machines

MDPI

Article

Weld-FHG-YOLO: A Lightweight Multi-Frequency Feature Fusion Network for Weld Keypoint Localization

Yunsong Yan [1], Xiaoning Meng [2], Wei Liu [3], Hougao Wang [3], Haiyang Liu [4], Chao Chen [1,5,6,*] and Fuxin Du [6]

[1] School of Rail Transportation, Shandong Jiaotong University, Jinan 250357, China; 25221003@stu.sdjtu.edu.cn
[2] Jinan Heavy Industries Group Co., Ltd., Jinan 250109, China; 15216407942@139.com
[3] Shandong Nuclear Power Equipment Manufacturing Co., Ltd., Yantai 265100, China; 6v2006@163.com (W.L.); wangli0212190@163.com (H.W.)
[4] School of Construction Machinery, Shandong Jiaotong University, Jinan 250357, China; haiyangliu@sdjtu.edu.cn
[5] Shantui Construction Machinery Co., Ltd., Jining 272073, China
[6] School of Mechanical Engineering, Shandong University, 17923 Jingshi Road, Jinan 250061, China; dufuxin@sdu.edu.cn
* Correspondence: chaochen@sdjtu.edu.cn

Abstract

Accurate weld keypoint localization is an important visual perception task for robotic welding, weld tracking, and intelligent manufacturing. However, weld keypoints in line-structured light images are usually small, weakly textured, and easily affected by reflections, noise, and spurious laser stripes. These factors make it difficult for lightweight detection models to balance localization accuracy and computational efficiency. To address this problem, this paper proposes Weld-FHG-YOLO, a lightweight multi-frequency feature fusion network for weld keypoint localization. The proposed model is built on the You Only Look Once version 11 nano (YOLO11n) framework and focuses on optimizing feature fusion and scale transformation in the Neck. Specifically, FasterC3K2 is introduced to replace the original C3K2 modules in the Neck, thereby reducing redundant computation during multi-scale feature fusion. In addition, a Haar Wavelet Decomposition and Group Shuffle Convolution (HWD-GSConv) downsampling fusion module is designed, in which Haar wavelet decomposition preserves low-frequency structural information and high-frequency details, while GSConv performs lightweight multi-frequency feature fusion. Experimental results show that Weld-FHG-YOLO achieves 2.301 M parameters and 6.016 GFLOPs, which are 11.26% and 6.83% lower than those of YOLO11n, respectively. Meanwhile, mAP@0.5:0.95 increases from 0.7337 to 0.7901, the Mean Center Error (MCE) decreases from 2.254 px to 2.131 px, and the CPU inference speed increases from 13.69 to 14.77 frames per second (FPS). These results indicate that the proposed method improves strict localization accuracy and localization stability while maintaining a lightweight computational profile, providing a practical visual perception approach for weld keypoint localization in resource-constrained intelligent manufacturing scenarios.

Keywords: weld keypoint localization; line-structured light images; robotic welding perception; lightweight detection network; multi-frequency feature fusion; haar wavelet downsampling; intelligent manufacturing

Check for updates

Academic Editor: Antonios Gasteratos

Received: 12 August 2026
Revised: 28 August 2026
Accepted: 31 August 2026
Published: 7 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Welding is a critical process in shipbuilding, pressure-vessel manufacturing, rail transit, and steel-structure production, where weld formation quality directly affects the service safety and production efficiency of structural components. In practical manufacturing, assembly errors, thermal deformation, workpiece positioning deviations, and process disturbances can cause the actual weld trajectory to deviate from the predefined path. If welding robots still rely on conventional teach-and-repeat trajectories, such deviations may lead to weld offset, incomplete penetration, undercutting, or repeated rework, thereby increasing material waste, downtime, and quality-inspection costs. Therefore, real-time and stable acquisition of weld position information in complex welding environments is essential for improving robotic welding automation and reducing manufacturing quality risks. Owing to their non-contact measurement, intuitive information representation, and relatively strong resistance to environmental interference, structured-light vision sensors have become an important technical route for weld position perception in intelligent welding robots. Yang et al. [1] systematically reviewed advances in structured-light sensing for intelligent welding robots and emphasized that visual perception is a key component for online weld tracking and path correction. Li et al. [2] used a laser vision sensor to recognize spatially intermittent weld trajectories, providing a visual measurement basis for automated skip-welding guidance. Zou et al. [3] developed a real-time weld seam tracking control system based on line-laser vision, demonstrating the practical value of active vision in welding trajectory correction.

Against this application background, this study focuses on keypoint localization in line-structured light weld images. Weld feature points are typically located at bending, fracture, intersection, or local mutation regions of the laser stripe. They provide essential front-end information for weld centerline extraction, start-point guidance, groove geometry recognition, and robotic path planning [4–6]. Therefore, weld keypoint localization is not an isolated image-recognition task, but a core perception step that directly supports welding trajectory correction and robot motion control. Although these feature points physically correspond to spatial pixel coordinates, direct coordinate regression can be easily disturbed by isolated noise in welding environments with strong interference and local stripe deformation. Accordingly, this study formulates weld feature point localization as an object detection task by annotating local regions that contain geometric stripe features. This formulation improves the robustness of feature point region recognition and is better suited to industrial single-stage detection frameworks.

Early studies on weld visual detection mainly relied on traditional image-processing methods, such as threshold segmentation, gray-scale centroid extraction, corner detection, and tracking filters [7–9]. However, when welding images contain spatter, smoke, arc reflections, or strong metallic reflections, these methods are prone to false edges and laser-stripe discontinuities, resulting in limited robustness under complex welding conditions [10].

With the development of deep learning, weld feature point localization has increasingly been formulated as object detection, semantic segmentation, or detection followed by post-processing, aiming to improve model adaptability to complex backgrounds and irregular laser-stripe patterns [11,12]. Gao et al. [13] proposed YOLO-Weld, which divided weld feature points into 16 categories and reformulated feature point localization as an object detection task, thereby improving detection robustness under extreme welding noise. Gao et al. [14] developed WeldLight, which combines lightweight design, noise augmentation, and attention for CPU-based weld classification and keypoint localization. Yang et al. [15] proposed StripePoint-YOLO, which combines a P2 detection head, SPDConv, a joint WIoU v3–NWD regression loss, and physics-driven data augmentation to improve multi-type weld keypoint localization under severe industrial interference. In addition, Li et al. [16] developed

a robotic welding guidance system based on an improved YOLOv5 algorithm and a RealSense depth camera, while Zhu et al. [17] applied an improved YOLOv5 model to steel-structure weld recognition. These studies indicate that deep learning, particularly YOLO-based detection models, can improve the automation level of weld visual detection.

Advances in general object detection and lightweight network design have provided a technical foundation for weld keypoint localization. YOLO enables efficient single-stage detection [18], while FPN supports multi-scale feature fusion [19]. SSD and Fast R-CNN represent single-stage and two-stage detection frameworks, whereas CenterNet and DETR provide center-based and Transformer-based localization paradigms [20–25]. MobileNet and MobileNetV2 reduce computational cost through depthwise separable convolution [26,27], while ShuffleNet and ShuffleNetV2 improve channel interaction and inference efficiency [28,29]. Related studies have further emphasized resource-constrained deployment and position-aware lightweight feature modeling [30,31]. Recent YOLO11 variants also demonstrate the adaptability of compact models to industrial vision tasks [32,33]. Nevertheless, these general-purpose designs do not explicitly address the high-frequency laser-stripe cues and center-localization sensitivity of weld keypoints.

Despite these advances, existing weld-specific methods have not fully resolved the joint requirements of preserving localization-sensitive stripe details and maintaining CPU-oriented computational efficiency. First, weld feature points are extremely small, and their discriminative information is highly concentrated around laser-stripe edges, bends, and local mutation regions. Recent studies have similarly highlighted the difficulty of reliable weld-keypoint localization in noisy and geometrically complex welding images, as well as the need for efficient task-specific feature extraction [34,35]. Conventional stride convolution or pooling-based downsampling may smooth these high-frequency details while reducing spatial resolution, leading to deviations in the predicted box center. Second, industrial welding systems are often deployed on CPUs or embedded industrial computers. Although simply increasing network depth or adding attention mechanisms may improve accuracy, it also increases the number of parameters, computational cost (GFLOPs), and memory access cost (MAC), making it difficult to meet the real-time requirements of closed-loop control.

To address these challenges, this paper proposes Weld-FHG-YOLO, a lightweight weld keypoint localization model based on YOLO11n. The Backbone and Detect Head remain unchanged, while the Neck is redesigned using two complementary modules. FasterC3K2 reduces redundant computation during multi-scale feature fusion [36], whereas HWD-GSConv preserves low- and high-frequency stripe information during downsampling through Haar wavelet decomposition and lightweight GSConv fusion [37,38].

The main contributions of this study are summarized as follows:

1. A lightweight weld keypoint localization network, named Weld-FHG-YOLO, is proposed for line-structured light weld images. The network is designed as a visual perception model for robotic welding and intelligent manufacturing scenarios, with emphasis on localization accuracy, model compactness, and CPU inference efficiency.
2. A FasterC3K2-based lightweight feature fusion module is introduced into the Neck of YOLO11n. By reducing redundant spatial computation during multi-scale feature fusion, the module decreases model parameters and computational complexity while maintaining effective feature representation.
3. HWD-GSConv downsampling fusion module is designed to enhance multi-frequency feature representation during scale transformation. Haar wavelet decomposition is used to preserve low-frequency structural information and high-frequency laser-stripe details, while GSConv performs lightweight fusion of the decomposed features.

4. The proposed model is systematically evaluated using baseline comparison, ablation experiments, comparisons with representative detection models, CPU inference-speed testing, MCE analysis, and weld-type-specific localization-error visualization. The results demonstrate that Weld-FHG-YOLO improves strict localization accuracy and localization stability while maintaining a lightweight computational profile.

The remainder of this paper is organized as follows. Section 2 introduces the image acquisition platform, dataset construction, and dataset partitioning. Section 3 presents the overall architecture of Weld-FHG-YOLO, including the FasterC3K2 module and the HWD-GSConv downsampling fusion module. Section 4 describes the experimental settings, evaluation metrics, baseline comparison, ablation study, model comparison, and visual error analysis. Section 5 discusses the task-specific rationale of the proposed modules and analyzes the balance among detection accuracy, localization stability, and model complexity. Section 6 concludes the paper and outlines future research directions.

