manual inspection methods are often time-consuming, subjective, and prone to human error, highlighting the need for automated solutions that can accurately classify weld quality. In this study, we focus on developing a deep learning–based approach using the YOLOv12 object detection framework to annotate and classify welds into three categories: (1) welding line, representing the seam continuity; (2) porosity, indicating trapped gas voids or bubbles within the weld; (3) spatters, referring to scattered molten droplets around the weld area; and (4) cracks, which are critical discontinuities that threaten structural integrity.

The implementation of automated weld classification has significant implications for productivity and cost-efficiency. Automated systems can provide rapid and consistent evaluations across large volumes of welds, enabling real-time monitoring and quality control within production pipelines. Furthermore, the ability of deep learning models to detect subtle variations that may not be visible to the human eye enhances the reliability of inspections compared to traditional methods.

The objective of this project is threefold: (1) explore a different annotation technique that best covers welded metal and its surrounding area, (2) build predictive models using YOLOv12 deep learning algorithm, and (3) to analyze and interpret detection outputs to evaluate weld integrity. The outcome of this study supports production optimization, promotes safer and more efficient welding practices, and contributes to the growing body of research in intelligent inspection systems within industrial and manufacturing contexts, thereby reducing the risks of structural failure and improving compliance with safety standards in engineering applications.

II. Related Work

The YOLOv10 deep learning model was employed in the study of Cengil [1] to automatically identify weld flaws. The model used bounding boxes to mark the locations of welds and classified them as good, bad, or defective. The Kaggle Welding Defect – Object Detection dataset, which already had YOLO-format box annotations, served as its training dataset. The model was able to identify the locations of flaws in each image thanks to these bounding boxes. The model's 0.939 precision and 0.91 recall demonstrated its ability to precisely and instantly identify weld flaws.

Truong et al., [2] additionally supports this method, which confirms that YOLOv10 was used for bounding box-based detection of weld defects. It describes how performance was evaluated by comparing expected and actual box positions using Intersection over Union (IoU). According to the study's findings, YOLOv10 offers a quick and accurate method for automatically identifying weld flaws.

III. DATASET OVERVIEW

The weld quality dataset used in this study comprises 17,063 images in total, generated through dataset augmentation (x3) from an initial collection of 7,109 images. The original dataset was divided into training, validation, and testing sets to support model development and performance evaluation. Augmentation techniques such as flipping, rotation, and contrast adjustment were applied to increase the dataset size while preserving the distribution of the classes. Annotation and segmentation were conducted through Roboflow, enabling consistent labeling across weld categories.

TABLE I. DATASET SPLIT SUMMARY

Set	Counts
Train	14,931
Valid	1,066
Test	1,066
Total	17,063

The dataset split follows a standard machine learning practice where most images are allocated to the training set, ensuring that the model has sufficient data to learn weld characteristics. A smaller validation set is used to fine-tune hyperparameters and monitor performance during training, while the test set provides an unbiased evaluation of the model's predictive capability. This structured division ensures both robustness and fairness in assessing model performance.

TABLE II. CLASS DISTRIBUTION

Class	Counts
Welding Line	12,153
Crack	3,084
Porosity	32,136
Spatters	47,600
Total	94,973

The dataset exhibits a noticeable variation in class distribution, with spatters representing the largest portion of annotated instances, followed by porosity, welding line, and crack. Despite this imbalance, each class provides distinct and visually meaningful examples: continuous seams for welding lines, small voids or bubbles for porosity, scattered molten droplets for spatters, and linear surface discontinuities for cracks. These annotations serve as ground truth labels for training and validating the YOLOv12 object detection model, ensuring accurate localization and recognition of key weld features critical to assessing weld integrity.

IV. DATASET DESCRIPTION

A. Preprocessing Overview

Before training the YOLOv12 model, the weld dataset underwent a series of preprocessing steps within Roboflow to enhance image quality and ensure consistency across all samples. These steps included Auto-Orient, Resize (640×640), and Auto-Adjust Contrast using Adaptive Equalization. Each transformation was applied with the objective of improving feature clarity, reducing noise, and optimizing images for deep learning workflows.

