Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Cross-Modal Detection-Segmentation Framework for Video-Based Process Monitoring in Wire Arc Additive Manufacturing

Sayak Halder
Indian Institute of Technology Kharagpur
Shahana Bano
s.bano@rgu.ac.uk

Robert Gordon University
Anil Prathuru
Robert Gordon University
Somasundar Kannan
Robert Gordon University
Shiva Sekar
Indian Institute of Technology Jammu
Deepika Nikam
Ulster University, Derry-Londonderry Campus
Sagar Nikam
Ulster University, Derry-Londonderry Campus

Research Article

Keywords: Wire Arc Additive Manufacturing, WAAM, Deep Learning, Computer Vision, Melt Pool Monitoring

Posted Date: September 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10935215/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Cross-Modal Detection-Segmentation Framework for Video-Based Process Monitoring in Wire Arc Additive Manufacturing

Sayak Halder[1], Shahana Bano[2*], Anil Prathuru[2], Somasundar Kannan[2], Shiva Sekar[3], Deepika Nikam[4], Sagar Nikam[4]

[1]Department of Industrial and Systems Engineering, Indian Institute of Technology Kharagpur, Kharagpur, India.
[2*]School of Computing, Engineering and Technology, Robert Gordon University, Aberdeen, United Kingdom.
[3]Laboratory for Advanced Manufacturing and Processing, Indian Institute of Technology Jammu, Jammu, India.
[4]Intelligent Systems Research Centre, Ulster University, Derry/Londonderry, United Kingdom.

*Corresponding author(s). E-mail(s): s.bano@rgu.ac.uk;
Contributing authors: sayakhalder572@gmail.com; a.prathuru@rgu.ac.uk; s.kannan1@rgu.ac.uk; shiva.sekar@iitjammu.ac.in; d.nikam@ulster.ac.uk; s.nikam@ulster.ac.uk;

Abstract

Wire Arc Additive Manufacturing (WAAM) process stability depends critically on monitoring the dynamic interplay between the deposited weld bead and the active molten pool. While deep learning has enabled significant progress in melt-pool analysis, robust real-time detection and pixel-accurate segmentation under arc glare, spatter, and thermal bloom remain unresolved. This study presents a comprehensive computer-vision pipeline for Tungsten Inert Gas (TIG) welding inspection that integrates lightweight object detection with promptable foundation-model segmentation. We systematically benchmark five modern detectors—YOLOv8n, YOLOv26n (RGB and grayscale), LW-DETR Tiny, and DINO-DETR—under a rigorous 5-fold stratified cross-validation protocol on a curated dataset of 335 industrial welding frames. YOLOv8n emerges as the optimal backbone, achieving a mean Average Precision of 0.8856 at IoU=0.50 and

an inference throughput of 140.7 FPS, substantially outperforming transformer-based alternatives on this small-scale industrial dataset. To circumvent costly pixel-level annotation, we propose a YOLOv8n-guided SAM2 framework that propagates temporally coherent segmentation masks from sparse bounding-box prompts. Qualitative evaluation demonstrates reliable tracking of the global weld region, yet SAM2 consistently merges the weld bead and molten pool due to the continuous physical boundaries and extreme luminance transitions inherent to arc welding—highlighting a fundamental domain gap in foundation models trained on discrete natural objects. Consequently, the extracted geometric descriptors are treated as approximate nominal features feeding a future Detection-Prompted Quality Network (DPQN) for weakly supervised video-level quality assessment. These findings establish a reproducible detection baseline and delineate the practical limits of zero-shot segmentation in welding environments, guiding future domain-adaptive architectures for autonomous WAAM quality assurance.

Keywords: Wire Arc Additive Manufacturing, WAAM, Deep Learning, Computer Vision, Melt Pool Monitoring

1 Introduction

Wire Arc Additive Manufacturing (WAAM), a form of Directed Energy Deposition (DED), has gained significant attention due to its capability to fabricate large-scale metallic components with high deposition rates and cost efficiency. Despite its advantages, challenges related to process stability, defect formation, and quality assurance persist, largely due to the complex and highly dynamic behaviour of the melt pool. Recent developments in artificial intelligence, particularly deep learning and computer vision, have enabled advanced monitoring and control strategies aimed at improving WAAM process reliability.

Vision-based monitoring has emerged as a promising approach for analysing melt pool characteristics and ensuring process quality. Unlike traditional sensor-based methods relying on arc current, voltage, or acoustic signals, vision systems provide direct and rich information about melt pool geometry and process stability. In this context, the WAAM-ViD framework introduces a significant advancement by proposing an angle-invariant vision-based monitoring pipeline capable of analysing melt pool images captured from multiple camera perspectives [1]. This study addresses a key limitation in existing works, where monitoring systems are often constrained to fixed camera orientations and specialised setups, thereby limiting generalisability across different industrial configurations. Furthermore, the introduction of the WAAM-ViD dataset, comprising multi-angle videos with detailed annotations and metadata, contributes to overcoming the scarcity of benchmark datasets in this domain [1].

The application of deep learning techniques, particularly semantic segmentation models such as DeepLabv3, has demonstrated high effectiveness in extracting melt pool features. In the WAAM-ViD framework, the segmentation model achieved a Dice Similarity Coefficient of 95.90%, highlighting its robustness under varying imaging

conditions and camera configurations [1]. Beyond segmentation, the study integrates a regression-based model, WAAM-ViDNet, to predict melt pool width using both image features and calibration data. This multimodal approach enables quantitative analysis of melt pool geometry, achieving a prediction accuracy of 88.71%, and marks an important transition from qualitative defect detection to geometry-aware process monitoring [1].

In addition to segmentation-based approaches, real-time anomaly detection using object detection algorithms has been widely explored. A notable contribution in this area is the use of the You Only Look Once (YOLO) algorithm for real-time monitoring of WAAM processes [2]. The YOLOv8-based framework demonstrates exceptional performance in detecting irregularities such as "regular" and "irregular" bead formations, achieving a mean average precision (mAP) of up to 99%. The key advantage of this approach lies in its high processing speed, with inference times as low as 15 milliseconds per frame, enabling real-time detection and potential process control [2]. Such capabilities are critical for industrial adoption, where timely identification of anomalies can significantly reduce material wastage and post-processing costs.

The YOLO-based framework also highlights the importance of integrating sensing systems, data processing, and decision-making units into a unified architecture. By incorporating multiple data sources, including melt pool images and process parameters, the system supports real-time feedback mechanisms for adaptive control [2]. However, challenges such as noise, spatter interference, and motion blur remain significant obstacles, affecting the robustness of vision-based models. These limitations indicate the need for enhanced data preprocessing techniques and more diverse datasets to improve model generalisation.

Another important line of research focuses on the use of convolutional neural networks (CNNs) for classification-based anomaly detection. The study on deep learning for anomaly detection in WAAM employs a CNN architecture inspired by ResNet-50 and VGG-16 to classify deposited beads into "regular" and "irregular" categories [3]. The model achieves satisfactory classification accuracy of approximately 90%, demonstrating the feasibility of using lightweight deep learning models even with limited datasets. This approach is particularly advantageous in scenarios where data acquisition is expensive and time-consuming, as is often the case in WAAM experiments [3].

Data augmentation techniques, including rotation and flipping, are employed to address data scarcity and improve model robustness [3]. While CNN-based classification models are effective for identifying the presence of defects, they do not provide detailed spatial information about defect location or geometry. Consequently, their applicability in advanced process monitoring systems is limited when compared to segmentation or object detection approaches.

A comparative analysis of the three approaches reveals a clear progression in the application of deep learning techniques in WAAM monitoring. CNN-based classification methods provide a foundation for defect detection but lack spatial and geometric detail. YOLO-based object detection approaches enable real-time localisation and classification of anomalies, offering improved capabilities for in-situ monitoring. In

contrast, segmentation and regression-based frameworks, such as WAAM-ViD, provide comprehensive insights into melt pool geometry and dimensional characteristics, supporting more advanced applications such as process optimisation and closed-loop control.

Foundational segmentation models are seen as a breakthrough offering high flexibility and performance. Such models are pretrained on multiple publicly available image datasets. One such model is Segment anything model (SAM) has recently gained much attention in object detection and tracking. Fine tuning SAM has shown significant promise with medical and industrial computed tomography images. However, it has known limitations in segmenting multiple classes and when small flaws and anomalies seen in material phase and defect related images. To address this, representative image generation using tools such as Generative Adversarial Networks (GANs) has been explored to enhance the training and segmentation performance of SAM [6]. More recently, SAM2 has been introduced with capabilities of spatio-temporal segmentation and object tracking across multiple frames of a video [7].

Despite these advancements, several research challenges remain. The limited availability of large and diverse datasets continues to hinder model generalisation across different materials, process parameters, and environmental conditions. Additionally, the robustness of vision-based systems is affected by practical issues such as arc glare, spatter, and occlusions. Computational efficiency is another critical factor, particularly for real-time industrial deployment, where resource constraints must be considered. Furthermore, while many studies focus on monitoring and detection, the integration of these systems into fully autonomous control frameworks remains an open research area.