2. Hardware Platform and Dataset

2.1. Hardware Platform and Image Acquisition Principle

The weld images used in this study were acquired using a robotic welding line-structured light vision acquisition system. The industrial camera used in the system was a HIKROBOT CB016 (Hikrobot, Hangzhou, China), with a 1/2.9-inch sensor, a resolution of 1440 × 1080 pixels, and a 25 mm focal-length lens. The central wavelengths of the single-line laser emitter and the narrowband filter were both approximately 660 nm. During image acquisition, the single-line laser projected a structured-light stripe onto the weld region, and the industrial camera captured the laser stripe image on the weld surface from an oblique viewing angle. The narrowband filter was used to suppress the influence of ambient light and arc reflection on imaging quality, while the control circuit ensured the stability of laser projection and image acquisition. When the structured light illuminated grooves, lap edges, or height-changing regions, the weld geometry caused bending, fracture, or intersection of the laser stripe. These local geometric mutations usually correspond to key feature points required for weld tracking and robotic path planning [1,13,14].

During image acquisition, the industrial camera and the single-line laser maintained a relatively fixed spatial pose, enabling stable two-dimensional representations of structured-light stripes across different weld types. Figure 1 shows representative weld images acquired by the system, including V-groove, butt, Y-groove, lap, and fillet welds. To further illustrate the visual characteristics and detection challenges of the collected data, the figure also provides example keypoint annotations for these five typical weld types.

As shown in Figure 1, different weld types exhibit distinct stripe morphologies and different numbers of keypoints. In addition, the images contain various sources of interference, such as spatter, smoke, local overexposure, strong metallic reflection, and stray laser stripes. Therefore, a weld keypoint detection model must not only accurately capture local mutation regions in structured-light stripes, but also maintain strong resistance to interference and real-time inference capability.

2.2. Weld Keypoint Dataset

In this study, a weld keypoint detection dataset was constructed for training and evaluating the proposed Weld-FHG-YOLO model. After image acquisition, each raw image was horizontally center-cropped to 1280 × 1080 pixels. This operation removed peripheral regions outside the effective weld observation area while preserving the original vertical resolution. Consequently, all images included in the final dataset had a resolution of 1280 × 1080 pixels. Each image contains one or more weld keypoints. Each keypoint was annotated using a 40 × 40-pixel bounding box centered on its manually identified

position. This size retained the surrounding laser-stripe geometry while limiting overlap with neighboring keypoints. The boxes represent standardized local regions rather than physical object boundaries. The annotations followed the YOLO detection format, in which each target was represented by a class index and normalized bounding-box coordinates. During inference, the predicted box center was taken as the keypoint coordinate. This formulation accommodates varying numbers and categories of keypoints across weld types and utilizes local structural context without requiring a predefined keypoint order.

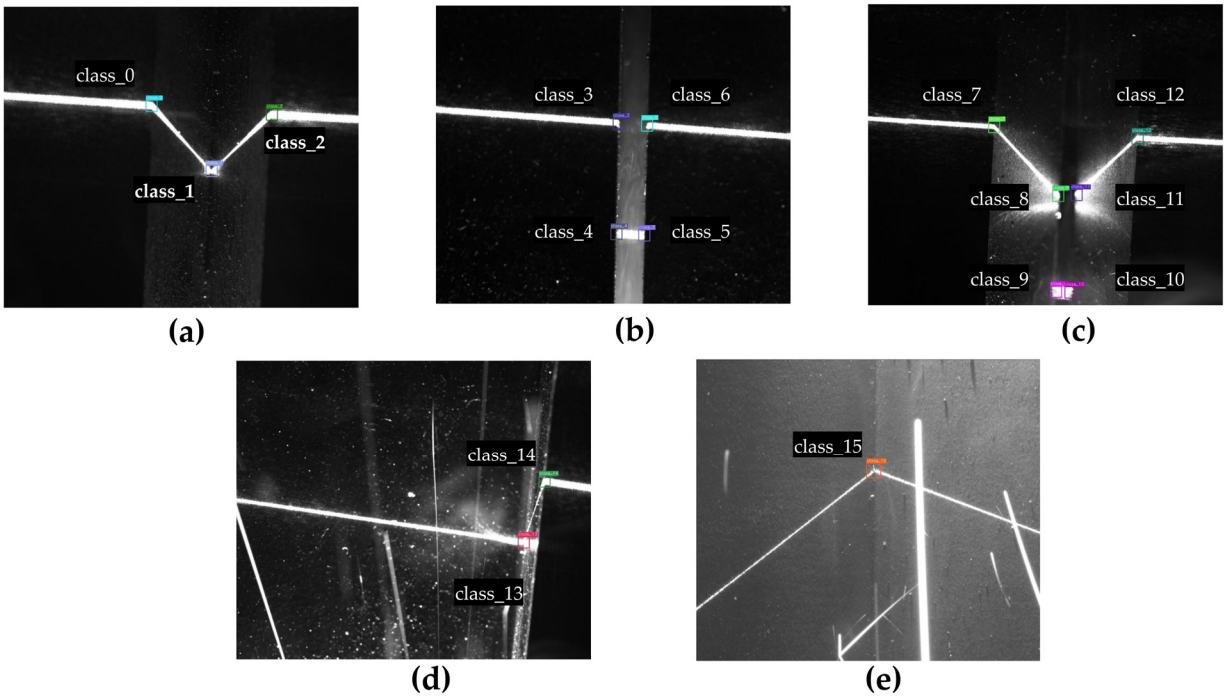


Figure 1. Representative annotated line-structured-light images of five weld types acquired using the vision sensor. Colored boxes indicate the annotated keypoint regions, while class_0 to class_15 denote the 16 keypoint categories defined in the dataset. (**a**) V-groove weld (class_0–class_2); (**b**) butt weld (class_3–class_6); (**c**) Y-groove weld (class_7–class_12); (**d**) lap weld (class_13–class_14); (**e**) fillet weld (class_15).

Unlike traditional image-processing-based methods for weld feature point extraction, this study formulates weld keypoint localization as an object detection task. The model takes weld structured-light images as input and outputs the class and location of the local region containing each keypoint. Because weld keypoints are usually small, their discriminative features are mainly concentrated around bending, fracture, or local mutation regions of the structured-light stripe. They are also susceptible to interference from spatter, smoke, and reflections. Therefore, this dataset reflects the small-target localization and complex-background interference challenges commonly encountered in industrial weld visual detection.

The dataset contains 16 categories of weld keypoints, with class indices ranging from class_0 to class_15. These categories correspond to keypoints associated with different weld types or different geometric locations. The strategy of subdividing multi-type weld feature points into 16 categories and localizing them in an object detection format is consistent with previous weld keypoint localization studies, such as YOLO-Weld [13]. To ensure fair model training and evaluation, all categories were included in the training, validation, and test sets, and the instance distribution of each category was kept as consistent as possible across the three subsets.

The dataset was manually annotated. For keypoints with clear boundaries, annotations were made according to the geometric mutation positions of the structured-light stripe. For keypoints affected by local noise or reflections, the annotation was determined by considering both the continuity of adjacent stripes and the weld geometric structure. This annotation strategy allows the center of each bounding box to be as close as possible to the actual weld keypoint position, providing a reliable basis for the subsequent center localization error analysis. This procedure reduces reliance on local brightness and helps control center-position uncertainty caused by noise and reflections. Detailed examples of keypoint locations for different weld types are shown in Figure 2.

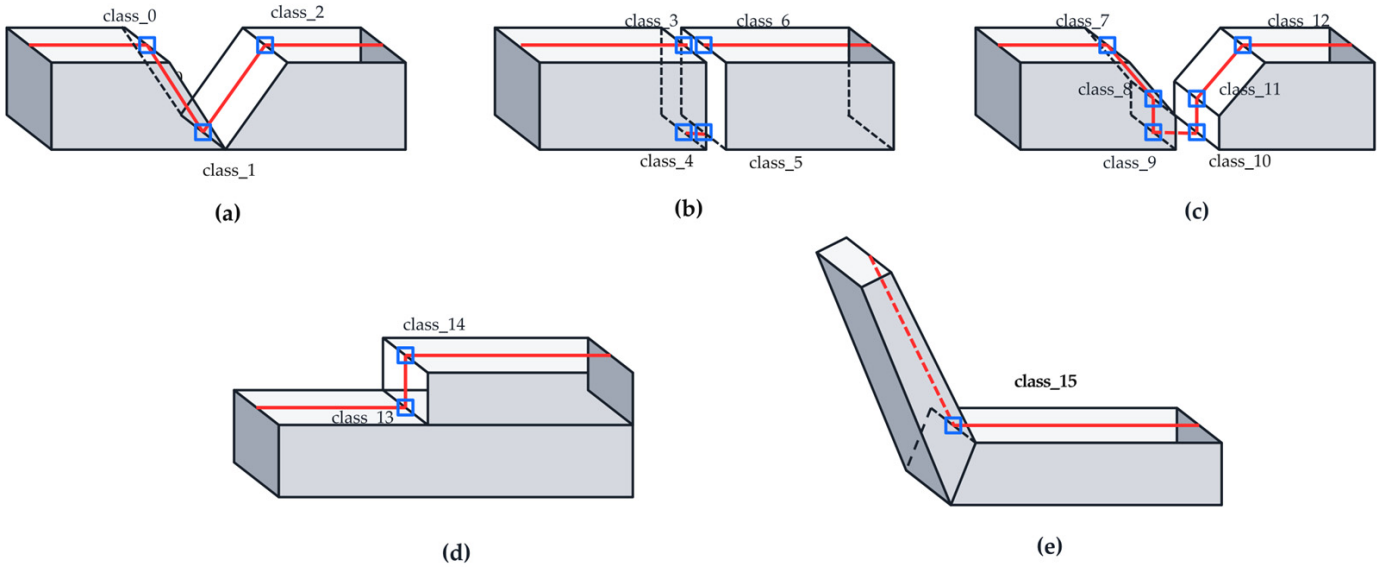


Figure 2. Schematic examples of feature point locations for different weld types. The V-groove, butt, Y-groove, lap, and fillet welds contain 3, 4, 6, 2, and 1 keypoint categories, respectively. (**a**) V-groove weld; (**b**) butt weld; (**c**) Y-groove weld; (**d**) lap weld; (**e**) fillet weld.

2.3. Dataset Partitioning and Class Distribution

During data acquisition, the sensor recorded multiple independent welding video sequences. Because the imaging process was continuous, adjacent frames often contained highly similar visual content. Randomly distributing such frames across different subsets could introduce correlations and produce overoptimistic performance estimates. To reduce this risk, all extracted frames were manually screened before dataset partitioning. Duplicate and highly similar images were removed, yielding a cleaned dataset of 6671 images.

The cleaned dataset was divided at the frame level rather than the video-sequence level. Stratified sampling based on combinations of keypoint categories was used to maintain similar class distributions across the three subsets. The dataset was divided into training, validation, and test sets at an approximate ratio of 70%:15%:15%.