B. Preprocessing Approaches

Three types of preprocessing approaches were done:

- **Auto-Orient:** Used to standardize the orientation of all images. Since weld photographs may be captured from different angles or devices, inconsistent orientation could cause the model to misinterpret features. By enforcing a uniform orientation, this preprocessing step ensured that the positional alignment of welds remained consistent across the dataset.
- **Resize (640x640):** Selected because YOLOv12 requires input images of fixed dimensions. The choice of 640×640 offers a balance between computational efficiency and feature resolution: it is small enough to reduce training time and GPU memory consumption, yet sufficiently large to retain fine-grained details of weld surfaces such as cracks or spatters. This resizing step guarantees compatibility with YOLOv12's detection framework while minimizing distortion across samples.
- **Auto-Adjust Contrast, Adaptive Equalization:** In enhancing image visibility, especially in cases where welds appeared blurred or poorly lit. Adaptive equalization redistributes pixel intensity values, improving contrast in localized regions rather than applying a uniform adjustment. This makes subtle features—such as surface irregularities, spatters, or small cracks—more distinguishable for the model. By mitigating the impact of blurry or uneven lighting conditions, this step improves the consistency of input data and supports better feature extraction during training.

C. Data Augmentation

To improve the robustness and generalization of the YOLOv12 model, data augmentation was applied to the weld dataset through Roboflow, generating multiple variations of each image while preserving the essential weld characteristics.

$$\textit{Output per training: 3} \tag{1}$$

Generating three outputs per training example significantly increases the effective dataset size, ensuring the model encounters a wider range of samples during training.

$$\textit{Rotation } (-15^\circ \textit{ to } +15^\circ) \tag{2}$$

Slight rotations simulate variations in camera angle or positioning during data collection. This ensures the model learns to recognize weld patterns regardless of minor shifts in orientation.

Brightness (0% to +15%) (3)

Brightness adjustment compensates for lighting inconsistencies across different image capture environments. Increasing brightness variation enables the model to remain effective under different illumination conditions.

Exposure (-15% to +15%) (4)

Exposure changes simulate both underexposed and overexposed conditions, helping the model adapt to welds photographed in challenging lighting scenarios without losing performance.

Blur (up to 1.5px) (5)

A mild blur was applied to mimic real-world cases where images may appear slightly out of focus. This helps the model remain resilient to minor blurriness while still learning to identify key weld features such as cracks or spatters.

Noise (up to 0.1% of pixels) (6)

Adding random noise prevents the model from overfitting to overly clean samples and improves its ability to generalize in noisy or imperfect environments, such as factory settings where images may include background interference.

Shear (±15° Horizontal, ±15° Vertical) (6)

Applying shear transformations slightly skews the image along horizontal or vertical axes, simulating variations in camera angle or perspective. This helps the model become more robust to viewpoint distortions that may occur when objects are captured from different positions or angles in real-world conditions.

This augmentation strategy reduces overfitting and enhances the model's ability to generalize by simulating real-world variations in weld imaging conditions. The most impactful transformations for improving weld classification were rotation, contrast-related adjustments, and blur, as they closely replicate the environmental and visual inconsistencies found in practical inspection scenarios.

V. Experiments and Results

This section presents the empirical evaluation of the deep learning model described earlier. By applying a variety of metrics across bounding box and image segmentation, we aim to compare the effectiveness, generalizability, and interpretability of this modeling approach. The results are supported by quantitative analysis and visualizations to offer a comprehensive view of model behavior and image detection in identifying weld structural integrity.

