sensors

MDPI

Article

WeldLight: A Lightweight Weld Classification and Feature Point Extraction Model for Weld Seam Tracking

Ang Gao [1,2,3], Anning Li [1,2,3], Fukang Su [1,2,3], Xinqi Yang [1,2,3], Wenping Liu [1,2,3,*], Fuxin Du [1,2,3] and Chao Chen [1,4]

[1] School of Mechanical Engineering, Shandong University, Jinan 250061, China; 202414363@mail.sdu.edu.cn (A.G.); 202214806@mail.sdu.edu.cn (A.L.); 202334436@mail.sdu.edu.cn (F.S.); 202314368@mail.sdu.edu.cn (X.Y.); dufuxin@sdu.edu.cn (F.D.); chaochen@sdjtu.edu.cn (C.C.)
[2] Key Laboratory of High-Efficiency and Clean Mechanical Manufacture, Ministry of Education, Shandong University, Jinan 250061, China
[3] State Key Laboratory of Advanced Equipment and Technology for Metal Forming, Beijing 100083, China
[4] School of Rail Transportation, Shandong Jiaotong University, Jinan 250357, China
* Correspondence: liuwp@sdu.edu.cn

Abstract

To address the issues of intense image noise interference and computational intensity faced by traditional vision-based weld tracking systems, we propose WeldLight, a lightweight and noise-resistant convolutional neural network for precise classification and positioning of welding seam feature points using single-line structured light vision. Our approach includes (1) an online data augmentation method to enhance training samples and improve noise adaptability; (2) a one-stage lightweight network for simultaneous positioning and classification; and (3) an attention module to filter features corrupted by intense noise, thereby improving stability. Experiments show that WeldLight achieves an F1-score of 0.9668 for seam classification on an adjusted test set, with mean absolute positioning errors of 1.639 pixels and 1.736 pixels on low-noise and high-noise test sets, respectively. With an inference time of 29.32 ms on a CPU platform, it meets real-time seam tracking requirements.

Keywords: feature point extraction; laser vision sensor; lightweight network; seam tracking; welding seam classification

check for updates

Academic Editor: Qiwu Luo

Received: 30 July 2025
Revised: 5 September 2025
Accepted: 6 September 2025
Published: 16 September 2025

Citation: 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. https://doi.org/10.3390/s25185761

Copyright: © 2025 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 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the wide application of robots in industrial manufacturing, a welding method based on teaching and reproduction has emerged. Traditional robot welding relies on workers to manually define an a priori torch position trajectory, which is inefficient and easily affected by some time-varying dimensional errors of workpieces, such as thermal deformation [1,2]. To track the welding seam in real time and realize the automatic adjustment of the welding robots, the researchers designed some advanced sensors to obtain the spatial position information and category information of the welding seam. With the rapid development of machine vision, structured light vision has found extensive application in intelligent robots [3,4] due to its high precision advantages [5] and the rich characteristic information it provides regarding welding processes [6].

As shown in Figure 1a,c, the characteristics of single-line structured light projection stripes vary significantly based on different welding seams. This variability is valuable for acquiring position information related to welding seam feature points and facilitates the

convenient classification of stripes. Such classification is crucial for adjusting the welding process parameters effectively. However, as shown in Figure 1b, common noises in the welding process, including splash, smog, and reflection light of arc, will be confused with stripes collected by industrial cameras. Several morphology-based methods have been proposed to realize the extraction of welding seam features. Li et al. [7] used a Kalman filter to track the meaningful laser stripes on the image through a window to reduce the influence of image noise during welding, and the laser stripes are decomposed into line–junction–line combination fragments to obtain the weld type and the position of feature points on the image. Yang et al. [8] applied a kernelized correlation filters' algorithm to realize seam tracking with feature point marks, which can adapt to different types of weld seams. Two common challenges exist in morphology-based research. One is that the tasks of feature point (or region) extraction and weld classification are usually two-stage. On the flip side, while these models possess a degree of resistance to noise, their tracking accuracy degrades under sustained high-noise conditions [9].