The training, validation, and test sets contained 4670, 1001, and 1000 images, respectively. Their corresponding numbers of keypoint instances were 14,967, 3206, and 3205. The training set was used for parameter learning, whereas the validation set was used for model selection and hyperparameter tuning. The test set was used only for final performance evaluation. Although manual filtering reduced direct duplication among the subsets, residual correlations between frames from the same video sequence cannot be completely excluded. The distribution of different weld types across the three subsets is presented in Table 1.

Table 1. Distribution of weld images and feature point instances in the dataset.

Dataset	V	Butt	Y	Lap	Fillet	Total	Proportion
Train	1018	990	883	876	903	4670	70%
Validation	218	212	189	188	194	1001	15%
Test	218	212	189	188	193	1000	15%
Total	1454	1414	1261	1252	1290	6671	100%

3. Methodology

To address the challenges of small weld keypoints, weak local textures, and the need for lightweight deployment in industrial scenarios, this study proposes a lightweight weld keypoint detection model based on YOLO11n, named Weld-FHG-YOLO. The proposed model does not modify the main Backbone or Detect Head of YOLO11n. Instead, it optimizes multi-scale feature fusion and scale transformation in the Neck. Specifically, FasterC3K2 is used to reduce redundant computation in the Neck feature fusion modules, whereas HWD-GSConv is introduced to enhance the preservation and fusion of low-frequency structural information and high-frequency details during downsampling. By combining these two structures, the model improves weld keypoint detection accuracy and localization stability while maintaining lightweight characteristics.

3.1. Network Architecture of Weld-FHG-YOLO

Weld-FHG-YOLO uses YOLO11n as the baseline network, and its overall architecture still consists of three main components: the Backbone, Neck, and Detect Head. The Backbone extracts multi-level semantic features, the Neck fuses and transfers features at different scales, and the Detect Head predicts the class and location of each weld keypoint based on the fused features. Considering that the discriminative information of weld keypoints is mainly concentrated around bending, fracture, and local mutation regions of structured-light stripes, this study focuses on the structural improvements on the Neck, which directly affects the quality of the input features provided to the Detect Head.

The improved Neck mainly involves two structural replacements. First, the original C3K2 modules in the Neck are replaced with FasterC3K2 to reduce repeated spatial convolution during multi-scale feature fusion. Second, some bottom-up stride convolutions in the Neck are replaced with HWD-GSConv, enabling feature maps to explicitly preserve low-frequency contour information and high-frequency edge details during scale transformation. This design avoids substantial modifications to the Backbone while targeting two key factors that are particularly important for weld keypoint detection: detail preservation and multi-scale feature fusion.

Specifically, FasterC3K2 serves as a lightweight feature fusion module by using partial convolution to reduce the parameter count and computational cost in the Neck. HWD-GSConv serves as a multi-frequency downsampling fusion module, where Haar wavelet decomposition and lightweight GSConv fusion are used to alleviate the loss of high-frequency information caused by conventional downsampling. These two modules are functionally complementary, enabling the Detect Head to receive multi-scale features that contain both semantic information and local details.

The two modules address distinct bottlenecks in weld keypoint localization. FasterC3K2 reduces redundant spatial computation during multi-scale feature fusion by applying partial convolution to only a subset of channels. HWD-GSConv alleviates the loss of localization-relevant details during scale transformation. Haar decomposition preserves low-frequency stripe structures and high-frequency edge details, while GSConv performs lightweight multi-frequency fusion. Their combination therefore balances model complex-

ity, inference efficiency, and localization accuracy through complementary mechanisms rather than simple structural stacking. The overall architecture is shown in Figure 3.

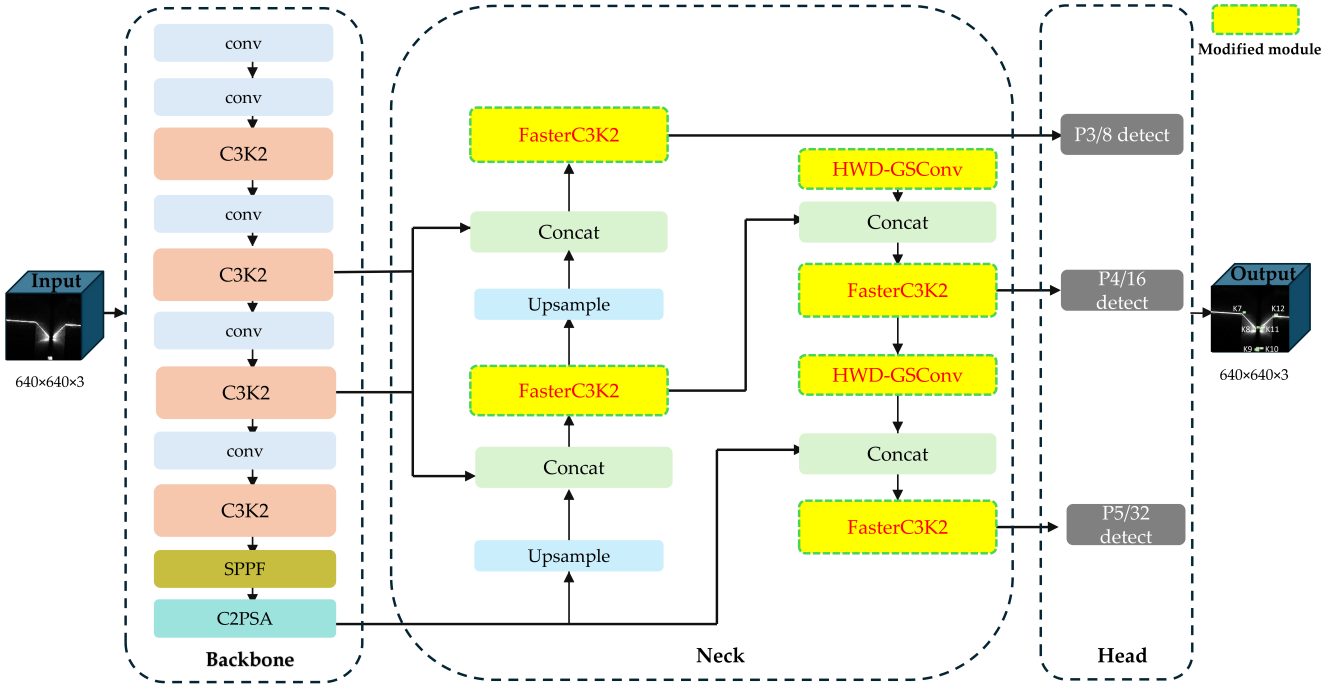


Figure 3. Overall architecture of the proposed Weld-FHG-YOLO network.

3.2. Baseline Network

This study uses YOLO11n as the baseline model for weld keypoint detection. YOLO-series models have advantages in end-to-end training and real-time inference [18,23], and YOLOv11n has recently been applied to lightweight scenarios such as industrial surface defect detection and agricultural robotic vision [32,33]. In YOLO11n, the Backbone extracts image features at different levels, the Neck fuses multi-scale features, and the Detect Head outputs object classes and bounding-box locations. Therefore, YOLO11n is suitable as a baseline network for weld keypoint detection under resource-constrained conditions.

Compared with conventional object detection tasks, weld keypoint detection relies more heavily on local details. Keypoints are usually located at bending, fracture, or local mutation regions of structured-light stripes, where the target regions are small and edge textures are relatively weak. These regions are also susceptible to interference from spatter, smoke, reflections, and high-intensity metallic surface highlights. Therefore, although directly using YOLO11n can provide good real-time performance, it may still suffer from insufficient detail representation, predicted-box center deviation, or missed detection of small targets in complex weld images.

The Neck is located between the Backbone and the Detect Head, and it directly affects the quality of multi-scale feature fusion and the representational capacity of the features delivered to the Detect Head. Multi-scale feature fusion structures, such as FPN, have demonstrated the importance of integrating cross-scale semantic and detail information for object detection [19]. Therefore, this study optimizes the Neck while keeping the main Backbone and Detect Head structures unchanged. This design preserves the feature extraction capability of the baseline network while more directly improving the multi-scale fusion and local detail preservation required for weld keypoint detection.

3.3. FasterC3K2 Lightweight Feature Fusion Module

In the YOLO11n Neck, C3K2 modules are used for multi-scale feature extraction and fusion. However, weld keypoints are spatially concentrated, and applying spatial convolution to every channel may introduce redundant computation. Inspired by the partial convolution strategy of FasterNet, FasterC3K2 divides the input channels into two subsets [36]. A 3×3 convolution is applied to one subset, while the remaining subset retains an identity path. The two branches are then concatenated and processed using a 1×1 convolution for channel interaction.

For weld keypoint localization, the identity branch preserves the original laser-stripe structure, while the spatially convolved branch extracts local geometric cues such as stripe edges, bends, intersections, and intensity mutations. The subsequent 1×1 convolution enables information interaction between the two branches. Consequently, FasterC3K2 reduces redundant spatial computation without discarding the local structural cues required to determine keypoint centers. FasterC3K2 replaces only the C3K2 modules in the Neck, while the Backbone remains unchanged. Its structure is illustrated in Figure 4.

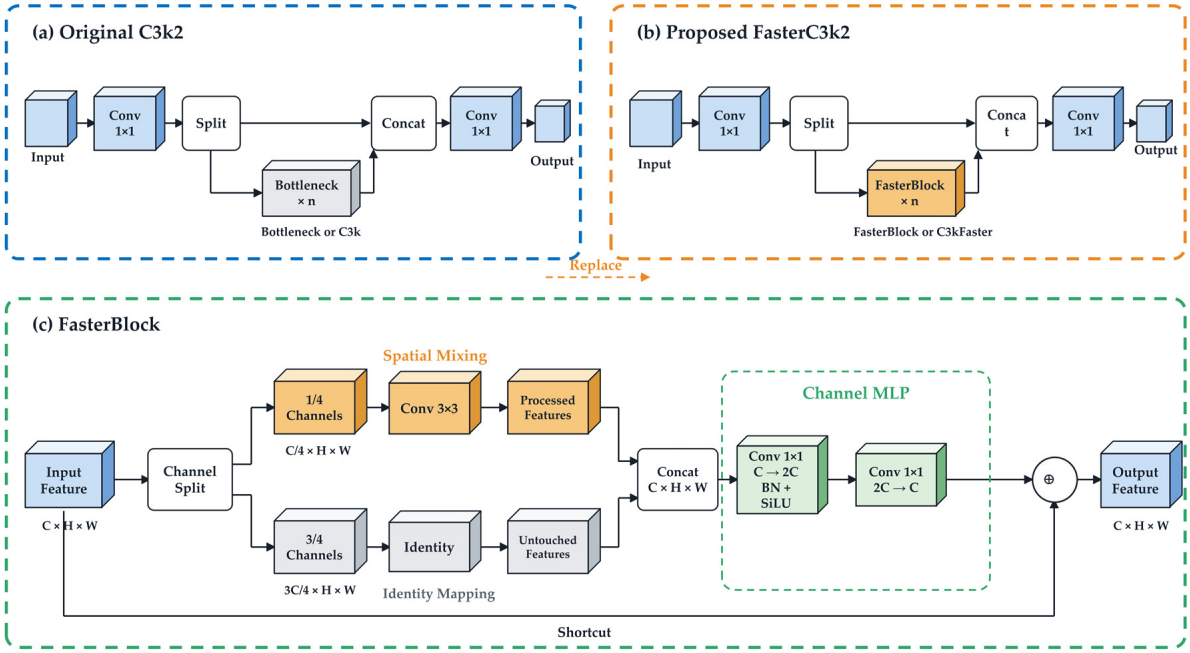


Figure 4. Structure of the FasterC3K2 lightweight feature fusion module. (**a**) Original C3k2; (**b**) Proposed FasterC3k2; (**c**) FasterBlock.