A. Evaluation Metrics

1) To assess performance across weld status image detection, the following metrics were employed:

a) Box(P): Measures the precision in how many detected objects are correct.

b) Box(R): Shows the recall of how many real objects were found.

c) mAP50: Evaluates detection accuracy when the predicted box overlaps the true box by at least 50%.

d) mAP50-95 (Box): A stricter measure that averages the accuracy across intersection-over-union (IoU) thresholds from 0.5 to 0.95.

e) Mask(P): Measures the precision in how many predicted segmentation masks are correct.

f) Mask(R): Measures the recall of how many actual defect regions were correctly identified.

g) mAP50 (Mask): Evaluates the segmentation accuracy when predicted masks overlap the true region by at least 50%.

h) mAP50-95 (Mask): Averages segmentation accuracy across intersection-over-union (IoU) thresholds from 0.5 to 0.95.

B. Model Performance Comparison

Model performance was quantitatively evaluated for the deep learning model (weld), highlighting the strengths and trade-offs among the two approaches: Box (Bounding Box), and Mask (Image Segmentation).

Box Results:

TABLE III. Bounding box evaluation

Class	Box(P)	Box(R)	mAP50(Box)	mAP50-95(Box)
All	0.8209	0.7520	0.8180	0.5756

Mask Results:

TABLE IV. Image segmentation evaluation

Class	Mask(P)	Mask(R)	mAP50(Mask)	mAP50-95(Mask)
All	0.7968	0.6924	0.7508	0.4218

The results indicate that the YOLOv12 model demonstrates strong overall detection and segmentation performance across all weld feature classes. With bounding box precision (0.8209) and recall (0.7520) yielding an mAP50 of 0.8180, the model effectively identifies and localizes welding lines, porosity regions, spatters, and cracks. The slightly lower mAP50-95 score (0.5756) reflects a modest drop in accuracy when stricter Intersection over Union (IoU) thresholds are applied, which is typical for fine-grained surface features such as cracks and porosity. In terms of segmentation, the model achieves a mask precision of 0.7968 and mAP50 of 0.7508, demonstrating its capability to outline weld features with reasonable accuracy. However, segmentation performance is somewhat lower than bounding box prediction, suggesting that delineating irregular and diffuse regions—such as scattered spatters or porous textures—poses a greater challenge compared to detecting their general location.

C. *Visual Results*

a) BoxF1 Curve: The model's ability to balance precision and recall at various confidence levels is demonstrated by this plot, which displays the F1 score against the confidence threshold. The model's best trade-off between accurately identifying and missing weld defects is represented by the peak point, which also serves as the ideal confidence threshold.

b) BoxP Curve: Precision versus confidence is shown in this graph. Precision usually increases as the confidence threshold rises because the model becomes more selective in its predictions. It may, however, detect fewer total objects, which would reduce recall.

c) BoxR Curve: This figure illustrates the relationship between recall and confidence, showing that recall falls as confidence rises. The model only makes predictions when it is more certain at higher thresholds, which results in more missed true instances but fewer false detections.

d) BoxPR Curve: The relationship between precision and recall is depicted by the Precision–Recall curve. Better overall model performance is indicated by a larger area under the curve (AUC), which demonstrates that the model can maintain high precision without significantly compromising recall.

e) Annotation and Distribution Analysis: The figure presents a comprehensive visualization of the dataset's annotation characteristics, providing insights into class distribution, object positioning, and size variation. The bar chart (top-left) illustrates the number of labeled instances for each class—crack, porosity, spatters, and welding line—revealing an imbalance where spatters and porosity appear most frequently, while cracks are relatively scarce. The overlay of bounding boxes (top-right) shows the overall spatial spread and size of annotations, indicating that most weld features are concentrated around the central region of the images. The heatmap (bottom-left) depicts the spatial distribution of object centers, with darker areas suggesting higher annotation density, again emphasizing a concentration near the image center where weld seams typically occur. Meanwhile, the width–height plot (bottom-right) displays the relative dimensions of bounding boxes, offering insight into the scale variability of different weld features. Overall, this visualization aids in understanding the dataset composition and spatial patterns, which is essential for optimizing YOLOv12's anchor box configuration and improving detection accuracy across diverse weld characteristics.