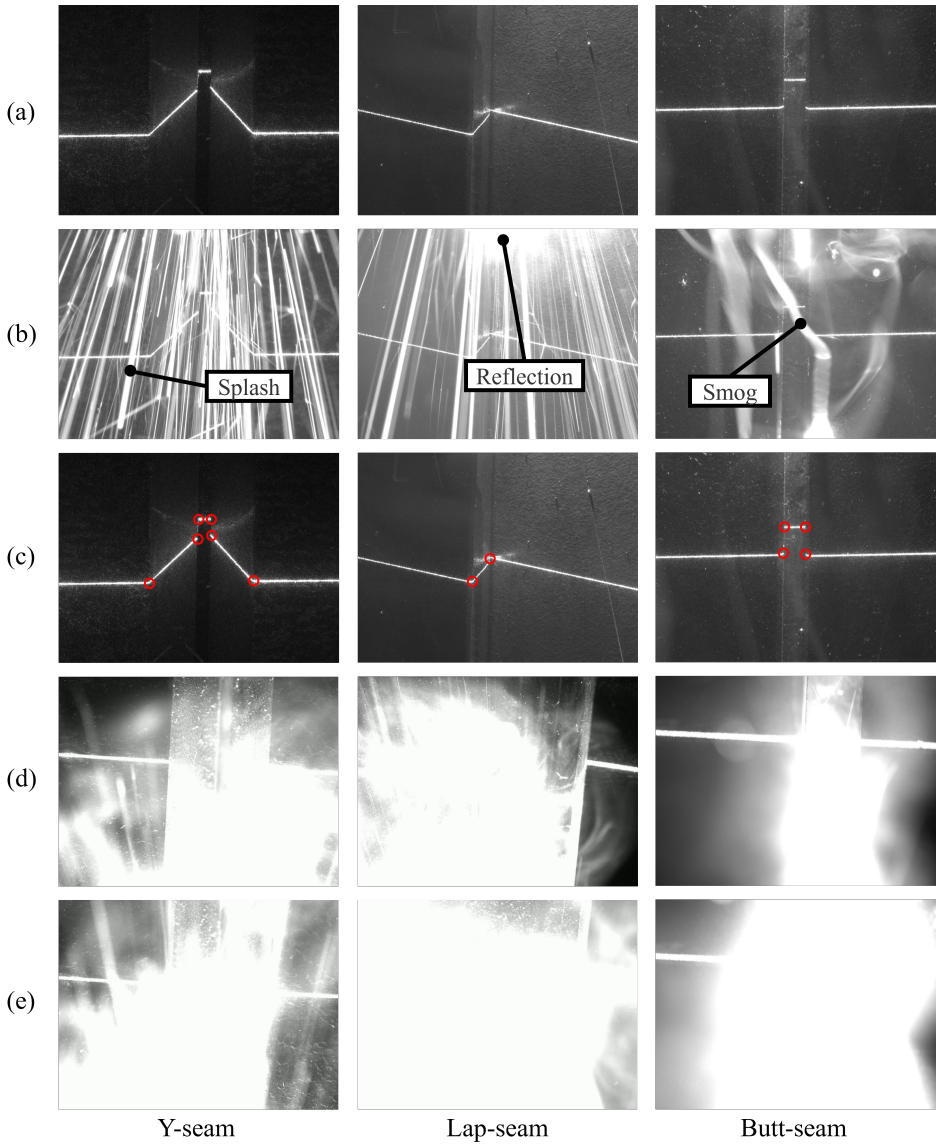

Figure 1. Discerning signal and noise characteristics in industrial camera images for y-seam, lap-seam, and butt-seam welds. (**a**) Noise-free structured light stripe image. (**b**) Structured light stripe image affected by noise. (**c**) Characteristic point of the structured light stripe. (**d**) Part of the feature points affected by noise. (**e**) All feature points affected by noise.

The evolution of deep learning technology in computer vision has led to the refinement of object detection and semantic segmentation, enabled by advances in computing hardware and the development of convolutional neural networks (CNNs). These tasks are frequently transferred to seam feature extraction tasks utilizing line structured light, contributing to the development of an anti-noise model for effective feature extraction, as evidenced by existing research.