3.4. HWD-GSConv Downsampling Fusion Module

Weld keypoint localization depends strongly on stripe edges, bends, intersections, discontinuities, and local intensity mutations. Conventional stride convolution may weaken these localization-sensitive details during scale transformation. Therefore, HWD-GSConv replaces the 3×3 stride convolutions at the P3-to-P4 and P4-to-P5 transformations in the YOLO11n Neck. The module consists of Haar Wavelet Downsampling and GSConv, as shown in Figure 5.

HWD first applies fixed Haar wavelet filters to perform multi-frequency decomposition of the input feature. Haar Wavelet Downsampling can preserve structural and detailed information while reducing the spatial resolution [37]. Four two-dimensional Haar filters are used to obtain the LL, LH, HL, and HH sub-bands:

$$\begin{aligned} F_{LL} &= \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad F_{LH} = \frac{1}{2}\begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}, \\ F_{HL} &= \frac{1}{2}\begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}, \quad F_{HH} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \end{aligned} \tag{1}$$

F_{LL} is the low-frequency filter, which is mainly used to extract the overall structure and contour information of the feature map; F_{LH}, F_{HL} and F_{HH} are high-frequency filters, which extract edge and local variation information in different directions. These filters use fixed coefficients and are stored as buffers; they are not updated during backpropagation. Therefore, Haar wavelet decomposition does not introduce additional trainable parameters.

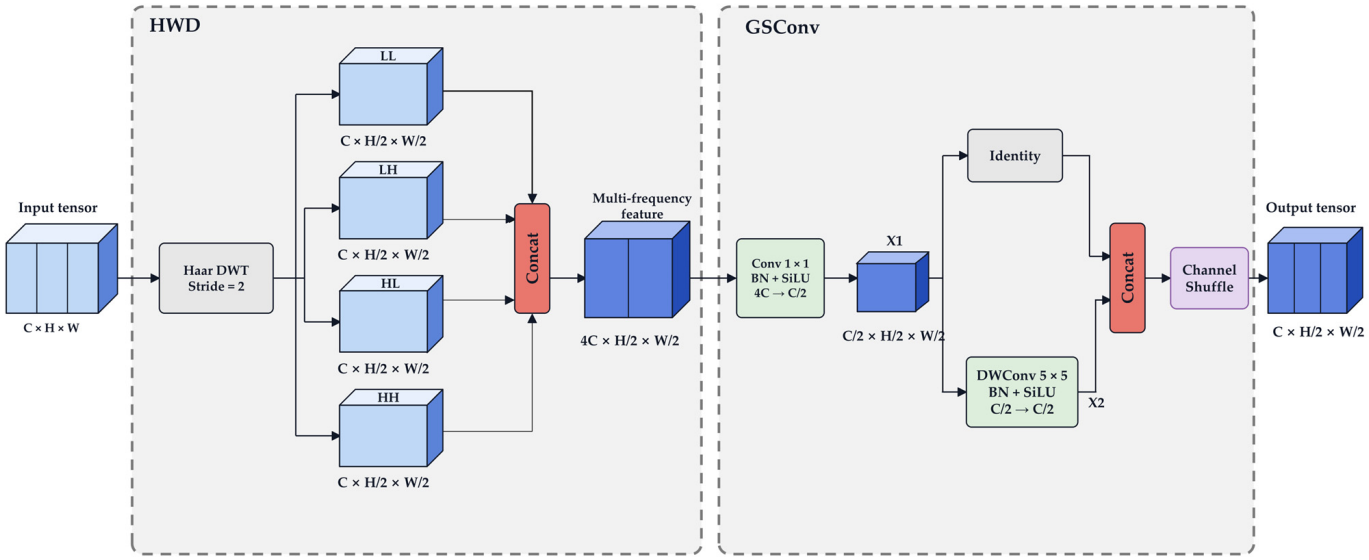


Figure 5. Architecture of the proposed HWD-GSConv downsampling fusion module.

Given an input feature $X \in R^{C\times H\times W}$, performs wavelet decomposition independently on each channel using grouped convolution, with the convolution stride set to 2. The four wavelet sub-bands are then concatenated along the channel dimension and can be expressed as follows:

$$X_w = Concat\left[(X * F_{LL})_{\downarrow 2}, (X * F_{LH})_{\downarrow 2}, (X * F_{HL})_{\downarrow 2}, (X * F_{HH})_{\downarrow 2}\right] \tag{2}$$

where $*$ denotes the convolution operation, and $\downarrow 2$ denotes spatial downsampling by a factor of 2. After wavelet decomposition and channel concatenation, the feature size changes from $C \times H \times W$ to $4C \times H/2 \times W/2$. Compared with directly using stride convolution for scale transformation, HWD explicitly maps low-frequency structural information and high-frequency details in different directions to the channel dimension while reducing spatial resolution, thereby providing a multi-frequency representation for subsequent feature fusion.

Since the number of output channels from HWD is expanded to four times that of the input, directly using standard convolution for channel compression and multi-frequency fusion would introduce a large number of parameters and high computational cost. Therefore, this study further introduces GSConv to perform lightweight fusion of the multi-frequency features generated by wavelet decomposition. GSConv was originally designed for lightweight Neck construction in real-time detectors and can achieve a favorable accuracy–speed trade-off between standard convolution and depthwise separable convolution [38]. First, a standard convolution compresses the number of channels from $4C$ to $C/2$. Then, depthwise convo-

lution is applied to the compressed features to supplement local spatial information with low computational cost. Finally, the outputs of the standard convolution branch and the depthwise convolution branch are concatenated, and Channel Shuffle is used to exchange information between different branches.

$$\begin{aligned} X_1 &= \delta[BN(Conv_{1\times1}(X_w))], \\ X_2 &= \delta[BN(DWConv_{5\times5}(X_1))], \\ Y &= Shuffle[Concat(X_1, X_2)], \end{aligned} \tag{3}$$

where $Conv_{1\times1}$ denotes standard pointwise convolution, $DWConv_{5\times5}$ denotes depthwise convolution, BN denotes batch normalization, $\delta(\cdot)$ denotes the SiLU activation function, and, $Shuffle(\cdot)$ denotes the channel shuffle operation. After GSConv fusion, the final output feature $\mathbf{Y}$ has a size $C \times H/2 \times W/2$, which is consistent with the output scale and channel number of the original stride convolution.

To analyze the lightweight characteristics of HWD-GSConv, the input and output channel numbers are assumed to be C, and the bias terms and batch normalization parameters are ignored. The main number of trainable parameters in the original 3×3 stride convolution is $9C^2$. In HWD, the Haar filters use fixed coefficients and contain no trainable parameters. Therefore, the trainable parameters of HWD-GSConv mainly come from the 1×1 standard convolution and the 5×5 depthwise convolution in GSConv. The main parameter counts of the two downsampling methods can thus be approximately expressed as follows:

$$P_{\text{Conv}} = 9C^2, \; P_{\text{HWD-GSConv}} \approx 2C^2 + \frac{25}{2}C. \tag{4}$$

According to Equation (4), as the number of channels increases, HWD-GSConv requires fewer trainable parameters than the original standard stride convolution. This module decomposes scale transformation into two stages: fixed Haar-based multi-frequency decomposition and lightweight GSConv fusion. The former extracts low-frequency structural information and high-frequency details, whereas the latter performs channel compression, local spatial modeling, and multi-frequency feature interaction. By combining these two stages, HWD-GSConv provides downsampled features that contain both structural semantics and local details for subsequent multi-scale fusion in the Neck, while keeping model complexity under control.

It should be noted that HWD-GSConv is not a completely lossless downsampling structure. Haar wavelet decomposition first explicitly preserves multi-frequency information in 4C channels, and GSConv then compresses these features to C channels. Therefore, the final output is still a learned compressed representation. In this study, HWD-GSConv is applied only to the two bottom-up scale transformation positions in the YOLO11n Neck, while the Backbone, the upsampling path of the Neck, and the Detect Head remain unchanged. This design preserves the main structure of the baseline network while specifically improving downsampling detail representation and lightweight feature fusion for weld keypoint detection.

4. Experiments and Analysis

4.1. Experimental Environment and Training Settings

To evaluate the effectiveness of Weld-FHG-YOLO for weld keypoint detection, experiments were conducted on the self-built weld keypoint dataset. To ensure comparability, all experiments were performed on the same computing platform. The hardware configuration and software environment are listed in Table 2.

All models were implemented using the Ultralytics YOLO framework. The input image size was uniformly set to 640×640. The number of training epochs was set to 200, with

a batch size of 16. The initial learning rate was 0.0001, and the learning rate decay factor was 0.001. AdamW was used as the optimizer. Mosaic data augmentation was disabled during training to avoid excessive disturbance to the spatial distribution of weld keypoints caused by strong augmentation. The number of workers was set to 4, and automatic mixed precision (AMP) training was enabled.

Table 2. Hardware configuration and software environment.

Experimental Environment	Configuration
CPU	AMD Ryzen 7 5800H
GPU	NVIDIA GeForce RTX 3060
Operating System	Windows 11
Python	3.10.20
PyTorch	2.5.1
Ultralytics	8.4.23
OpenCV	4.13.0
CUDA	12.1

The final performance of all models was evaluated on the same test set. The evaluation included detection accuracy, model complexity, CPU inference speed, and keypoint center localization error. CPU FPS was measured by sequential single-image inference. MCE was calculated as the Euclidean distance between the center of the matched predicted bounding box and the center of the corresponding ground-truth annotation.