Deep learning and computer vision have significantly enhanced the capabilities of WAAM process monitoring. The development of robust, generalisable, and computationally efficient models, along with the creation of benchmark datasets, will be crucial for advancing towards fully autonomous and intelligent WAAM systems. Future research should emphasise the integration of monitoring systems with real-time control mechanisms to achieve adaptive and self-optimising manufacturing processes.

1.1 Contributions:

The main contributions of this work are summarised as follows:

1. An automated computer vision pipeline for TIG welding inspection is developed to detect weld beads and molten pools from industrial videos. The workflow integrates data acquisition, frame extraction, annotation, model training, and evaluation, enabling intelligent welding process monitoring.
2. A curated and manually annotated TIG welding dataset is constructed from industrial videos under both good and defective conditions. Redundant frames are removed, and the remaining images are annotated with bounding boxes, providing a reliable benchmark for object detection in welding environments.
3. A comprehensive benchmarking of modern object detection models is performed. Five detectors YOLOv8n, YOLOv26n (RGB), YOLOv26n (Grayscale), LW-DETR

Tiny, and DINO-DETR are evaluated using 5-fold cross-validation, enabling systematic comparison of convolutional and transformer-based approaches.
4. The impact of grayscale image representation is investigated to evaluate whether luminance-based input improves detection performance while reducing complexity, providing insights into the role of color in welding image analysis.
5. Transformer-based detectors (LW-DETR Tiny and DINO-DETR) are evaluated on a limited industrial dataset, highlighting their performance limitations and data dependency in welding inspection tasks.
6. An efficient detection backbone is identified by jointly analyzing accuracy, localization, computational cost, and inference speed. YOLOv8n demonstrates a favorable balance and is selected for future real-time welding inspection and quality assessment.

2 Methodology

2.1 Experimental Setup

Figure. 1 shows the setup used for the wire arc additive manufacturing (WAAM) trials. The deposition path was controlled using a robotic arm into which the deposition setup was integrated. A Fronius TPS 400i welding unit, which operates under Cold Metal Transfer (CMT) process modes in a 2-step control configuration was used to supply the energy required for material deposition. The feedstock material used for deposition is ER307 stainless steel wire, while the substrate material is mild steel. A gas mixture composed of 93% argon, 5% carbon dioxide, and 2% oxygen, ensured arc stability and protection of the molten material from atmospheric contamination as well as providing a shielding environment. WAAM process operating conditions, such as welding current, arc voltage, wire feed rate, and travel speed, are controlled through the integrated process control system. The material deposition process was monitored using a high dynamic range (HDR) monochrome camera pointed towards the deposition zone and was operated at 30 frames per second (fps). A frame size was of 1280 x 1024 pixels and a pixel size of 0.0068 mm were used for the recordings. Critical process phenomena such as the arc shape and morphology, melt pool temporal evolution and shape, spatter formation, droplet transfer, bead solidification and growth were captured under various WAAM process operational conditions. Process behaviour and deposition consistency were evaluated from individual frames. Figure. 1 demonstrates an image extracted from the captured video of the various process phenomena under selected deposition conditions. The experimental matrix covers process parameter combinations (current, voltage) chosen to induce both stable and unstable bead deposition. In some cases, inadequate shielding led to degraded surface qualitydue to formation of impurities as well as contaminations. The captured images also included intermittent material transfer yielding non-uniform bead geometry and deposition discontinuities. In some cases, spatter particles due to excessive molten material expulsion were noticed due to unstable arc conditions.

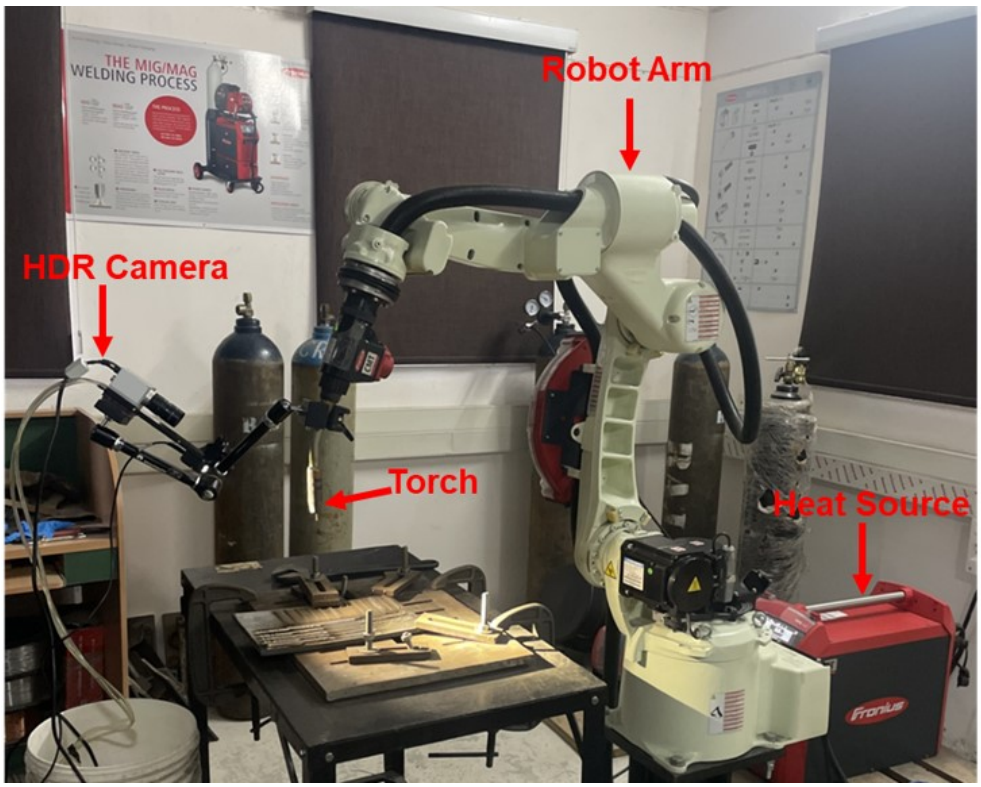


Fig. 1 Experimental Setup

2.2 Data Acquisition

The dataset was constructed from grayscale videos of the Tungsten Inert Gas (TIG) welding process captured using a fixed industrial vision system positioned above the welding region. The camera remained stationary throughout data acquisition, providing a consistent top-down view of the weld bead and molten pool. The videos were recorded at an approximate frame rate of 25 FPS, with a spatial resolution of 1280 $\times$ 1024 pixels and a bitrate of approximately 3002 kbps. A total of 11 welding videos were collected, comprising seven good-quality welds and four defective welds. Videos exhibiting transitions between good and defective welding conditions were intentionally excluded to ensure that each recording represented a single, consistent welding condition.

The acquired dataset is denoted as

$$D = \{V_i\}_{i=1}^{11}, \tag{1}$$

where each video sequence V_i belongs to either the good-weld class (C_g) or the defective-weld class (C_d). The dataset contains

$$|C_g| = 7, \qquad |C_d| = 4. \tag{2}$$

Each video sequence is represented as

$$V_i = \{f_t\}_{t=1}^{T_i}, \tag{3}$$

where $f_t \in \mathbb{R}^{1024\times1280}$ denotes a grayscale frame captured at time instant t, and T_i is the total number of frames in video V_i.

2.3 Frame Extraction

Individual image frames were extracted from the recorded videos using OpenCV. Initially, 552 frames were obtained. As consecutive frames in welding videos exhibit high temporal correlation, many extracted images contained little or no observable movement of the welding torch or process state. Such frames provide limited discriminative information while unnecessarily increasing annotation effort and the risk of overfitting. Furthermore, a subset of frames was affected by severe arc-flash glare and saturation, resulting in limited visibility of the weld bead and molten pool and reducing annotation reliability. Therefore, temporally redundant and flash-affected frames were manually removed, resulting in a final dataset of 335 representative images.

The extracted frame dataset is defined as

$$F = \{f_i\}_{i=1}^{552}. \tag{4}$$

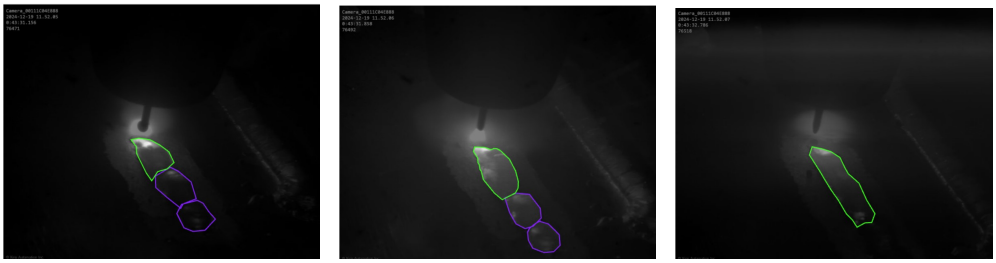
Following manual removal of temporally redundant and severely flash-affected frames, the representative frame subset is

$$F_r \subset F, \qquad |F_r| = 335. \tag{5}$$

The retained images preserve substantial variation in weld geometry, molten pool appearance, illumination, and process dynamics, while reducing redundancy within the dataset.