Gao et al. [10] built the YOLO-WELD model based on YOLOv5 in the task of detecting weld feature points. RepVGG is used as the backbone network, and the NAM attention mechanism and lightweight head layer, RD-Head, are introduced to improve the detection effect. Deng et al. [11] improved CenterNet, used DenseNet as the backbone network, and adjusted the head layer to separate the feature point position regression task from the weld classification task. This operation prevented the general feature detection network from associating different categories with feature points on a laser stripe image. Liu et al. [12] reported a method to extract the feature points of multi-layer and multi-pass welding seams by using a conditional generative adversarial network (CGAN) and improving the CNN model. Cacarion et al. [13] presented a detection model without NMS (non-maximum suppression) and DETR (detection transformer), which considers CNN as a feature extractor, and used an improved Transformer [14] to realize the further mining of the global information of the input picture. Because of the self-attention mechanism of the Transformer, the DETR incurs considerable computational complexity (in this paper, DETR is used as a typical algorithm to compare the inference time of the model).

However, these methods based on neural networks require considerable computing resources; in turn, they necessitate higher computing performance from the central control equipment integrated into welding robot systems. Notably, such systems commonly lack GPU acceleration support. To address this challenge, lightweight neural network techniques have been extensively explored in computer vision. Typical approaches include pruning, knowledge distillation, weight quantization, and especially structural simplification. A variety of effective lightweight backbones have emerged, ranging from MobileNet [15–18] and ShuffleNet [19,20] to newer designs such as Shvit [21] and LSNet [22]. Alongside these developments, many studies have begun integrating attention modules [23–25] into lightweight architectures to counterbalance the performance degradation often induced by aggressive model simplification.

Liu et al. [26] extended the concept of depthwise separable convolution from MobileNet to 3-D convolution. They combined the 2-D input image with temporal information from image sequences to assess the current welding state. Additionally, three attention mechanisms were introduced to enhance the model's robustness against welding noise. Ma et al. [27] developed the WeldNet model specifically for welding tasks, encompassing starting point detection and seam classification. The model's architecture employs ShuffleNet as the backbone network. However, although these methods adopt lightweight network architectures and demonstrate real-time performance on GPUs, their CPU performance remains limited, and they require large-scale training datasets, thereby hindering the cost-effective deployment of neural networks in industrial applications.

There are also some feature extraction methods based on the semantic segmentation model. Zou et al. [28] developed a lightweight laser stripe image segmentation model by replacing the backbone with ShuffleNetv2. They achieved laser stripe segmentation and improved the model by a pruning operation based on trainable parameters of the Batch Normalization (BN) layer, combined with a welding seam tracking algorithm. Despite maintaining a high inference speed on the CPU, this method does not directly provide the specific position of the welding seam in a one-stage manner.

Additionally, an efficient ViT (Vision Transformer) [29] based on cascaded grouping attention was proposed, a substitute for multi-head self-attention (MHSA). The structure of this attention mechanism can have fewer parameters and obtain higher inference speed because it is memory-efficient.

Utilizing a lightweight model in welding tasks can decrease the number of network parameters and floating-point operations (FLOPs), making it more suitable for deployment on edge computing devices with limited computational capabilities. Nevertheless, this comes at the cost of diminished performance in abstract feature extraction, potentially weakening the model's noise robustness.

To address the issues above, the objectives of this paper are as follows:

1. To propose a one-stage model for (1) extracting the position of light stripe feature points of weld structures and (2) performing weld classification (e.g., identifying groove types) by analyzing distinctive geometric patterns in the light stripe.
2. The proposed method can be used to enhance the model's robustness to image noise during welding without affecting the real-time performance of the model.
3. To reduce the model's parameters and computational cost can be deployed on the computing platform where GPU acceleration is unavailable, such as an embedded industrial computer (EIC), and at the same time maintain a high inference speed to meet the real-time requirements of welding tracking systems in the industrial field.
4. Compared with typical one-stage target detection models, the study aims to test and validate the feature point positioning performance under different noise conditions, weld classification performance, and real-time performance of the proposed model.

The paper comprises five sections. Section 2 details the hardware devices and datasets employed in the study, while Section 3 provides a structural description of WeldLight. Following this, Section 4 outlines the specific configuration of the experiment and results obtained from the comparative verification of various models. Section 5 presents the obtained conclusions.