4.2. Evaluation Metrics

This study comprehensively evaluated each model using Precision (P), Recall (R), mAP@0.5, mAP@0.5:0.95, Params, GFLOPs, CPU FPS, and MCE. Precision measures the proportion of correctly detected positive samples among all samples predicted as positive, and is calculated as follows:

$$P = \frac{TP}{TP + FP} \tag{5}$$

Recall measures the proportion of correctly detected targets among all ground-truth targets, and is calculated as follows:

$$R = \frac{TP}{TP + FN} \tag{6}$$

where TP denotes the number of correctly detected targets, FP denotes the number of false positive detections, and FN denotes the number of false negative detections. The mean average precision (mAP) is the mean value of AP across all categories and is used to comprehensively evaluate detection accuracy. It is calculated as follows:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{7}$$

where N denotes the number of categories, and AP_i denotes the average precision of the i-th category. In this study, mAP@0.5 and mAP@0.5:0.95 were used to evaluate detection performance. mAP@0.5 represents the mean average precision at an IoU threshold of 0.5, whereas mAP@0.5:0.95 represents the mean average precision averaged over IoU thresholds from 0.5 to 0.95 with a step size of 0.05.

To further evaluate the localization accuracy of weld keypoints, MCE was used to measure the mean Euclidean distance between the center of the predicted bounding box and the center of the corresponding ground-truth annotation. It is calculated as follows:

$$MCE = \frac{1}{M}\sum_{j=1}^{M}\sqrt{\left(x_j^p - x_j^g\right)^2 + \left(y_j^p - y_j^g\right)^2} \quad (8)$$

where M denotes the number of matched keypoint targets used for evaluation, $\left(x_j^p, y_j^p\right)$ denotes the center coordinates of the j-th predicted bounding box, and $\left(x_j^g, y_j^g\right)$ denotes the center coordinates of the corresponding ground-truth bounding box. A lower MCE indicates a smaller localization error for weld keypoints. Params and GFLOPs were used to evaluate model complexity, where lower values indicate a more lightweight model. To improve the engineering interpretability of localization errors, the pixel-level MCE was converted into an approximate physical distance using the calibrated line-structured-light imaging system. Pixel coordinates were mapped to the camera coordinate system using the camera intrinsic parameters and calibrated laser-plane equation. Under the fixed sensor-to-workpiece geometry, the local equivalent scale was approximately 0.05 mm/pixel [15]. Therefore, the physical localization error was calculated as:

$$\mathrm{MCE}_{mm} \approx 0.05 \times \mathrm{MCE}_{px}. \quad (9)$$

where MCE_{mm} and MCE_{px} denote the localization errors in millimeters and pixels, respectively.

CPU FPS was used to measure inference speed in a CPU environment and reflects the deployment potential of the model in industrial scenarios.

4.3. Comparison with the Baseline Model

The proposed Weld-FHG-YOLO was first compared with the baseline YOLO11n model. The test results are presented in Table 3.

Table 3. Comparison between YOLO11n and Weld-FHG-YOLO on the test set.

Model	Params (M)	GFLOPs	P	R	mAP@0.5	mAP@0.5:0.95	CPU FPS	MCE (px/mm)
YOLO11n	2.593	6.457	0.9898	0.9841	0.9938	0.7337	13.69	2.254/0.113
Weld-FHG-YOLO	2.301	6.016	0.9899	0.9910	0.9940	0.7901	14.77	2.131/0.107

As shown in Table 3, compared with YOLO11n, Weld-FHG-YOLO reduced the number of parameters from 2.593 M to 2.301 M, corresponding to a reduction of approximately 11.3%. GFLOPs decreased from 6.457 to 6.016, representing a reduction of approximately 6.8%. Meanwhile, mAP@0.5:0.95 increased from 0.7337 to 0.7901, with an improvement of 5.64 percentage points. These results indicate that the improved model achieves better overall detection performance under stricter localization thresholds while reducing model complexity.

In terms of keypoint localization error, the MCE of Weld-FHG-YOLO decreased from 2.254 px (0.113 mm) to 2.131 px (0.107 mm), indicating that the predicted bounding-box centers were closer to the ground-truth keypoint positions. In addition, CPU FPS increased from 13.69 to 14.77, suggesting that the proposed method maintains good CPU inference efficiency while improving detection accuracy and reducing localization error.

4.4. Ablation Experiments

To evaluate the contribution of each improved module to weld keypoint detection performance, several ablation experiments were designed using YOLO11n as the baseline model. All models were trained and evaluated using the same dataset split, training settings, and test set. The experimental results are shown in Table 4.

Table 4. Ablation results of different improved YOLO11n models on the test set.

Model	Params (M)	GFLOPs	P	R	mAP@0.5	mAP@0.5:0.95	CPU FPS	MCE (px/mm)
YOLO11n	2.593	6.457	0.9898	0.9841	0.9938	0.7337	13.69	2.254/0.113
YOLO11n + GSConv Neck	2.503	6.343	0.9912	0.9906	0.9943	0.7766	12.22	2.145/0.107
YOLO11n + HWD Neck	2.491	6.326	0.9932	0.9919	0.9944	0.7875	13.53	2.149/0.107
YOLO11n + FasterC3K2	2.442	6.196	0.9894	0.9900	0.9943	0.7747	15.11	2.161/0.108
YOLO11n + HWD-GSConv	2.452	6.277	0.9922	0.9905	0.9944	0.7844	14.64	2.154/0.108
YOLO11n + FasterC3K2 + HWD	2.340	6.065	0.9885	0.9865	0.9937	0.7787	14.97	2.164/0.108
Weld-FHG-YOLO	2.301	6.016	0.9899	0.9910	0.9940	0.7901	14.77	2.131/0.107

As shown in Table 4, the different improved modules had positive effects on model performance. After introducing GSConv Neck alone, the number of parameters decreased from 2.593 M to 2.503 M, GFLOPs decreased from 6.457 to 6.343, and mAP@0.5:0.95 increased from 0.7337 to 0.7766. These results indicate that GSConv can improve feature fusion in the Neck while reducing model complexity. When HWD Neck was introduced alone, mAP@0.5:0.95 further increased to 0.7875, and MCE decreased from 2.254 px to 2.149 px. This result suggests that Haar wavelet downsampling helps preserve detail information related to weld edges and keypoint locations.

When FasterC3K2 was used alone, the number of parameters decreased to 2.442 M, GFLOPs decreased to 6.196, and CPU FPS increased to 15.11, while mAP@0.5:0.95 reached 0.7747. These results indicate that FasterC3K2 effectively reduces redundant computation during Neck feature fusion and maintains good detection performance under a lightweight design. When HWD-GSConv was used alone, mAP@0.5:0.95 reached 0.7844, and MCE decreased to 2.154 px. This indicates that combining wavelet-based multi-frequency decomposition with lightweight GSConv fusion helps enhance detail representation in downsampled features. With the joint effect of FasterC3K2 and HWD-GSConv, the final Weld-FHG-YOLO model achieved the best overall performance.

To further visualize the performance trade-off among different ablation models, an accuracy–complexity trade-off plot was drawn, as shown in Figure 6. In this plot, the horizontal axis represents GFLOPs, the vertical axis represents mAP@0.5:0.95, and the marker size represents CPU FPS. Weld-FHG-YOLO is located in the upper-left region of the plot, indicating that it achieves the highest mAP@0.5:0.95 among all ablation models while maintaining the lowest computational complexity and relatively high inference speed. This result further confirms the effectiveness of the proposed improvement strategy.

To further examine the reliability of the reported improvements, repeated-run experiments were conducted for the YOLO11n baseline and the proposed Weld-FHG-YOLO. Each model was trained three times under the same dataset split and training settings, and the obtained weights were evaluated on the same test set. The mean and standard deviation of mAP@0.5:0.95, MCE, and CPU FPS are reported in Table 5. Here, MCE is reported in both pixels and millimeters according to the calibration relationship of 1 pixel = 0.05 mm.

As shown in Table 5, Weld-FHG-YOLO shows more stable overall performance than YOLO11n across the repeated runs. The average mAP@0.5:0.95 increases from 0.7296 to 0.7854, while the average MCE decreases from 2.243 px to 2.142 px, indicating improved detection accuracy and center localization stability. In addition, the relatively small standard deviations suggest that the performance gain is not dependent on a single training run. The average CPU FPS of Weld-FHG-YOLO is also slightly higher than that of YOLO11n, which further indicates that the proposed model improves localization accuracy while maintaining a modest advantage in CPU inference efficiency.

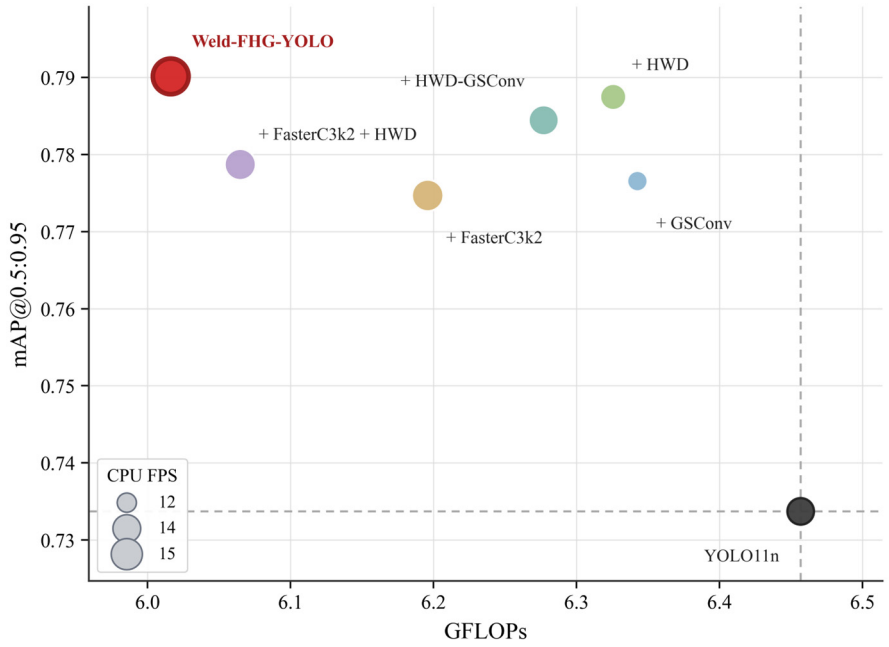


Figure 6. Accuracy–complexity trade-off analysis of different ablation models.

Table 5. Repeated-run stability analysis of YOLO11n and Weld-FHG-YOLO on the test set.