2.4 Annotation Procedure

All images were manually annotated using the Roboflow annotation platform by a single annotator as shown in Fig. 2. Two object categories were defined **Bead**: the deposited weld bead formed during the welding process. **Molten Pool**: the active molten metal region immediately beneath the welding arc.



(a) Initial bead annotation (b) Molten pool annotation (c) Glare and occlusions

Fig. 2 Manual annotation of weld bead and molten pool using Roboflow. Axis-aligned bounding boxes are used to label both classes under varying visual conditions, including arc glare and partial occlusion.

Axis-aligned bounding boxes were manually annotated to tightly enclose each visible object while ensuring consistency across the dataset. Particular care was taken when labelling frames affected by strong arc glare, specular reflections, and partial occlusions. Two distinct annotation classes were defined: weld bead and molten pool. The final dataset comprises 196 weld bead instances and 285 molten pool instances, yielding a total of 481 annotated objects. All annotations were exported in YOLO format for detector training.

The class distribution is given by

$$N_{\text{bead}} = 196, \qquad N_{\text{pool}} = 285, \tag{6}$$

resulting in a total of

$$N_{\text{obj}} = N_{\text{bead}} + N_{\text{pool}} = 481 \tag{7}$$

annotated objects.

Polygon Annotation was converted to Bounding Boxes although the dataset was annotated at the polygon level in Roboflow, providing precise instance segmentation masks for both weld bead and molten pool classes, all five baseline detectors evaluated in this work operate on axis-aligned bounding boxes rather than pixel-level masks. The conversion from polygon to bounding box was performed automatically within Roboflow during dataset export by computing the tightest axis-aligned enclosing rectangle for each polygon annotation. Specifically, the bounding box is defined by the minimum and maximum x and y coordinates across all polygon vertices, producing a rectangle that fully contains the annotated region without clipping.

For a polygon annotation represented by the vertex set

$$P = \{(x_i, y_i)\}_{i=1}^{n}, \tag{8}$$

the corresponding axis-aligned bounding box is defined by

$$x_{\min} = \min_i(x_i), \qquad x_{\max} = \max_i(x_i), \tag{9}$$

$$y_{\min} = \min_i(y_i), \qquad y_{\max} = \max_i(y_i). \tag{10}$$

The enclosing bounding box is therefore given by

$$B = (x_{\min}, y_{\min}, x_{\max}, y_{\max}), \tag{11}$$

with width and height

$$w = x_{\max} - x_{\min}, \qquad h = y_{\max} - y_{\min}. \tag{12}$$

This conversion is lossless in the sense that no spatial extent of the original annotation is discarded; the bounding box is a strict superset of the polygon mask. The YOLO-format normalised centre coordinates and dimensions recorded in each label file correspond exactly to this enclosing rectangle. The resulting bounding boxes were verified against the source images to ensure that no box extended beyond image boundaries and that aspect ratios remained consistent with the expected elongated bead geometry and approximately circular molten pool geometry across the dataset.

The decision to use bounding boxes rather than polygon supervision is deliberate and domain-motivated rather than driven by convenience. The primary detection objective in welding monitoring is spatial localisation, i.e., reliably identifying the positions of the bead and molten pool in each frame so that downstream components can extract spatially aligned crops for quality assessment. Pixel-precise boundary delineation, while geometrically informative, is not required for this objective. Furthermore, instance segmentation models incur significantly higher parameter counts and inference costs compared to detection models with comparable backbones, which conflicts with the requirement for near real-time deployment on edge hardware proximal to the welding station.

In addition, the polygon annotations exhibit inter-frame inconsistency in this dataset. The weld bead boundary in active welding frames is often partially obscured by arc glow, spatter, and thermal bloom, resulting in annotations that follow visible luminance edges rather than the underlying geometric boundary. Using such noisy polygon masks as supervision for segmentation would introduce label noise, whereas bounding box regression is inherently more robust, as the enclosing rectangle is less sensitive to individual vertex placement errors. The polygon annotations are therefore retained as a higher-fidelity supervision resource for future geometric feature extraction work described in Section 5, where the performance of segmentation models trained on full polygon labels can be compared against the SAM2 zero-shot approach, which was found to be insufficient in preliminary experiments.

2.5 Image Preprocessing and Augmentation

Prior to model training, all images were resized to 640 X 640 pixels to ensure a consistent input resolution across all evaluated detection architectures.

The original image dataset is represented as

$$I = \{I_i\}_{i=1}^{N}, \tag{13}$$

where I_i denotes an individual welding image. To ensure a consistent input size across all detectors, each image was resized according to

$$I_i' = \mathcal{R}(I_i, 640, 640), \tag{14}$$

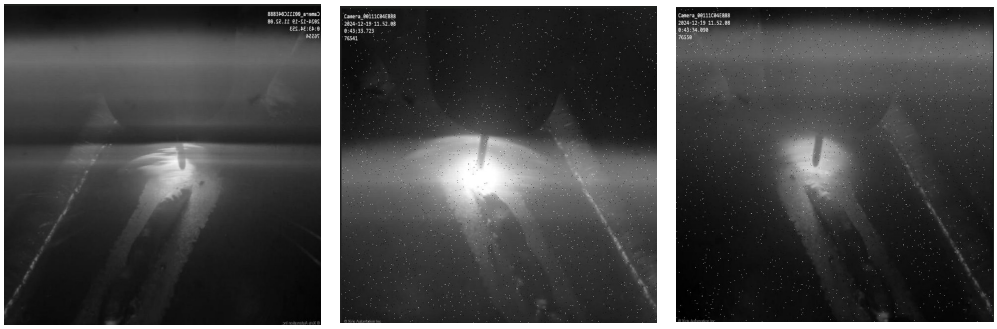
where $\mathcal{R}(\cdot)$ denotes the image resizing operation and $I_i' \in \mathbb{R}^{640\times640\times c}$ is the resulting image, with $c = 3$ for RGB images and $c = 1$ for grayscale images.

Data augmentation was performed using the Roboflow platform to improve model robustness to variations commonly encountered in welding conditions. The augmentation pipeline included Horizontal image flipping, Gaussian blur, Random image noise, Saturation adjustment and Image resizing to 640 X 640 pixels as shown in Fig 3. For the grayscale experiments, RGB images were converted to a single-channel luminance representation prior to training, whereas the RGB experiments retained the original three-channel inputs.

For the grayscale experiments, RGB images were converted to a single-channel luminance representation using

$$Y = 0.299R + 0.587G + 0.114B, \tag{15}$$

where R, G, and B denote the red, green, and blue intensity channels, respectively. No additional image enhancement or histogram equalization techniques were applied.



(a) Horizontal flipping (b) Gaussian blur (c) Random noise injection

Fig. 3 Data augmentation pipeline applied during training

2.6 Cross-Validation Protocol

To obtain statistically reliable performance estimates, a 5-fold stratified cross-validation protocol was employed. Frame-level stratification ensured that each fold contained a representative distribution of both target object classes. For each fold, four subsets were used for training and one subset was used for validation, with the reported performance corresponding to the mean and standard deviation across all five folds.

Although frame-level cross-validation provides an effective benchmark for comparing detector architectures, neighbouring frames extracted from the same video may

remain temporally correlated. Consequently, future work will investigate video-level cross-validation to provide a more conservative assessment of model generalization.

The annotated dataset is defined as

$$D = \{(x_i, y_i)\}_{i=1}^{N}, \tag{16}$$

where x_i denotes an input welding frame and y_i denotes the corresponding object annotations. The dataset is partitioned into $K = 5$ mutually exclusive folds,

$$D = \bigcup_{k=1}^{K} D_k, \qquad D_i \cap D_j = \emptyset,\ i \neq j, \tag{17}$$

where stratified sampling is employed to preserve the class distribution within each fold. For the kth cross-validation iteration, the training and validation sets are given by

$$D_{\text{train}}^{(k)} = D \setminus D_k, \tag{18}$$

and

$$D_{\text{val}}^{(k)} = D_k, \tag{19}$$

respectively. The evaluation metric obtained for fold k is represented by $M^{(k)}$. Overall performance is reported as the mean

$$\mu_M = \frac{1}{K} \sum_{k=1}^{K} M^{(k)}, \tag{20}$$

and standard deviation

$$\sigma_M = \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} \left(M^{(k)} - \mu_M\right)^2}. \tag{21}$$

2.7 Baseline Model Evaluation

As part of this study, a rigorous performance baseline for weld bead and molten pool detection was established through the systematic evaluation of five object detection architectures under a unified experimental protocol: **YOLOv8n, YOLOv26n (RGB), YOLOv26n (grayscale), LW-DETR Tiny, and DINO-DETR**. These models span a diverse range of detection paradigms, including lightweight single-stage convolutional detectors and transformer-based approaches. Each model was trained to simultaneously localise two target classes, namely deposited weld beads and active molten pools, which serve as key indicators of process state in TIG welding.

The detection task is formulated as a two-class object detection problem,

$$C = \{c_1, c_2\} = \{\text{weld bead}, \text{molten pool}\}, \tag{22}$$

where each image x_i contains a set of annotated objects

$$Y_i = \{(b_j, c_j)\}_{j=1}^{N_i}, \tag{23}$$

with b_j denoting the bounding-box coordinates and $c_j \in C$ the associated class label. The objective of each detector is to predict

$$\hat{Y}_i = \{(\hat{b}_j, \hat{c}_j, s_j)\}_{j=1}^{M_i}, \tag{24}$$

where $\hat{b}_j$, $\hat{c}_j$, and s_j represent the predicted bounding box, class label, and confidence score, respectively.