2. Hardware Platform and Dataset

2.1. Hardware Platform and Measuring Principle

As shown in Figure 2, the self-designed welding seam positioning and tracking sensor based on single-line structured light includes one industrial camera, a single-line laser transmitter, a narrow band filter, and others. The model of the used industrial camera is CB016 from HIKROBOT with a sensor size of $1/2.9''$ and 1440×1080 resolution, equipped with a lens with a focal length of 25 mm. The wavelength of both the single-line laser transmitter and narrow band filter is about 660 nm.

As shown in Figure 3, a pixel coordinate frame (PCF) and a camera coordinate frame (CCF) are established in the line structured light imaging system illustrated in the figure; a random feature point in 2-D PCF can be transformed into 3-D CCF by

$$p_c = z_c M^{-1} q_p, \tag{1}$$

where p_c is the coordinate of points on the laser stripe relative to CCF. q_p is the 2-D coordinate of the feature point of the welding seam on the PCF, with the upper left corner of the image as the origin of the coordinate and the right and downward directions as positive, respectively. z_c denotes the projection value on the Z axis of the feature point with respect to CCF, which is an unknown variable. M denotes the intrinsic matrix of the camera, which can be calculated by [30].

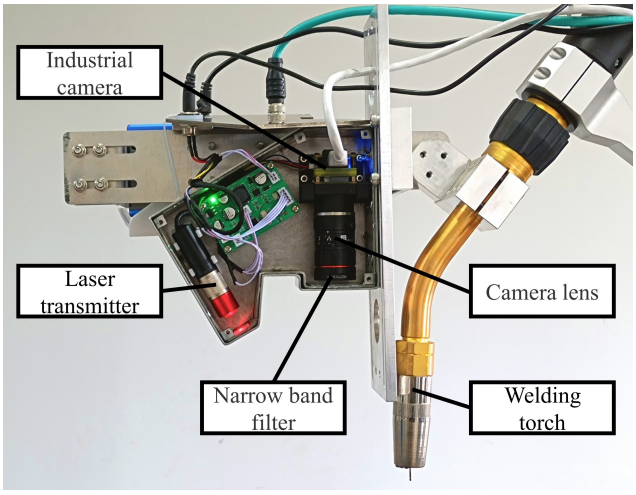

Figure 2. Laser vision sensor experimental platform.

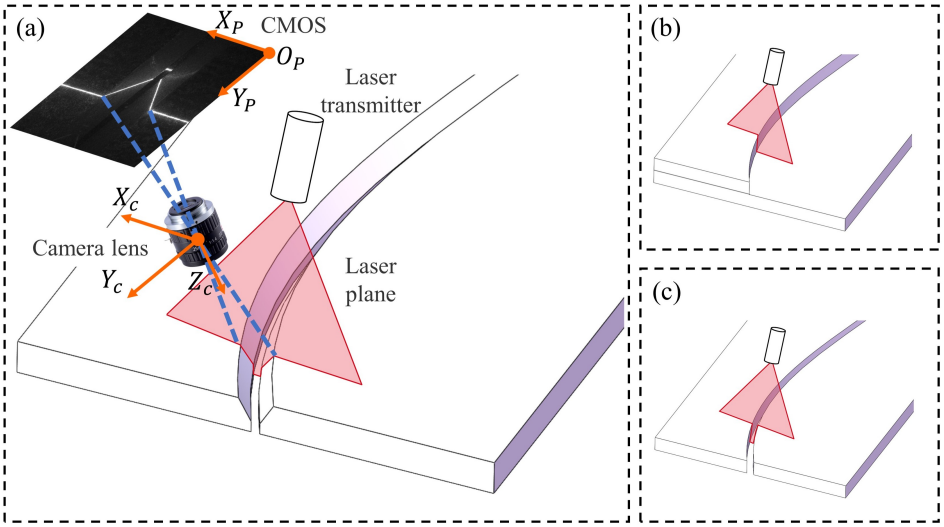

Figure 3. The effect diagram of structured light projection on different welds. (**a**) illustrates the projection effect on a Y-seam and embodies the measurement principle of a single-line structured light vision sensor, and (**b**,**c**) depicts the projection effect of structured light on the lap-seam and butt-seam, respectively.