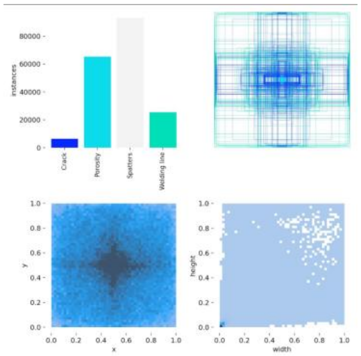
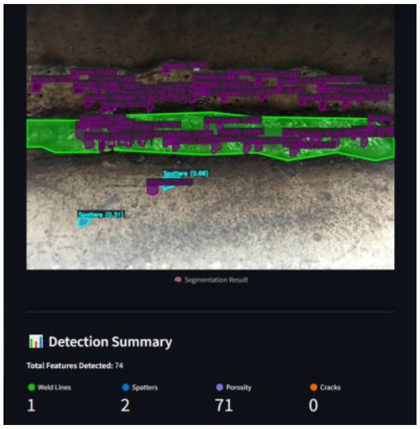


Fig. 1. Class and Spatial Distribution

VI. Deployment

The trained weld detection model was implemented/deployed using Streamlit. With modifications that allows it to predict multiple instances of a single presented image.



📊 Detection Summary

Total Features Detected: 74

● Weld Lines	● Spatters	● Porosity	● Cracks
1	2	71	0

Fig. 2. Streamlit App Detection

VII. Discussion

This section interprets the results in greater depth, addressing model performance, dataset limitations, and the balance between different evaluation metrics in weld defect prediction.

- Best-Performing Classes:

  The model achieved its highest precision, recall, and mAP values for spatters and porosity, indicating strong capability in identifying these visually distinct and frequently occurring features. In contrast, crack detection showed lower precision and recall, likely due to the subtle and fine-grained nature of cracks and their limited representation in the dataset. Overall, the YOLOv12 model performed slightly better in bounding box detection than in mask segmentation, suggesting that it is more effective at locating weld features than precisely outlining their boundaries.

- Precision vs. Recall Trade-Off:

  Despite variation across classes, the model maintained a balanced trade-off between precision and recall. High precision in detecting porosity and spatters indicates confident and accurate predictions, while the slightly higher recall for welding line detection shows the model's ability to capture most instances, even at the cost of a few false positives. In practice, higher precision is preferred for automated quality assurance to minimize false alarms, whereas higher recall is more suitable for safety-critical inspections where missing a defect is unacceptable.

VIII. Recommendations

This study structured the train and test dataset mainly with image segmentation over bounding boxes. In this technical approach tested, it was found that defect class performed the least among other classes with abundant data instances in the sourced dataset from Kaggle.

To further improve predictive accuracy and model generalization, the following recommendations are proposed:

- Dataset Expansion:

  To address class imbalance and aid the model in learning the distinctive characteristics of different defect types, future work should concentrate on including more images, especially for the defect class. Improving generalization and decreasing bias toward more prevalent classes, like good and bad welds, can be achieved by increasing the number of training samples.

- Quality Improvement:

  Furthermore, it is crucial to make sure that the images are sharp, well-lit, and properly focused because dark or blurry images can make it more difficult for the model to recognize and segment weld areas. The robustness of the model and detection performance will be further enhanced by applying consistent imaging conditions and preprocessing for clarity.

References

[1] E. Cengil, "Weld defect detection with YOLOv10," Bitlis Eren University, Department of Computer Engineering, Bitlis, Türkiye – 13100. Malatya Turgut Ozal University Journal of Engineering and Natural Sciences, vol. 5, no. 2, pp. 77–81, 2024. doi: 10.46572/naturengs.1592956.

[2] V. D. Truong, Y. Wang, C. Won, and J. Yoon, "A deep learning-based machine vision system for online monitoring and quality evaluation during multi-layer multi-pass welding," Sensors (Basel), vol. 25, no. 16, p. 4997, Aug. 2025. doi: 10.3390/s25164997.