The five baseline architectures were selected to span the primary design axes relevant to this detection task, enabling a principled comparison rather than an arbitrary survey. **YOLOv8n** was selected as the established lightweight baseline, representing the current standard for real-time single-stage detection on resource-constrained hardware. It has demonstrated strong performance across industrial inspection tasks in recent literature and serves as the direct predecessor to the detector adopted in the proposed DPQN pipeline, making its standalone performance the most practically relevant reference.

YOLOv26n was evaluated in both its standard **RGB configuration and a grayscale input** variant to investigate a domain specific hypothesis. Welding imagery is dominated by luminance contrast rather than chromatic information, as arc emission and thermal radiation produce high-intensity, near monochromatic regions against a dark background. If the grayscale variant performs comparably to or better than the RGB model, this would justify a single channel input design in deployment scenarios with camera or bandwidth constraints, and further indicate that colour information provides negligible discriminative value for this task. This finding would directly inform sensor selection in industrial welding monitoring systems. The grayscale conversion was implemented by collapsing the pretrained first convolutional layer's RGB weights using ITU-R BT.601 luminance weighting, thereby preserving the relative importance of each channel learned during ImageNet pretraining rather than discarding it through random reinitialisation.

DINO-DETR and LW-DETR Tiny were included to evaluate the extent to which the set prediction formulation and global self-attention mechanisms of transformer-based detectors provide advantages over convolutional single-stage models for this task. Transformer-based detectors have demonstrated strong performance in dense and occluded object scenarios in general-domain benchmarks. In welding frames, the partial spatial overlap between the molten pool and the leading edge of the weld bead represents a comparable configuration, where global attention may outperform local anchor-based reasoning.

LW-DETR Tiny was selected in preference to larger DETR variants due to its parameter count being comparable to YOLOv8n, enabling a fair assessment of architectural differences without confounding effects from model capacity. The inclusion of DINO-DETR (see Fig. 4) further enables evaluation of a custom architecture built on a large pretrained vision foundation model, assessing whether the semantic representations encoded in the DINOv3 ViT-S+16 backbone transfer effectively to the

specialised domain of weld imagery. This is of broader relevance given the increasing interest in applying foundation models to industrial inspection tasks with limited annotated data.

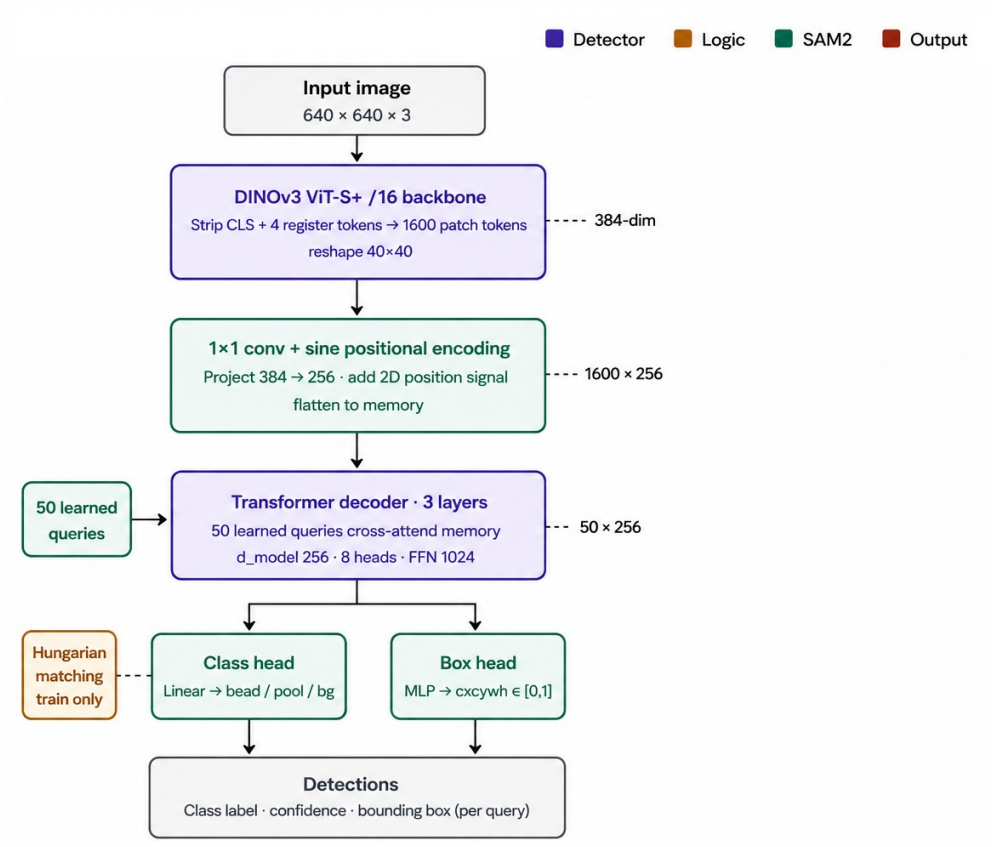


Fig. 4 DINO-DETR Architecture

2.8 Evaluation Metrics

Detection performance was evaluated using mean Average Precision at an intersection-over-union threshold of 0.50 (mAP50) and across the COCO IoU range of 0.50–0.95 (mAP50-95). The former serves as the primary metric for assessing detection capability, while the latter evaluates localisation quality under progressively stricter overlap requirements. Precision, recall, and F1-score are additionally reported to characterise the precision–recall trade-off of each model.

For a predicted bounding box B_p and ground-truth bounding box B_g, the Intersection-over-Union (IoU) is defined as

$$\text{IoU} = \frac{|B_p \cap B_g|}{|B_p \cup B_g|}, \tag{25}$$

where $|\cdot|$ denotes the area of a region. Detection precision and recall are computed as

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{26}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{27}$$

where TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively. The F1-score, representing the harmonic mean of precision and recall, is given by

$$F_1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{28}$$

Average Precision (AP) is computed as the area under the precision-recall curve,

$$AP = \int_0^1 P(R)\, dR, \tag{29}$$

where $P(R)$ denotes precision as a function of recall. The mean Average Precision (mAP) is obtained by averaging AP across all object classes,

$$mAP = \frac{1}{C} \sum_{c=1}^{C} AP_c, \tag{30}$$

where C denotes the number of classes.

For $mAP50$, a detection is considered correct when IoU ≥ 0.50. The $mAP50$–95 metric averages AP over multiple IoU thresholds ranging from 0.50 to 0.95 with an increment of 0.05,

$$mAP_{50-95} = \frac{1}{10} \sum_{t=0}^{9} AP_{(0.50+0.05t)}. \tag{31}$$

Inference throughput is measured in frames per second (FPS),

$$FPS = \frac{N_f}{T}, \tag{32}$$

where N_f denotes the number of processed frames and T is the total inference time in seconds.

To assess deployment suitability for real-time welding monitoring, computational efficiency was also evaluated for the YOLO-family detectors using the number of

trainable parameters, computational complexity in GFLOPs, and inference throughput measured in frames per second (FPS). FPS measurements were obtained on an NVIDIA Tesla T4 GPU using a batch size of one and an input resolution of 640 × 640 pixels.

The WAAM dataset comprises 11 high-speed arc welding videos collected under varying process conditions. In total, 335 welding frames were annotated, capturing variability in both process parameters and illumination levels. All experiments were conducted using a 5-fold cross-validation protocol with frame-level stratified splitting to ensure a representative distribution of both target classes within each fold. Input images were uniformly resized to 640 × 640 pixels. YOLO-based detectors were trained for 100 epochs, whereas LW-DETR Tiny and DINO-DETR were trained for 150 epochs to account for the slower convergence characteristics typically associated with transformer-based architectures. All experiments were performed on a single NVIDIA Tesla T4 GPU (15 GB VRAM) within the Kaggle computing environment. Detection performance is reported as mean ± standard deviation across the five folds.

Table 1 Detection performance on the welding dataset. Values are reported as mean ± standard deviation across five folds. Best values are highlighted in bold

Model	mAP50	mAP50-95	Precision	Recall	F1
YOLOv8n	**0.8856 ± 0.0362**	0.5424 ± 0.0529	**0.8796 ± 0.0579**	**0.8647 ± 0.0434**	**0.8699 ± 0.0289**
YOLOv26n-RGB	0.8581 ± 0.0355	**0.5520 ± 0.0260**	0.7804 ± 0.0647	0.8064 ± 0.0491	0.7903 ± 0.0331
YOLOv26n-Grayscale	0.8263 ± 0.0646	0.5212 ± 0.0439	0.7330 ± 0.0871	0.7724 ± 0.0852	0.7498 ± 0.0741
LW-DETR Tiny	0.4407 ± 0.0327	0.3011 ± 0.0295	0.4835 ± 0.0542	0.4536 ± 0.0402	0.4640 ± 0.0181
DINO-DETR-our model	0.2916 ± 0.0229	0.1532 ± 0.0116	0.3269 ± 0.0616	0.3713 ± 0.0397	0.3286 ± 0.0180

Table 2 Computational efficiency comparison of YOLO-family detectors.