Model	mAP@0.5:0.95	MCE (px/mm)	CPU FPS
YOLO11n	0.7296 ± 0.0072	2.243 ± 0.018/0.112 ± 0.00090	13.40 ± 0.94
Weld-FHG-YOLO	0.7854 ± 0.0047	2.142 ± 0.009/0.107 ± 0.00045	14.45 ± 0.55

4.5. Comparison with Other Detection Models

To further evaluate the overall performance of Weld-FHG-YOLO, it was compared with YOLOv10n, YOLO11n, YOLOv12n, StripePoint-YOLO, and SSD300-VGG16. All models were trained and evaluated on the same training, validation, and test sets. A unified test script was used to calculate Precision, Recall, mAP, CPU FPS, and MCE. The results are presented in Table 6.

Table 6. Comparison with other detection models on the test set.

Model	Params (M)	GFLOPs	P	R	mAP@0.5	mAP@0.5:0.95	CPU FPS	MCE (px /mm)
SSD300-VGG16	25.751	31.174	0.9620	0.9399	0.9780	0.7422	4.49	2.597/0.130
StripePoint-YOLO	1.864	8.811	0.9838	0.9772	0.9926	0.6885	10.42	2.357/0.118
YOLOv10n	2.713	8.425	0.9768	0.9774	0.9936	0.7745	14.67	2.089/0.104
YOLO11n	2.593	6.457	0.9898	0.9841	0.9938	0.7337	13.69	2.254/0.113
YOLOv12n	2.571	6.496	0.9929	0.9905	0.9943	0.7325	12.02	2.170/0.109
Weld-FHG-YOLO	2.301	6.016	0.9899	0.9910	0.9940	0.7901	14.77	2.131/0.107

To preserve the standard architecture and official pretrained-weight configuration of each comparison model, SSD300-VGG16 used its native input resolution of 300 × 300, whereas the YOLO-series models used an input resolution of 640 × 640. It should be noted that the lower input resolution may weaken the ability of SSD300 to represent small weld keypoints. Therefore, this experiment compares the overall performance of different detection frameworks under their respective standard configurations, rather than strictly isolating the effect of input resolution on detection performance.

To examine the convergence behavior of different models during training, the curves of Precision, Recall, mAP@0.5, and mAP@0.5:0.95 over the training epochs were further plotted, as shown in Figures 7 and 8.

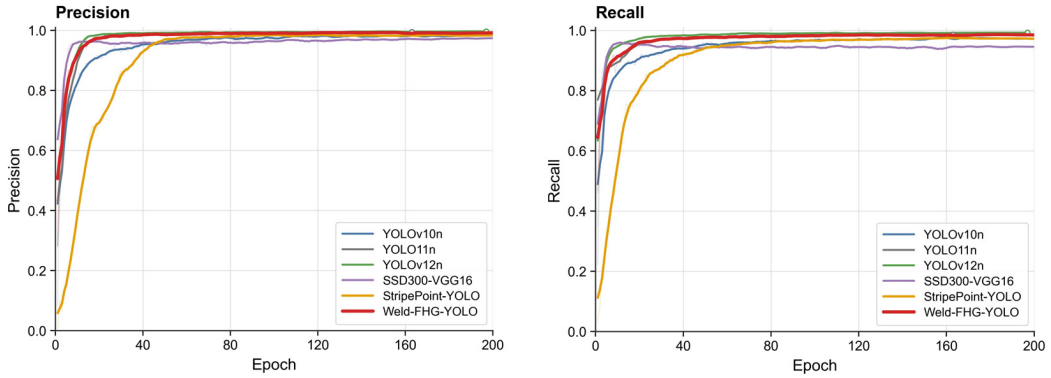


Figure 7. Precision and recall curves of different detection models on the validation set. (**a**) Precision. (**b**) Recall.

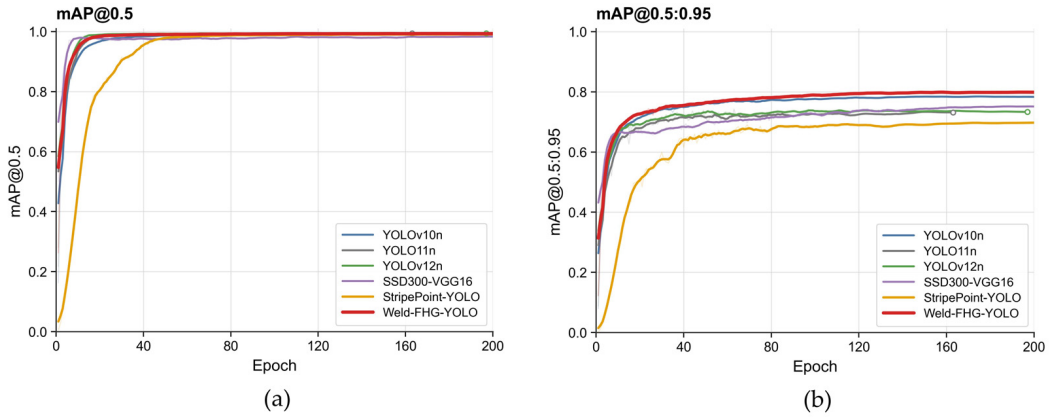


Figure 8. mAP curves of different detection models on the validation set. (**a**) mAP@0.5; (**b**) mAP@0.5:0.95.

As shown in Table 6, Weld-FHG-YOLO achieved an mAP@0.5:0.95 of 0.7901, which was the highest among all comparison models. It exceeded YOLOv10n, YOLO11n, YOLOv12n, StripePoint-YOLO, and SSD300-VGG16 by 1.56, 5.64, 5.76, 10.16, and 4.79 percentage points, respectively. Meanwhile, Weld-FHG-YOLO contained 2.301 M parameters and required 6.016 GFLOPs. Its GFLOPs were the lowest among the compared models, while its parameter count was the second lowest after StripePoint-YOLO, while its CPU FPS reached 14.77. In terms of localization error, Weld-FHG-YOLO achieved an MCE of 2.131 px, which was lower than those of YOLO11n, YOLOv12n, StripePoint-YOLO, and SSD300-VGG16, but slightly higher than that of YOLOv10n (2.089 px). Overall, the proposed model achieved a better balance among detection accuracy, model complexity, and CPU inference efficiency.

4.6. Visualization and Error Analysis

To analyze the keypoint localization capability of the model from both the overall error distribution and representative detection results, the center localization error distributions of different weld types on the complete test set were first calculated, as shown in Figure 9.

In Figure 9a–e represent V-groove, butt, Y-groove, lap, and fillet welds, respectively. Each scatter point represents the center localization error of a matched keypoint. The box

indicates the interquartile range of the error distribution, while the horizontal line and diamond inside the box represent the median and mean values, respectively.

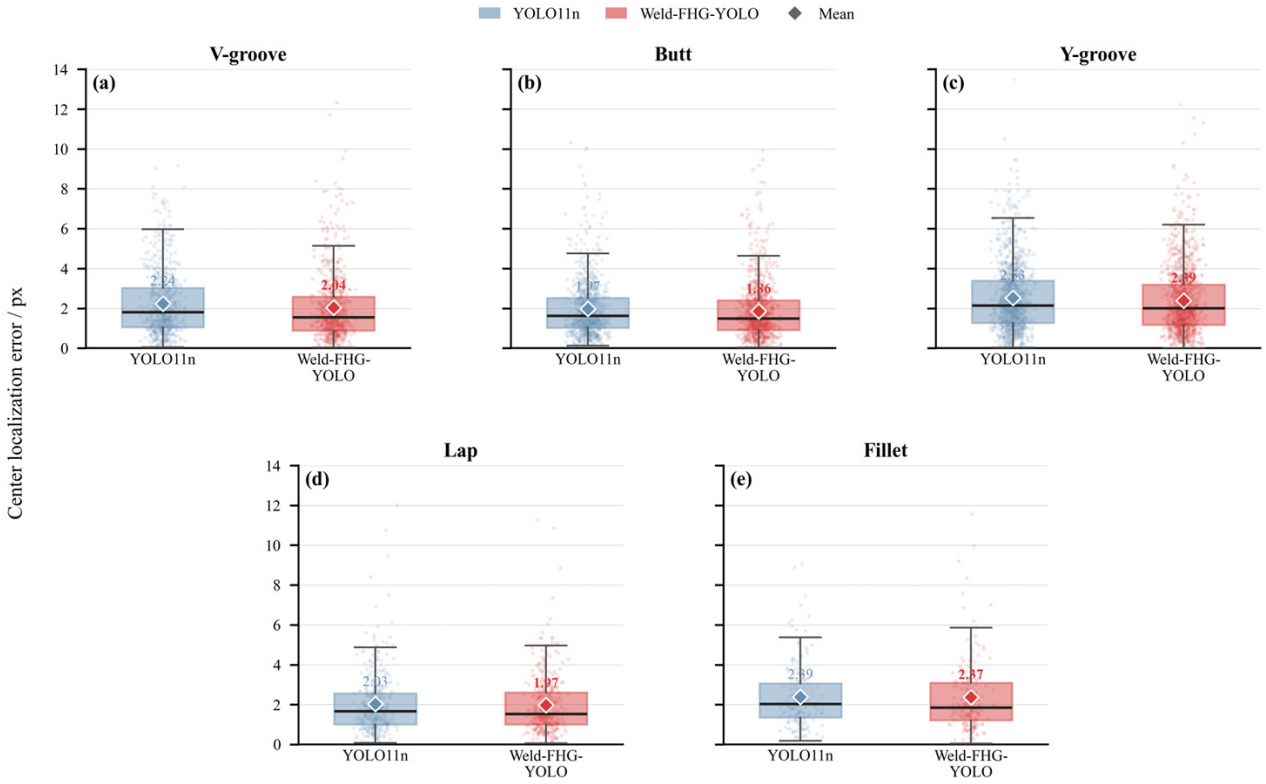


Figure 9. Distribution of center localization errors for five weld types on the test set. (**a**) V-groove; (**b**) butt; (**c**) Y-groove; (**d**) lap; (**e**) fillet weld.

As shown in Figure 9, Weld-FHG-YOLO achieved lower mean center localization errors than YOLO11n for all five weld types. For V-groove, butt, Y-groove, lap, and fillet welds, the mean errors decreased from 2.24, 1.97, 2.53, 2.03, and 2.39 px to 2.04, 1.86, 2.39, 1.97, and 2.37 px, respectively. Among them, the reduction was most evident for V-groove welds, indicating that the improved model can more accurately localize multiple keypoints near groove edges. The overall localization error for Y-groove welds was higher than that for the other weld types, which may be related to the larger number of keypoints and more complex local geometry.

To further validate the localization superiority and environmental robustness of the proposed method, qualitative comparisons and hard-sample visualizations were conducted under various complex industrial conditions.