The laser plane equation of the line structured light with respect to the CCF can be determined by

$$\mathrm{n}^{\mathrm{T}} \cdot p_c + \mathrm{d} = 0m \tag{2}$$

where $\{\mathrm{n}, \mathrm{d}\}$ stands for the normal vector of the plane and the distance between the origin of CCF and the laser plane, respectively, which can be calculated by the calibration method proposed in [31] to obtain the general equation of the structured light plane relative to the CCF. According to (1) and (2), the coordinates of feature points concerning CCF can be represented by

$$p_c = -\frac{\mathrm{d}}{\mathrm{n}^T M^{-1} q_p} M^{-1} q_p. \tag{3}$$

Equation (3) will be employed in Section 4.3 to assess the comprehensive performance of the model and hardware platform on feature point positioning.

2.2. Welding Image Dataset

In complex and variable industrial settings, neural networks often require new training data to adapt to evolving environments, and precise annotation typically demands substantial human effort. Consequently, developing models suitable for small-batch datasets becomes necessary. A total dataset of 1500 images (1440 × 1080 pixels), comprising Y-seam, lap-seam, and butt-seam welding stripes captured by CMOS cameras, was collected. We divided the test images into low-noise and high-noise subsets to better evaluate the robustness of the network under different operating conditions. The low-noise subset consists of noise-free and slightly noisy images, in which light stripes remain clear, and the minor disturbances can still be reliably recognized using traditional morphological methods, as illustrated in Figure 4a,b. The high-noise subset includes images with extensive splash and fume noise that severely interfere with recognition and even obscure critical features. Such cases are generally difficult to identify using morphological methods, as shown in Figure 4c, and they account for approximately 50% of the test set. The dataset was manually annotated, where the positions of occluded keypoints in the high-noise subset were estimated based on the continuity of light stripe variations.