Model	Parameters (M)	GFLOPs	FPS
YOLOv8n	3.01	4.10	**140.7**
YOLOv26n (RGB)	2.50	2.89	98.4
YOLOv26n (Grayscale)	2.50	2.89	92.6

Table 1 summarises the detection performance obtained through 5-fold cross-validation. In addition to detection accuracy, computational efficiency was evaluated

for the YOLO-based detectors, as shown in Table 2. We have to balance the mAP50 and mAP50-95 as the balancing increase the detection threshold and area coverage simultaneously. Among all evaluated detectors, YOLOv8n achieves the strongest overall performance, attaining the highest mAP50 (0.8856), precision (0.8796), recall (0.8647), and F1-score (0.8699). Notably, YOLOv8n also exhibits relatively low cross-fold variance, indicating stable generalisation across varying imaging conditions. Although YOLOv26n (RGB) achieves a marginally higher mAP50-95 score (0.5520 versus 0.5424), YOLOv8n consistently outperforms all competing models across the remaining metrics. Given that reliable object detection rather than extremely precise boundary localisation constitutes the primary objective of the proposed pipeline, YOLOv8n was selected as the detector for all subsequent experiments.

The comparison between the RGB and grayscale variants of YOLOv26n provides insight into the role of colour information in welding imagery. While welding scenes are visually dominated by intensity variations arising from arc emission and molten material, the RGB model consistently outperforms its grayscale counterpart across all evaluation metrics. Specifically, the RGB variant achieves higher mAP50, mAP50-95, precision, recall, and F1-score, while also exhibiting lower cross-fold variability. These findings suggest that residual chromatic information, although subtle, contributes useful cues for distinguishing weld beads and molten pools from surrounding background structures.

Transformer-based detectors perform substantially below the YOLO-family models in the present setting. Among the transformer architectures, LW-DETR Tiny achieves an mAP50 of 0.4407, considerably outperforming DINO-DETR (0.2916), yet still falling well short of the 0.83–0.89 range achieved by the convolutional detectors. One plausible explanation is that DETR-family architectures typically require larger and more diverse training datasets to learn robust spatial representations through self-attention mechanisms. Given the limited dataset size of 335 annotated frames, the available training data may be insufficient for effective transformer-based optimisation despite the extended training schedule. While LW-DETR demonstrates that lightweight transformer designs can partially mitigate this limitation, neither transformer-based detector approaches the accuracy achieved by the YOLO models under the current experimental conditions.

Computational efficiency further strengthens the case for selecting YOLOv8n. Despite containing slightly more parameters and computational complexity than YOLOv26n, YOLOv8n achieves the highest inference throughput at 140.7 FPS, exceeding YOLOv26n (RGB) and YOLOv26n (Grayscale) by approximately 43% and 52%, respectively. This result demonstrates that YOLOv8n provides not only superior detection accuracy but also the computational efficiency required for real-time welding process monitoring applications.

Representative detection outputs are shown in Fig. 5 and 6. The detector successfully localises both weld beads and molten pools across a wide range of illumination conditions, weld geometries, and process states. Accurate localisation is maintained despite strong reflections, varying arc intensities, and changing bead morphology. Most localisation inaccuracies occur under severe arc glare and partial occlusion,

where object boundaries become less distinct. Nevertheless, qualitative inspection confirms that the detector remains robust under the majority of operating conditions encountered within the dataset.

Overall, the results demonstrate that lightweight convolutional detectors remain highly effective for small-scale industrial vision datasets. The combination of strong detection accuracy, stable cross-validation performance, and high inference throughput makes YOLOv8n the most suitable detector to generate region proposals in the proposed weld quality assessment pipeline.

A limitation of the present evaluation is the use of frame-level cross-validation. Since neighbouring frames originate from continuous welding sequences, temporal correlations may exist between training and validation samples, potentially leading to optimistic performance estimates. Nevertheless, the adopted protocol provides a consistent and reproducible benchmark for comparing detector architectures and establishing a robust detection baseline for subsequent stages of the proposed framework.

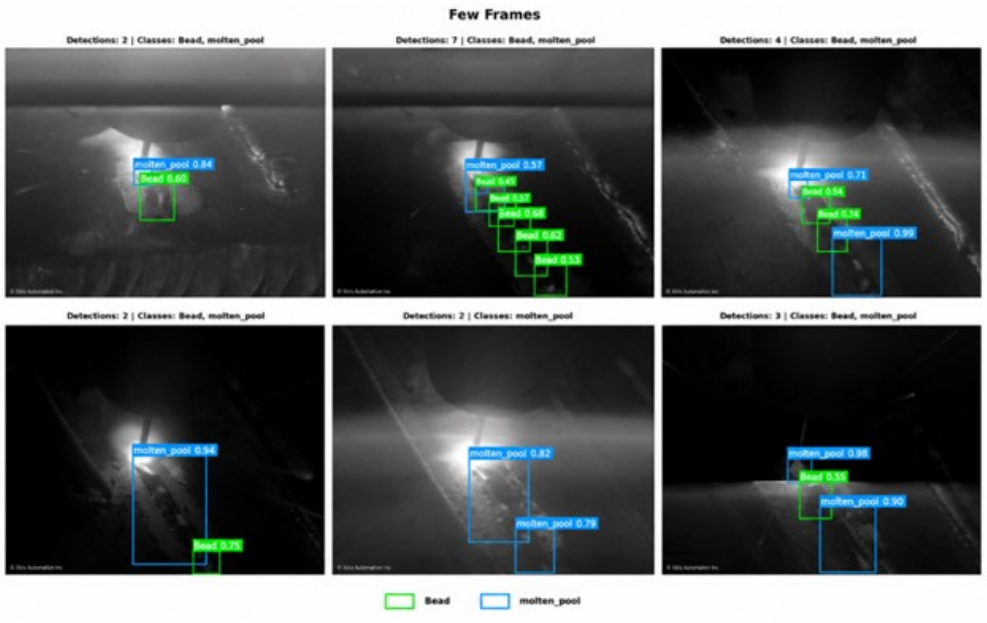
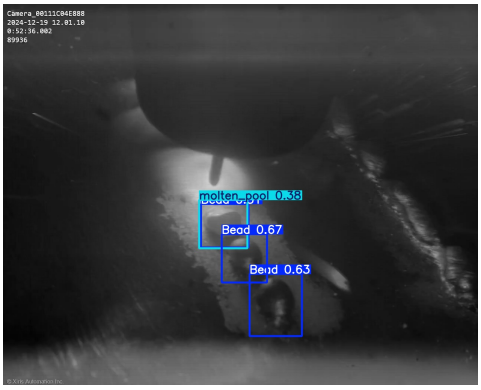


Fig. 5 Representative YOLOv8n detection results across diverse welding conditions, illumination levels, and bead geometries.



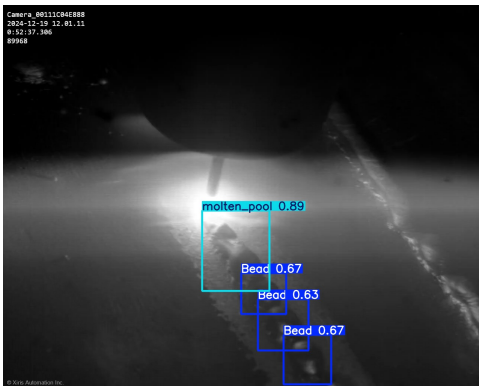
(a) Bead detection



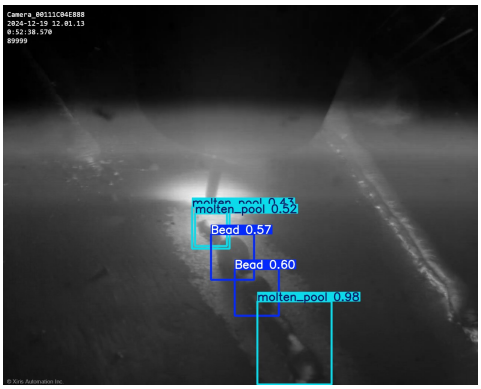
(b) Molten pool detection



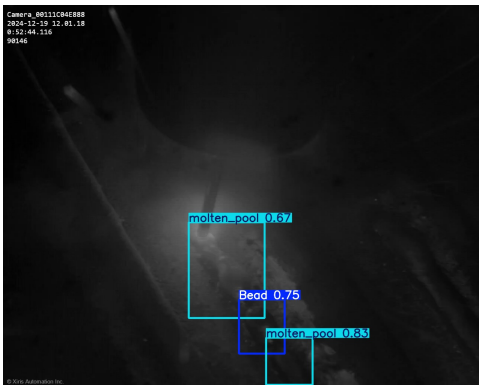
(c) Molten pool detection



(d) Molten and Bead detection



(e) Molten and Bead detection



(f) Molten and Bead detection

Fig. 6 Representative YOLOv8n detection results across diverse welding condition

3 SAM2: Segment Anything Model 2

Although YOLOv8-segmentation provides instance segmentation in a single inference pass, it was not adopted for geometric feature extraction in this work for two primary reasons. First, YOLOv8-segmentation produces a per-frame polygon mask independently for each detected instance, with no mechanism to associate or accumulate mask geometry across frames. However, a weld bead is not a discrete object that appears fully formed, but rather a continuously growing deposit whose morphology evolves throughout the welding process. Consequently, meaningful geometric descriptors such as bead width, deposited cross-sectional area, fill ratio, and aspect ratio consistency can only be reliably computed over the temporal extent of bead formation, rather than from isolated frames.