Figure 10 illustrates the qualitative comparison of keypoint localization results between the baseline YOLO11n and the proposed Weld-FHG-YOLO across five representative scenarios. In Figure 10, the first row shows the localization results of YOLO11n, while the second row shows those of Weld-FHG-YOLO. Green circles and red crosses indicate the ground-truth keypoint centers and predicted keypoint centers, respectively, the values next to the matched keypoints denote the center localization errors between the predicted and ground-truth centers. For the selected V-groove, butt, Y-groove, lap, and fillet weld samples, the average image-level MCE decreased from 3.62, 2.97, 3.41, 2.55, and 3.76 px (corresponding to 0.181, 0.149, 0.171, 0.128, and 0.188 mm) for YOLO11n to 1.65, 2.14, 2.50, 1.04, and 1.31 px (0.083, 0.107, 0.125, 0.052, and 0.066 mm) for Weld-FHG-YOLO, respectively.

To explicitly demonstrate the boundary performance under extreme industrial disturbances, Figure 11 presents the visualization results and quantitative metrics on challenging samples suffering from severe arc glare, intense spatter, and degraded contrast.

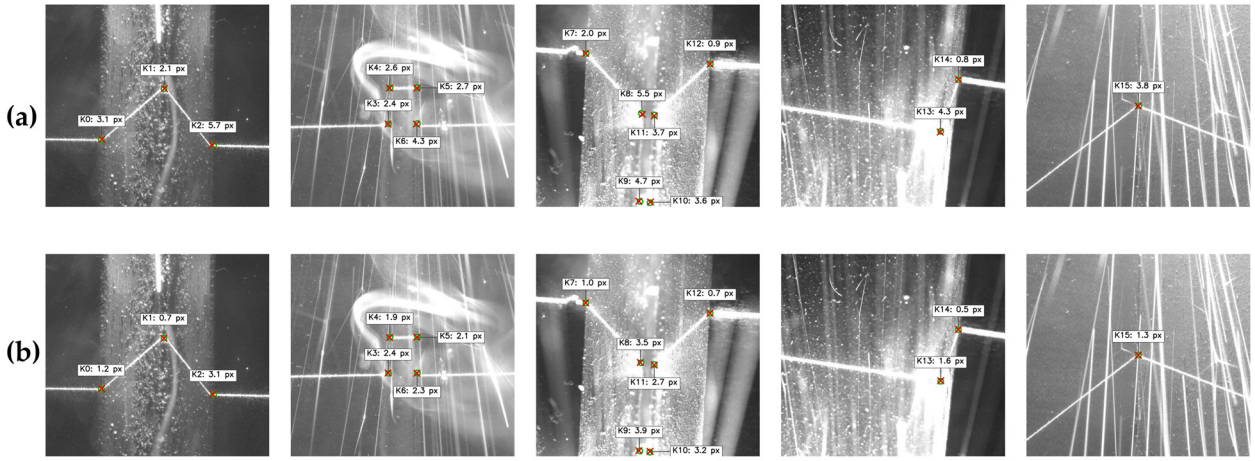


Figure 10. Typical comparison of keypoint localization results. (**a**) YOLO11n; (**b**) Weld-FHG-YOLO.

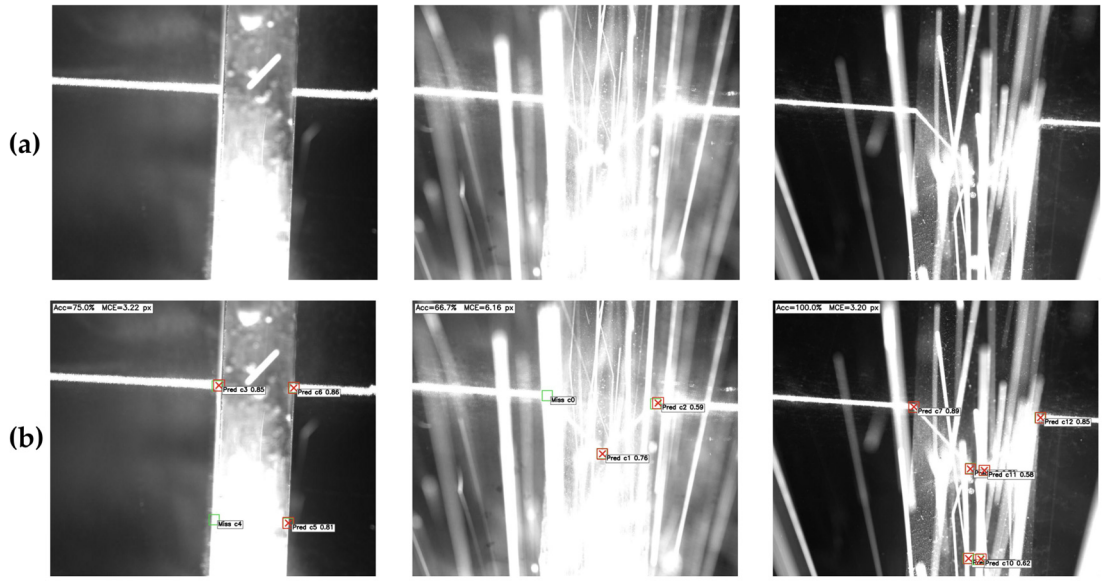


Figure 11. Visualization results of keypoint detection on challenging samples under severe industrial disturbances. (**a**) Corresponding original input images. (**b**) Keypoint detection, confidence scores, and corresponding MCE metrics for representative weld seam scenarios (e.g., strong arc glare and intense spatter). Green boxes denote ground-truth annotations, red boxes denote network predictions (with red crosses indicating keypoint centers), and green labels indicate missed detections (Miss). The callout boxes in the upper-left corner of each image display the overall accuracy (Acc) and Mean Center Error (MCE).

As observed in Figure 11b, even when image features are heavily corrupted by intense arc glare or dense spatter, Weld-FHG-YOLO maintains strong localization stability. Although minor missing detections (Miss) occasionally occur under extreme local saturation (such as Miss c0 in the second scenario), the network successfully captures the remaining critical keypoints with reasonable center deviations. These results indicate that the proposed method maintains a degree of localization robustness under the evaluated noise conditions.

Taken together, the visualization results show that Weld-FHG-YOLO achieved lower center localization errors across all five weld types. The error distributions on the complete test set reflect the overall improvement in localization performance, whereas the representative difficult samples further demonstrate the model's localization behavior under complex geometric structures and visual interference. In particular, under multi-keypoint

groove structures and conditions involving reflection, overexposure, and stray laser stripes, the predicted centers of the improved model were generally closer to the ground-truth annotations. This indicates that Weld-FHG-YOLO provides better keypoint localization accuracy and stronger adaptability to complex weld imaging scenarios.

5. Discussion

The experimental results indicate that optimizing the Neck structure can effectively improve weld keypoint localization. Unlike conventional general object detection scenarios, weld feature points in line-structured light images are usually small, locally concentrated, and easily affected by weak texture, overexposure, surface reflection, and stray laser stripes. Therefore, the quality of multi-scale feature fusion directly affects the localization accuracy of weld feature points.

The effectiveness of FasterC3K2 mainly comes from its reduction of redundant spatial computation. In the ablation experiment, introducing FasterC3K2 alone reduced the number of parameters and GFLOPs while maintaining good detection performance, indicating that this module achieves a reasonable balance between lightweight design and feature representation.

The effectiveness of HWD-GSConv mainly comes from its ability to alleviate detail loss during downsampling. Conventional stride convolution reduces the spatial resolution of feature maps, but it may also smooth or weaken high-frequency information such as weld-stripe edges, bends, and local mutations. These high-frequency cues are important for keypoint localization.

The combination of the two modules is not a simple stacking of structures, but a targeted response to two different bottlenecks in the Neck. FasterC3K2 reduces redundant computation during multi-scale feature fusion, whereas HWD-GSConv preserves multi-frequency details during downsampling. The former contributes to lightweight feature fusion, while the latter contributes to detail preservation and feature stability. In the ablation study, each module improved model performance when used individually, and their combination achieved better overall performance. This indicates that the two improvement directions are complementary.

The improvement in mAP@0.5:0.95 was more pronounced than the reduction in MCE. Compared with YOLO11n, mAP@0.5:0.95 increased from 0.7337 to 0.7901, whereas MCE decreased from 2.254 to 2.131 px. The MCE reduction was 0.123 px, corresponding to approximately 5.46% or 0.006 mm. This difference arises because the two metrics evaluate different aspects of model performance. mAP evaluates detection quality across multiple IoU thresholds, whereas MCE measures residual center displacement only between matched predictions and ground-truth targets. Since YOLO11n already achieved a relatively low center error, the remaining margin for MCE reduction was limited. Therefore, the MCE result represents modest refinement of localization rather than a large physical improvement.

From the perspective of intelligent manufacturing, accurate weld keypoint localization can provide geometric references for robotic weld tracking, trajectory correction, and path planning. The proposed Weld-FHG-YOLO focuses on the visual perception front end rather than a complete robotic welding control system. Compared with the YOLO11n baseline, the model improves strict localization accuracy and reduces model complexity, which is beneficial for weld visual perception under resource-constrained industrial conditions. The visualization results further show that Weld-FHG-YOLO produces more stable keypoint localization in challenging cases with reflection, overexposure, noise, and stray laser stripes. However, occasional missed detections still occurred under severe local saturation, indicating that extreme feature corruption remains challenging. These characteristics indicate its potential value for robotic welding perception and intelligent manufacturing applications.

It should also be noted that the model with only FasterC3K2 achieved a slightly higher CPU inference frame rate than the final complete model. This is mainly because HWD-GSConv introduces wavelet decomposition and multi-channel feature fusion operations, which bring a small amount of additional computational cost. Therefore, Weld-FHG-YOLO does not simply pursue the highest inference speed. Instead, it aims to balance localization accuracy, localization stability, parameter count, and computational cost, thereby achieving better overall performance. The final model achieved a CPU inference speed of 14.77 FPS, corresponding to approximately 67.7 ms per image. However, this measurement covers only model inference and excludes image acquisition, data communication, robot control, and trajectory-update latency. Therefore, the present results indicate potential for CPU-based front-end weld perception but do not establish closed-loop real-time welding performance.

Although the proposed method shows promising performance, several aspects require further investigation. First, the present dataset was partitioned at the frame level after manually removing duplicate and highly similar images. Although this preprocessing reduced direct overlap among the subsets, residual correlations between frames originating from the same video sequence cannot be completely excluded. Future work will adopt sequence-disjoint partitioning across independently recorded videos to provide a stricter evaluation of model generalization. Second, the manually annotated 40×40-pixel bounding boxes may contain residual center-position uncertainty, particularly under severe reflection, spatter, or local saturation. Such residual deviations may act as label noise during training and potentially weaken localization stability under severe visual interference. They may also affect the reported MCE because the annotated centers are used as reference positions. Since all models were trained and evaluated using the same annotation protocol, the comparison remains methodologically consistent; nevertheless, small differences in MCE should be interpreted with appropriate caution. Repeated annotation or multi-annotator consistency analysis should be used to quantify this uncertainty in future work. Third, the experiments were conducted on a self-built line-structured-light weld image dataset. The generalization ability under different sensors, welding materials, and manufacturing conditions should therefore be further validated. Finally, this study does not include closed-loop robotic welding experiments. Future work will integrate the proposed model into an actual robotic welding platform and evaluate its effects on trajectory tracking accuracy, processing latency, and welding quality.