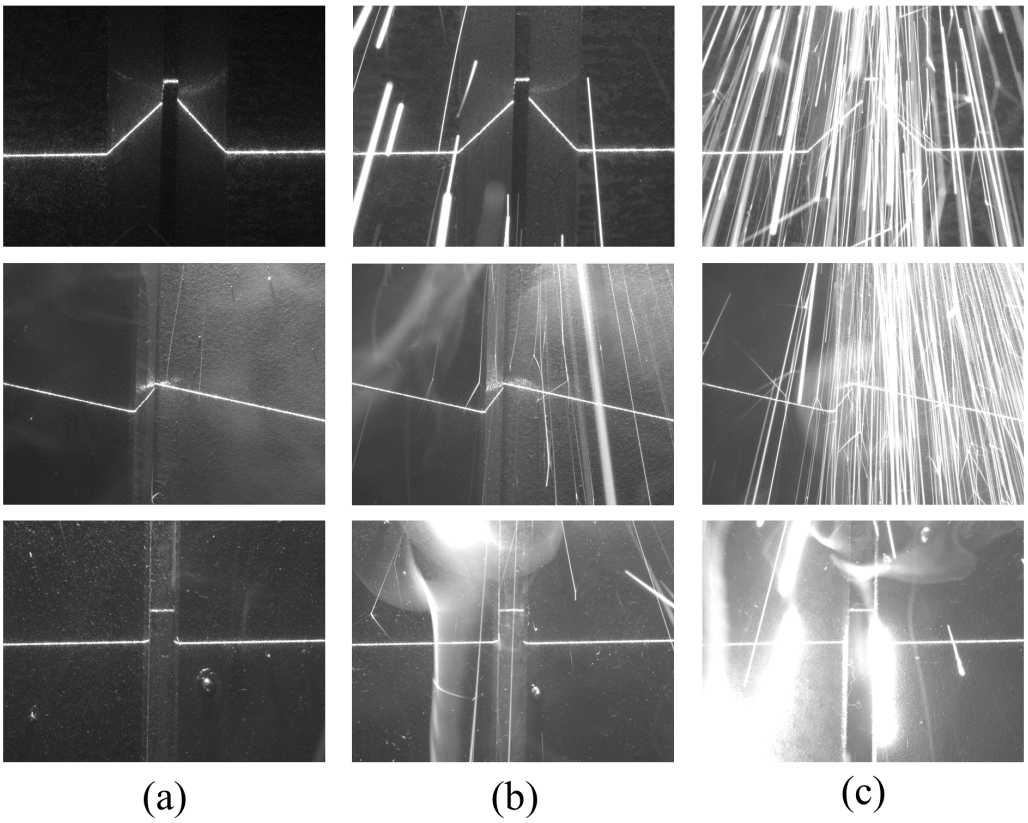

Figure 4. Images with different noise levels in the dataset. (**a**) Noise-free image; (**b**) slightly noisy image; (**c**) high-noise image.

It is worth noting that we excluded extremely noisy examples in the high-noise subset (e.g., Figure 1e), which may result from occasional extreme welding noise, highly reflective components, or hardware design limitations. In such cases, most light-stripe features become invisible, making single-image recognition ineffective. Fortunately, such extreme noise conditions are rarely encountered in typical welding scenarios. When they do arise, mitigation can be achieved through specialized hardware improvements (e.g., adding optical filters or adjusting laser wavelengths) or through downstream processing. The latter may involve trackers that integrate temporal information to predict the current occluded

point or, alternatively, skipping the frame and continuing with welding path fitting. In this study, Kalman filtering was adopted for such predictions, although its details are not elaborated here.

The dataset division is detailed in Table 1. Although the evaluation in this work is restricted to three weld seam types, the approach is not confined to these cases; with appropriate specification of feature points, it can be generalized to a wider variety of seam geometries.

Table 1. Partition of the dataset.

Dataset	Y	Lap	Butt	Total
Train	415	404	396	1215
Val	33	45	57	135
Test	52	51	47	150
Total	500	500	500	1500

2.3. Online Data Augmentation

The quality of neural network training is highly dependent on the quality and scale of the collected dataset; however, data collection and annotation are often labor-intensive and time-consuming. Therefore, in many real-world applications, data augmentation is commonly employed to simulate variations in working environments, sensor models, and sensor positions, thereby enabling effective training on limited data. For example, previous studies [32,33] designed data augmentation techniques tailored to their respective application domains, which improved the dataset quality and enhanced the performance of neural network models.

In this study, we similarly adopt data augmentation to enhance the robustness of the neural network model. Specifically, data augmentation is applied to increase the diversity of training samples and improve the model's generalization ability through techniques such as image flipping, warping, affine transformations, and variations in hue, saturation, and value. In addition, common noise patterns observed in welding processes are also simulated to further improve the training quality of the network.

Splashes are the main interference factor in detecting feature points captured by the camera. To mitigate the impact of splash noise on the neural network model, a method involving the random generation of splashes is introduced into the traditional online data augmentation step to enhance the learning of splash features by the network model. So, in every epoch, the images loaded from the dataset undergo different forward passes and backward propagations. It is crucial to note that this process is distinct from the offline splash addition introduced in this work, as detailed in [10]. As illustrated in Algorithm 1, for each noise parameter, we initialize an empty splash image and generate *num* random lines. Each line is assigned a random starting point near the light source, a random direction, and random attributes including width and intensity. The line is then randomly segmented and rendered on the splash image. After applying Gaussian blur to the splash image, it is superimposed on the original image. Finally, the original images are blended with a noise template and subjected to a sequence of augmentation operations, including cropping, scaling, affine transformations, and brightness adjustments. The overall effect of this noise generation process is shown in Figure 5.