Second, the molten pool partially overlaps the leading edge of the bead in many frames, leading YOLOv8-segmentation to produce fragmented or merged masks under independent per-frame inference. This further reduces the reliability of frame-level geometric measurements. To address these limitations, a YOLOv8n + SAM2 pipeline was adopted, wherein YOLOv8n provides bounding-box prompts at bead onset and SAM2 propagates a temporally coherent mask tracklet across subsequent frames.

From these tracklets, five nominal geometric features are extracted and temporally aggregated per bead: mean bead width (horizontal mask extent in pixels, normalised by frame width), cumulative deposited area (sum of mask areas across the tracklet), fill ratio (mask area relative to bounding-box area, serving as a proxy for surface regularity), aspect ratio (width-to-height ratio of the mask bounding box), and width deviation (standard deviation of per-frame width measurements, capturing longitudinal uniformity).

For a bead tracklet consisting of T segmented frames, let M_t denote the binary mask at frame t and let (w_t, h_t) represent the width and height of the corresponding mask bounding box. The mask area is computed as

$$A_t = \sum_{x,y} M_t(x, y), \tag{33}$$

where $M_t(x, y) \in \{0, 1\}$. The mean bead width is defined as

$$\bar{W} = \frac{1}{T} \sum_{t=1}^{T} \frac{w_t}{W_f}, \tag{34}$$

where W_f denotes the image width used for normalization. The cumulative deposited area across the tracklet is

$$A_{\text{cum}} = \sum_{t=1}^{T} A_t. \tag{35}$$

The fill ratio is calculated as

$$F = \frac{1}{T} \sum_{t=1}^{T} \frac{A_t}{w_t h_t}, \tag{36}$$

where $w_t h_t$ corresponds to the area of the enclosing bounding box. The average aspect ratio is given by

$$AR = \frac{1}{T} \sum_{t=1}^{T} \frac{w_t}{h_t}. \tag{37}$$

Width deviation, used to quantify longitudinal bead uniformity, is computed as

$$\sigma_W = \sqrt{\frac{1}{T-1} \sum_{t=1}^{T} \left(\frac{w_t}{W_f} - \bar{W} \right)^2}. \tag{38}$$

These features are physically meaningful, as high-quality weld beads typically exhibit stable width, high fill ratio, and low width deviation, whereas defective depositions arising from insufficient current, excessive travel speed, or wire feed instability manifest as irregular cross-sections, tapered profiles, and increased width variability.

The resulting geometric descriptor provides an interpretable, physics-informed complement to the visual appearance pathway in the DPQN classifier, offering additional discriminative capability in cases where beads appear visually similar but differ geometrically. However, limitations were observed in the YOLOv8n + SAM2 pipeline on this dataset. Due to the continuous nature of bead formation, SAM2's object memory mechanism designed for objects with stable boundaries undergoing motion struggled to reliably propagate masks across arc-flash frames and rapid morphological variations characteristic of welding. This resulted in inconsistent tracklets in approximately half of the videos. Accordingly, the extracted geometric features should be interpreted as approximate nominal descriptors rather than precise measurements. The downstream DPQN classifier accounts for this uncertainty through its weakly supervised Multiple Instance Learning (MIL) formulation, which is designed to be robust to such noise.

3.1 Segmentation for Geometric Feature Extraction using Foundation Model

Although object detection accurately localises weld beads and molten pools, bounding boxes alone provide limited geometric information for downstream weld quality assessment. Quantitative characteristics such as bead width, molten pool area, contour irregularity, perimeter, and shape evolution require pixel-level segmentation rather than coarse object localisation. These geometric descriptors are expected to play an important role in the future classification stage of the proposed intelligent welding inspection framework.

A straightforward approach is to train a fully supervised instance segmentation model such as YOLOv8-Segment. However, such methods require manually annotated pixel-level masks, which are significantly more expensive and time-consuming to produce than bounding-box annotations. Since the current dataset contains only bounding-box annotations, generating high-quality segmentation masks for all weld bead and molten pool instances would require substantial additional effort.

To address this limitation, the Segment Anything Model 2 (SAM2) is investigated as a promptable foundation segmentation model. Unlike conventional segmentation networks, SAM2 does not require task-specific segmentation training and can generate segmentation masks directly from sparse prompts. By combining the accurate object localisation provided by YOLOv8n with the prompt-based segmentation capability of SAM2, pixel-level masks can be obtained without manual segmentation annotation.

3.2 YOLOv8n guided SAM2 video Segmentation

The segmentation framework builds upon the detection pipeline presented in the previous section. First, YOLOv8n detects the weld bead and molten pool in each welding frame. Instead of prompting SAM2 for every individual frame, the predicted bounding boxes are provided as prompts every 15 frames, after which SAM2 propagates the segmentation masks across intermediate frames using its internal temporal memory mechanism. Given a video sequence

$$V = \{I_t\}_{t=1}^{T}, \tag{39}$$

YOLOv8n generates a set of detections

$$B_t = \{b_t^j\}_{j=1}^{N_t}, \tag{40}$$

where b_t^j denotes the jth bounding box in frame t. Bounding-box prompts are provided at intervals of

$$\Delta t = 15, \tag{41}$$

frames. SAM2 then propagates the corresponding segmentation masks

$$M_t = \{m_t^j\}_{j=1}^{N_t} \tag{42}$$

through intermediate frames using its temporal memory mechanism. The overall workflow is illustrated in Fig. 7.

This hybrid approach significantly reduces computational overhead while preserving temporal consistency across video sequences. Compared with frame-wise segmentation, the memory-based propagation mechanism enables smoother mask evolution throughout the welding process and eliminates the need for repeated prompting in every frame.

Representative segmentation results are presented in Fig. 8, where it shows the original welding frames, presents the corresponding YOLOv8n detections, and the segmentation masks generated by SAM2. The displayed frames demonstrate the temporal evolution of segmentation results across an entire welding sequence.

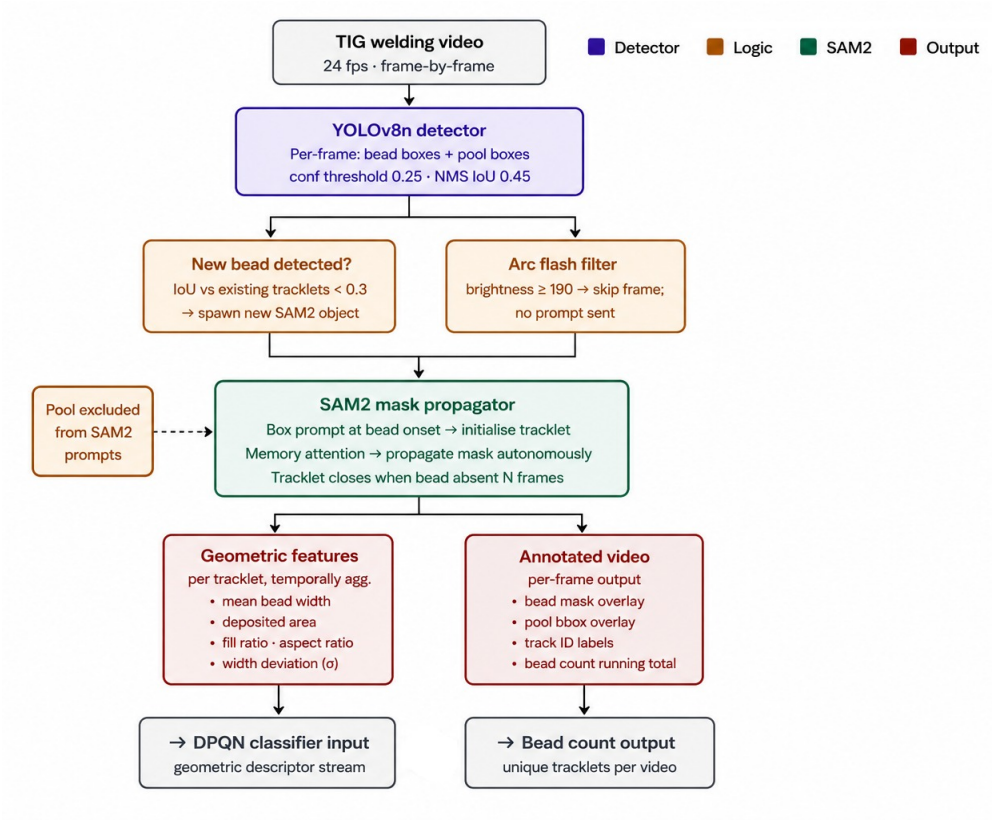


Fig. 7 YOLOv8n Guided SAM2 Video Segmentation Architecture

3.3 Results

Qualitative evaluation demonstrates that the proposed YOLOv8n-guided SAM2 framework successfully tracks the overall welding region throughout the video sequence. Once initial prompts are provided, SAM2 effectively propagates segmentation masks across subsequent frames while adapting to gradual changes in weld geometry, molten pool evolution, and illumination conditions. The resulting masks exhibit strong temporal continuity, making them suitable for estimating global geometric properties of the welding process.