6. Conclusions

This paper proposes Weld-FHG-YOLO, a lightweight multi-frequency feature fusion network for weld keypoint localization in line-structured light images. The model is built on the compact YOLO11n framework, while the main Backbone and Detect Head are kept unchanged. To improve the balance between localization accuracy and model lightweightness, the Neck is optimized using two complementary modules. FasterC3K2 reduces redundant computation during multi-scale feature fusion, and HWD-GSConv enhances multi-frequency detail preservation during downsampling.

Experimental results on a 16-class weld keypoint dataset show that Weld-FHG-YOLO achieves better overall performance than the YOLO11n baseline. The number of parameters decreased from 2.593 M to 2.301 M, and GFLOPs decreased from 6.457 to 6.016. Meanwhile, mAP@0.5:0.95 increased from 0.7337 to 0.7901, MCE decreased from 2.254 px to 2.131 px, and CPU FPS increased from 13.69 to 14.77. The ablation study, comparisons with other detection models, and visual localization-error analysis further support the effectiveness of the proposed method on the present dataset in lightweight feature fusion and keypoint localization stability.

https://doi.org/10.3390/machines14091022

Overall, Weld-FHG-YOLO achieves a favorable balance among strict localization accuracy, model complexity, and CPU inference efficiency. It provides a lightweight visual perception approach with potential application in robotic weld tracking, welding path planning, and resource-constrained industrial visual detection. However, because the present evaluation is limited to a self-built dataset and does not include closed-loop robotic welding experiments, its generalizability and end-to-end applicability remain to be validated. Future work will expand the dataset under more diverse welding conditions and evaluate the proposed method on embedded platforms and robotic welding systems to further improve its robustness and real-time applicability in complex manufacturing environments.

Author Contributions: Conceptualization, Y.Y. and C.C.; methodology, Y.Y.; software, Y.Y.; validation, Y.Y., X.M., W.L. and H.W.; formal analysis, Y.Y., W.L. and H.W.; investigation, Y.Y., X.M., W.L. and H.W.; resources, C.C. and H.L.; data curation, Y.Y.; writing—original draft preparation, Y.Y.; writing—review and editing, C.C., H.L. and F.D.; visualization, Y.Y.; supervision, C.C.; project administration, C.C.; funding acquisition, C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This paper is supported by the Shandong Provincial Natural Science Foundation (Grant No. ZR2024QE326).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: The authors would like to thank all the participants who contributed to this research.

Conflicts of Interest: The author Xiaoning Meng was employed by the company Jinan Heavy Industries Group Co., Ltd., Jinan 250109, China; the authors Wei Liu and Hougao Wang were employed by the company Shandong Nuclear Power Equipment Manufacturing Co., Ltd., Yantai 265100, China; the author Chao Chen was employed by the company Shantui Construction Machinery Co., Ltd., Jining 272073, China. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Yang, L.; Liu, Y.; Peng, J. Advances techniques of the structured light sensing in intelligent welding robots: A review. *Int. J. Adv. Manuf. Technol.* **2020**, *110*, 1027–1046. [CrossRef]
2. Li, G.; Hong, Y.; Gao, J.; Hong, B.; Li, X. Welding seam trajectory recognition for automated skip welding guidance of a spatially intermittent welding seam based on laser vision sensor. *Sensors* **2020**, *20*, 3657. [CrossRef] [PubMed]
3. Zou, Y.; Wang, Y.; Zhou, W.; Chen, X. Real-time seam tracking control system based on line laser visions. *Opt. Laser Technol.* **2018**, *103*, 182–192. [CrossRef]
4. Ma, Y.; Fan, J.; Deng, S.; Luo, Y.; Ma, X.; Jing, F.; Tan, M. Efficient and accurate start point guiding and seam tracking method for curve weld based on structure light. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–10. [CrossRef]
5. He, Y.; Xu, Y.; Chen, Y.; Chen, H.; Chen, S. Weld seam profile detection and feature point extraction for multi-pass route planning based on visual attention model. *Robot. Comput.-Integr. Manuf.* **2016**, *37*, 251–261. [CrossRef]
6. Yang, L.; Li, E.; Long, T.; Fan, J.; Liang, Z. A high-speed seam extraction method based on the novel structured-light sensor for arc welding robot. *IEEE Sens. J.* **2018**, *18*, 8631–8641. [CrossRef]
7. Zeng, J.; Cao, G.; Li, W.; Sun, J.; Lv, J.; Chen, B. Feature point extraction based on contour detection and corner detection for weld seam. *J. Phys. Conf. Ser.* **2018**, *1074*, 012161. [CrossRef]
8. Xiao, R.; Xu, Y.; Hou, Z.; Chen, C.; Chen, S. An adaptive feature extraction algorithm for multiple typical seam tracking based on vision sensor in robotic arc welding. *Sens. Actuators A Phys.* **2019**, *297*, 111533. [CrossRef]
9. Fan, J.; Deng, S.; Ma, Y.; Zhou, C.; Jing, F.; Tan, M. Seam feature point acquisition based on efficient convolution operator and particle filter in GMAW. *IEEE Trans. Ind. Inform.* **2021**, *17*, 1220–1230. [CrossRef]
10. Du, R.; Xu, Y.; Hou, Z.; Shu, J.; Chen, S. Strong noise image processing for vision-based seam tracking in robotic gas metal arc welding. *Int. J. Adv. Manuf. Technol.* **2019**, *101*, 2135–2149. [CrossRef]

11. Yang, L.; Fan, J.; Huo, B.; Li, E.; Liu, Y. Image denoising of seam images with deep learning for laser vision seam tracking. *IEEE Sens. J.* **2022**, *22*, 6098–6107. [CrossRef]
12. Lu, J.; Yang, A.; Chen, X.; Xu, X.; Lv, R.; Zhao, Z. A seam tracking method based on an image segmentation deep convolutional neural network. *Metals* **2022**, *12*, 1365. [CrossRef]
13. Gao, A.; Fan, Z.; Li, A.; Le, Q.; Wu, D.; Du, F. YOLO-Weld: A Modified YOLOv5-Based Weld Feature Detection Network for Extreme Weld Noise. *Sensors* **2023**, *23*, 5640. [CrossRef] [PubMed]
14. Gao, A.; Li, A.; Su, F.; Yang, X.; Liu, W.; Du, F.; Chen, C. WeldLight: A Lightweight Weld Classification and Feature Point Extraction Model for Weld Seam Tracking. *Sensors* **2025**, *25*, 5761. [CrossRef] [PubMed]
15. Yang, M.; Li, S.; Sun, X.; Wang, H.; Gao, A.; Du, F.; Chen, C. StripePoint-YOLO: Task-Adaptive Detection of Multi-Type Weld Seam Keypoints in Noisy Industrial Welding Scenes. *Sensors* **2026**, *26*, 4980. [CrossRef] [PubMed]
16. Li, M.; Huang, J.; Xue, L.; Zhang, R. A guidance system for robotic welding based on an improved YOLOv5 algorithm with a RealSense depth camera. *Sci. Rep.* **2023**, *13*, 21299. [CrossRef] [PubMed]
17. Zhu, S.; Zhang, L.; Zhao, J.; Hu, X.; Luo, X. Research on steel structure weld seam recognition algorithm based on improved YOLOv5. *Sci. Rep.* **2025**, *15*, 34276. [CrossRef] [PubMed]
18. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016*; IEEE: New York, NY, USA, 2016; pp. 779–788.
19. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017*; IEEE: New York, NY, USA, 2017; pp. 936–944.
20. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single shot multibox detector. In *Computer Vision—ECCV 2016*; Springer: Cham, Switzerland, 2016; pp. 21–37.
21. Zhou, X.; Wang, D.; Krahenbuhl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
22. Girshick, R. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015*; IEEE: New York, NY, USA, 2015; pp. 1440–1448.
23. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
24. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Computer Vision—ECCV 2020*; Springer: Cham, Switzerland, 2020; pp. 213–229.
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [CrossRef]
26. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
27. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted residuals and linear bottlenecks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: New York, NY, USA, 2018; pp. 4510–4520.
28. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: New York, NY, USA, 2018; pp. 6848–6856.
29. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In *Computer Vision—ECCV 2018*; Springer: Cham, Switzerland, 2018; pp. 116–131.
30. Liu, L.; Xu, Z. Optimizing lightweight neural networks for efficient mobile edge computing. *Sci. Rep.* **2025**, *15*, 22056. [CrossRef] [PubMed]
31. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021*; IEEE: New York, NY, USA, 2021; pp. 13713–13722.
32. Ma, C.; Chen, J.; Pan, Y. DSC-YOLOv11: An efficient method for detecting industrial surface defects. *Measurement* **2026**, *280*, 121910. [CrossRef]
33. Xu, L.; Li, H.; Tang, Z.; He, Y.; Yu, Y.; Zi, L.; Jiang, H.; Wang, Y. A multi-ripeness citrus detection model based on OR-YOLOv11. *Smart Agric. Technol.* **2026**, *14*, 102254. [CrossRef]
34. Cheng, J.; Jin, H.; Qian, X. Real-time weld seam feature extraction in construction sites. *Autom. Constr.* **2024**, *160*, 105330. [CrossRef]
35. Mobaraki, M.; Ahani, S.; Gonzalez, R.; Yi, K.M.; Van Heusden, K.; Dumont, G.A. Vision-based seam tracking for GMAW fillet welding based on keypoint detection deep learning model. *J. Manuf. Process.* **2024**, *117*, 315–328. [CrossRef]
36. Chen, J.; Kao, S.-H.; He, H.; Zhuo, W.; Wen, S.; Lee, C.-H.; Chan, S.-H.G. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023*; IEEE: New York, NY, USA, 2023; pp. 12021–12031.

37. Xu, G.; Liao, W.; Zhang, X.; Li, C.; He, X.; Wu, X. Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. *Pattern Recognit.* **2023**, *143*, 109819. [CrossRef]
38. Li, H.; Li, J.; Wei, H.; Liu, Z.; Zhan, Z.; Ren, Q. Slim-neck by GSConv: A lightweight-design for real-time detector architectures. *J. Real-Time Image Process.* **2024**, *21*, 62. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.