However, an important limitation was consistently observed. Although SAM2 accurately segments the overall welding region, it often struggles to distinguish the deposited weld bead from the active molten pool. Instead of producing two well-separated semantic regions, the generated masks frequently merge both structures into a single continuous segment or produce unstable boundaries near the transition between molten and solidified material as shown in Fig.9.

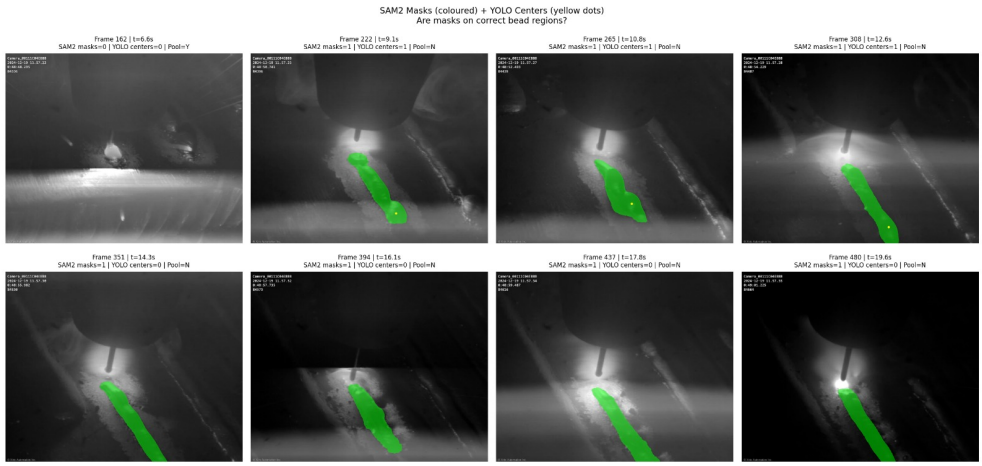
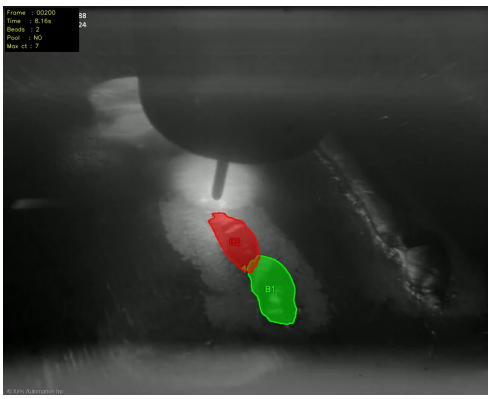


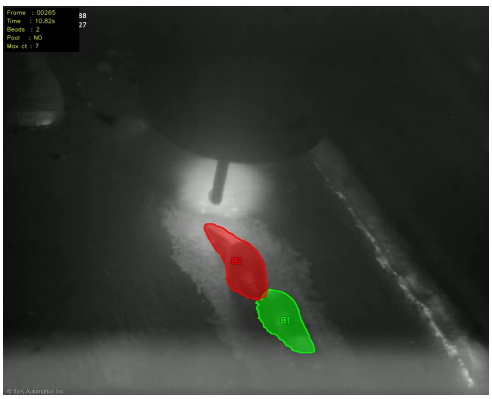
Fig. 8 Representative segmentation results across a welding sequence showing original frames, YOLOv8n detections, and SAM2 masks, illustrating temporal evolution

This behavior is closely related to the physical characteristics of the welding process. Unlike conventional object segmentation tasks involving visually distinct objects, the weld bead and molten pool form a continuous physical phenomenon. During welding, the molten pool gradually solidifies into the deposited bead, resulting in smooth intensity transitions rather than clearly defined object boundaries. Furthermore, intense arc radiation, thermal blooming, specular reflections, and local overexposure significantly reduce boundary contrast, making semantic separation inherently difficult even for human annotators. The observed limitation is also consistent with the design philosophy of foundation segmentation models. SAM2 is primarily trained on large-scale natural image datasets containing discrete objects with well-defined semantic boundaries, such as humans, vehicles, animals, and everyday objects. Consequently, its learned representations are well suited for separating spatially independent instances but are less effective for industrial processes involving continuous material flow and evolving object morphology.

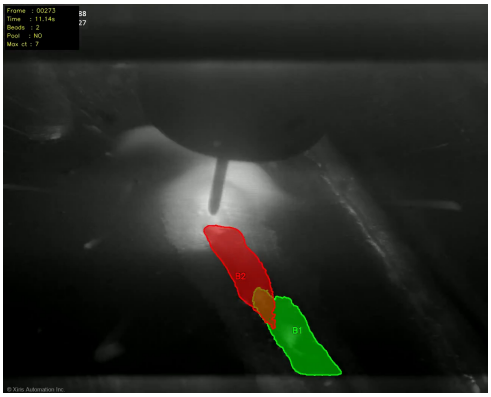
Despite these challenges, the generated segmentation masks remain valuable for extracting global geometric descriptors, including overall weld region shape, contour, area, perimeter, and temporal evolution of the weld profile. These features provide complementary information beyond object detection and are expected to support future weld quality classification models. Overall, this study highlights both the potential and current limitations of applying foundation segmentation models to industrial welding inspection. While the combination of YOLOv8n and SAM2 enables prompt-based video segmentation without requiring manually annotated masks, accurate semantic separation between weld beads and molten pools remains an open challenge. Future work will investigate domain-adaptive fine-tuning, specialised industrial segmentation architectures, and hybrid detection–segmentation frameworks to improve boundary discrimination while preserving temporal consistency.



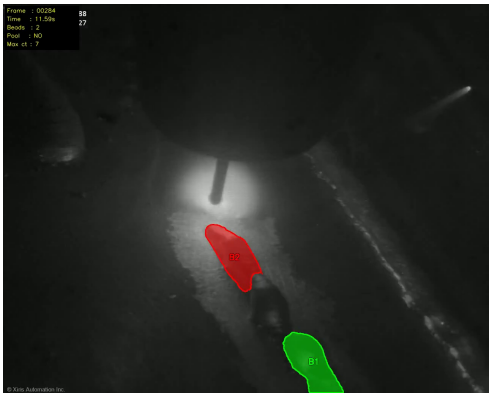
(a) Initial weld region segmentation



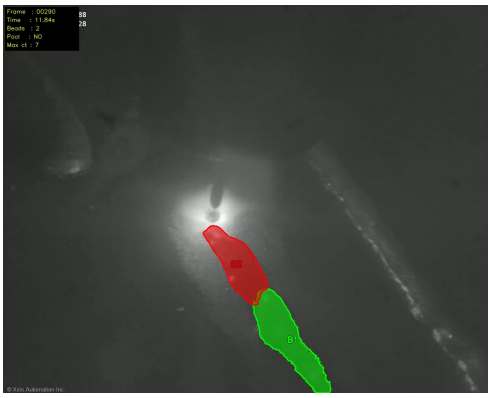
(b) Early bead – pool overlap



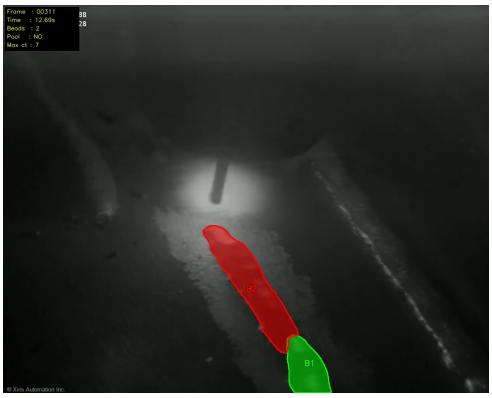
(c) Merged bead and molten pool mask



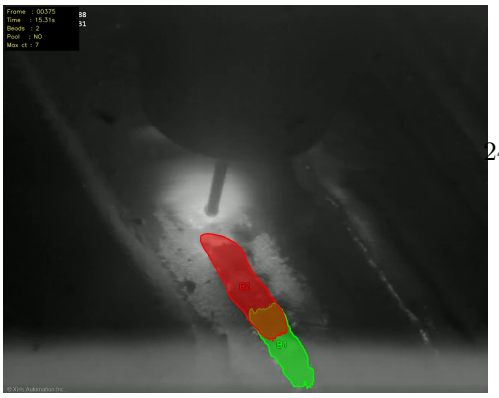
(d) Boundary ambiguity under arc glare



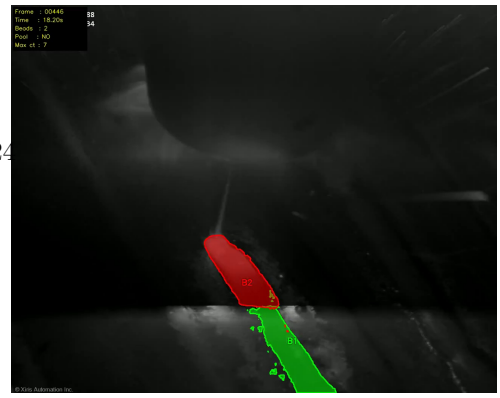
(e) Partial separation attempt



(f) Unstable segmentation boundary



(g) Blended bead and pool region



(h) Inconsistent final mask output

Fig. 9 Qualitative examples of SAM2-based segmentation outputs across welding frames. While the overall welding region is detected, the model often merges the weld bead and molten pool or produces unstable boundaries, particularly under arc glare and dynamic transitions.

4 Conclusion

This work establishes a rigorous, reproducible baseline for real-time weld bead and molten pool detection in TIG welding through the systematic evaluation of five modern object detection architectures under 5-fold stratified cross-validation. Our results demonstrate that lightweight convolutional detectors, specifically YOLOv8n, retain a decisive advantage on small-scale industrial datasets, achieving an mAP50 of 0.8856 with stable cross-fold variance and an inference throughput of 140.7 FPS markedly superior to transformer based alternatives that suffer from data starvation. The investigation further reveals that residual chromatic information in RGB welding imagery provides non-negligible discriminative cues, as evidenced by the consistent performance gap between the RGB and grayscale variants of YOLOv26n. To address the annotation bottleneck inherent to pixel-level supervision, we introduced a YOLOv8n-guided SAM2 pipeline that leverages sparse bounding-box prompts and temporal memory propagation to generate continuous segmentation tracklets without manual mask annotation. While this hybrid approach successfully tracks the global weld region and extracts interpretable geometric descriptors such as bead width, deposited area, and fill ratio—its applicability is bounded by a fundamental domain mismatch: SAM2's object-centric representations, learned from natural images with discrete semantic boundaries, struggle to disentangle the weld bead and molten pool, which constitute a continuous physical phenomenon with smooth intensity transitions. This limitation underscores the need for domain-adaptive or physics-informed segmentation architectures in manufacturing environments. Several constraints warrant acknowledgement. The current dataset comprises eleven videos and 335 frames; although temporally pruned to reduce redundancy, its scale remains modest relative to general-domain benchmarks. The frame-level cross-validation protocol, while standard for architectural comparison, may yield optimistic estimates due to temporal correlations among neighbouring frames. Future work will transition to video-level cross-validation and extend the dataset to encompass broader process parameters, materials, and defect morphologies. Looking ahead, the proposed Detection-Prompted Quality Network (DPQN) offers a principled pathway toward end-to-end weld quality assessment. By integrating YOLOv8n with BoT-SORT tracking, multi-branch MobileNetV3-Small feature extraction, FiLM-conditioned process parameter fusion, and a weakly supervised MIL head, the DPQN framework promises to translate spatially localised detections into video-level quality scores without requiring costly frame-level labels. Ultimately, this research delineates both the immediate feasibility and the practical boundaries of deploying generic deep learning and foundation models in WAAM inspection, charting a course toward robust, adaptive, and fully autonomous welding quality assurance systems.

5 Future Work

An immediate extension of the proposed detection pipeline is the full realization of a Detection-Prompted Quality Network (DPQN), a multi-stage architecture for welding quality assessment built upon the YOLOv8n detector, identified as the strongest

baseline. In this framework, the trained YOLOv8n model is integrated with a BoT-SORT multi-object tracker to generate temporally consistent tracklets of weld beads and molten pools across video sequences. BoT-SORT leverages Kalman filter–based motion prediction combined with appearance re-identification, making it particularly suitable for TIG welding scenarios characterized by rapid morphological variations in the weld bead and intermittent visibility of the molten pool due to arc flash.

For each tracked frame, three spatially distinct image crops are extracted: (i) a 224×224 bead patch centered on the detected bead, (ii) a 224×224 molten pool patch aligned with the corresponding pool detection, and (iii) a 224×224 downsampled full-frame patch capturing global contextual information, including the surrounding thermal environment. These inputs are processed in parallel using a shared MobileNetV3-Small backbone pretrained on ImageNet, with only the final two layers fine-tuned. The resulting feature vectors are concatenated and linearly projected to form a unified visual token per frame. The shared-weight configuration is essential given the limited dataset size (eleven videos), as it mitigates overfitting across branches while enabling joint representation learning from bead appearance, molten pool dynamics, and scene-level context.

To incorporate process-awareness, welding parameters (voltage, current, wire feed speed, and travel speed) are injected using Feature-wise Linear Modulation (FiLM). A lightweight multilayer perceptron (MLP) generates scaling (γ) and shifting (β) vectors, which are applied element-wise to the visual token. This conditioning mechanism allows the model to adapt its interpretation of visual features according to the operating regime, addressing ambiguities where identical visual patterns may correspond to different quality outcomes under varying process conditions.

The sequence of FiLM-modulated tokens is subsequently processed by a lightweight transformer encoder comprising two layers, two attention heads, a model dimension of 128, and a dropout rate of 0.4. This component captures temporal dependencies in bead formation while maintaining computational efficiency. For final prediction, a weakly supervised Multiple Instance Learning (MIL) head with noisy-OR pooling aggregates frame-level predictions into a video-level quality score. Notably, this formulation requires only video-level annotations (good/bad), eliminating the need for frame-level labels, which are costly and difficult to obtain in industrial environments.

The proposed DPQN framework would be evaluated using leave-one-video-out cross-validation across the eleven available videos. Furthermore, an ablation study would systematically remove key components (FiLM conditioning, the molten pool branch, the temporal transformer, and the MIL head) to quantify their individual contributions to the overall classification performance.

6 Declarations

6.1 Funding

Author Sagar Nikam acknowledges the funding support provided by the Department for Science, Innovation and Technology (DSIT), through International Science Partnership Funding (ISPF) awarded to Ulster University via the Department for the Economy (DfE), Northern Ireland.

6.2 Competing 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.

6.3 Author Contributions

Sayak Halder: conceptualisation, formal analysis, investigation, methodology, validation, writing—original draft. **Shahana Bano:** conceptualisation, supervision, investigation, methodology, project administration, validation, writing—drafting and editing **Anil Prathuru:** conceptualisation, methodology, project administration, validation, writing—drafting and editing. **Somasundar Kannan:** formal analysis, writing—drafting and editing. **Shiva Sekar:** methodology, data acquisition, validation. **Deepika Nikam:** methodology, data acquisition, validation. **Sagar Nikam:** funding acquisition, conceptualisation, methodology, data acquisition, writing—drafting and editing.

6.4 Availability of Data and Materials

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

References

[1] K. W. Kim, A. Kamerkar, T.-E. Chiu, I. Abdi, J. Qin, W. Suder, and S. Asif, "WAAM-ViD: Towards universal vision-based monitoring for wire arc additive manufacturing," Cranfield University, United Kingdom, 2023.

[2] M. Chandra, S. Rajak, and K. E. K. Vimal, "Deep learning-based framework for the observation of real-time melt pool and detection of anomaly in wire-arc additive manufacturing," *Materials and Manufacturing Processes*, vol. 39, no. 6, pp. 761–777, 2024. doi: 10.1080/10426914.2023.2254386.

[3] M. Chandra, A. Kumar, S. K. Sharma, K. Kazmi, and S. Rajak, "Deep learning for anomaly detection in wire-arc additive manufacturing," *Welding International*, vol. 37, no. 8, pp. 457–467, 2023. doi: 10.1080/09507116.2023.2252733.

[4] Tabassum, A., and Ziabari, A. (2024). "Adapting Segment Anything Model (SAM) to Experimental Datasets via Fine-Tuning on GAN-based Simulation: A Case Study in Additive Manufacturing." http://arxiv.org/abs/2412.11381

[5] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," Ultralytics, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[6] N. Ravi *et al.*, "SAM 2: Segment Anything in Images and Videos," *arXiv preprint arXiv:2408.00714*, 2024.

[7] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "SAM 2: Segment Anything in Images and Videos," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. doi: 10.48550/arXiv.2408.00714.

[8] A. Kirillov *et al.*, "Segment Anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026.

[9] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 213–229. doi: 10.1007/978-3-030-58452-8_13.

[10] H. Zhang *et al.*, "DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[11] A. Aharon, R. Orfaig, and S. Bobrovsky, "BoT-SORT: Robust Associations Multi-Pedestrian Tracking," *arXiv preprint arXiv:2206.14651*, 2022.

[12] A. Howard *et al.*, "Searching for MobileNetV3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1314–1324.

[13] B. Li *et al.*, "Exploring Few-Shot Defect Segmentation in General Industrial Scenarios with Metric Learning and Vision Foundation Models," *arXiv preprint arXiv:2502.01216*, 2025.

[14] A.-M. Nica-Sorin, H. Zahran, and V. S. Bujoreanu, "Leveraging Segment Anything Model (SAM) for Weld Defect Detection in Industrial Ultrasonic B-Scan Images," *Sensors*, vol. 25, no. 1, p. 277, 2025. doi: 10.3390/s25010277.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[16] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.