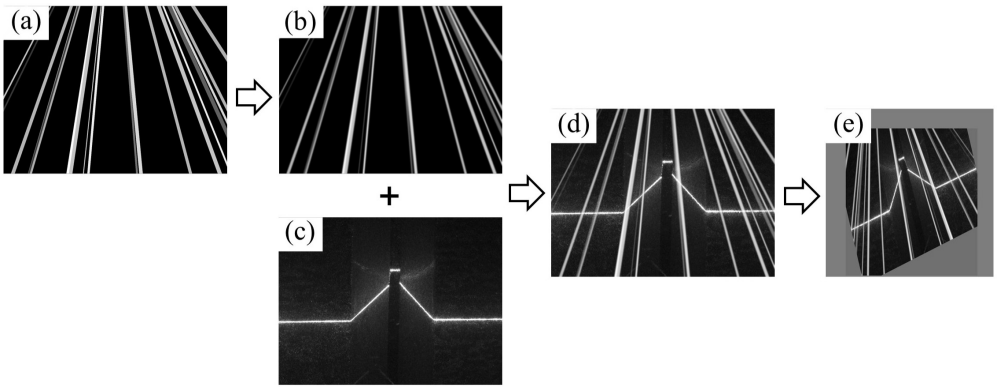

Figure 5. The operation flow of the proposed data augmentation method. (**a**) is a randomly generated simulated splash light noise map; (**b**) is the result of Gaussian filtering on (**a**); (**c**) is the original image from the dataset; (**d**) is the result of superimposing (**b**) on (**c**); (**e**) indicates the results of other data augmentation operations.

Algorithm 1 Random Noise Generation Method

procedure RANDOMNOISE($Image, num$)
  for each parameter p in *noise_parameter* **do**
    $splash_image \leftarrow \mathbf{0}$ ▷ zero matrix with same size as $Image$
    for $i \leftarrow 1$ to num **do**
      $P_{start} \leftarrow$ GENERATEPOINT(light source region)
      $d \leftarrow$ RANDOMDIRECTION()
      $(w, \alpha) \leftarrow$ RANDOMWIDTHINTENSITY()
      $L \leftarrow$ CONSTRUCTLINE(P_{start}, d, w, α)
      $L \leftarrow$ TRUNCATELINE(L)
      DRAWLINE($splash_image, L$)
    end for
    $splash_image \leftarrow$ GAUSSIANBLUR($splash_image$)
    $Image \leftarrow Image + splash_image$
  end for
  return $Image$
end procedure

3. WeldLight Structure

3.1. Overall Structure

As shown in Figure 6, WeldLight can be divided into three parts: the backbone layer, the neck layer, and the head layer.

The backbone layer is responsible for extracting features and outputting feature map branches with different downsample factors, which are then delivered to the neck layer. MobileNetV3-Small is chosen in the proposed work. In the MobileNetV3 block, a sequence consisting of a 3×3 depthwise convolution (DWC) with a BN layer and different nonlinearities (i.e., h-swish and ReLU), followed by a channel-fusing 1×1 pointwise convolution (PWC) with a BN layer, is proposed to replace one standard 3×3 convolution with BN and ReLU. Notably, a similar DP block (Figure 6) is used in the neck and head layers of WeldLight, employing the ReLU6 function as the nonlinearity to better suit feature-point localization. In terms of convolution computation, the quantitative relation between DP block and one standard 3×3 convolution is given by (4).

$$\frac{k \times k \times C_{in} \times H \times W + C_{in} \times C_{out} \times H \times W}{k \times k \times C_{in} \times C_{out} \times H \times W} = \frac{1}{C_{out}} + \frac{1}{k^2}, \tag{4}$$

where the numerator and denominator of the fraction represent the computations of the DP block and the standard 3×3 convolution, respectively, k denotes the kernel size, and C_{in}, C_{out}, H, W denote the input channels, output channels, and the spatial dimensions of the input feature map. Generally, there are many output channels of convolution operation, especially the number of channels in the deep layer, which is much larger than the size of the convolution kernel. Consequently, the calculation can be reduced to about $\frac{1}{k^2}$.

In the neck layer, a feature pyramid network (FPN) [34] construction is applied. Due to many stages of backbone producing feature maps of the different resolutions with different semantic levels, the deepest three branches produced by the backbone with downsampling strides specifically of $\{8, 16, 32\}$ with respect to the input image will be fused by FPN, as shown in Figure 6. Though more branches merged in the neck layer indicates the model can integrate deep abstract semantic information and superficial coarse-grained feature information to achieve better small-scale feature detection capability, the number of branches should not increase without limit, because more branches will seriously slow down the inference speed [20]. Through experimentation, the optimal branch number of three was determined to minimize computational complexity and latency while preserving a certain level of positioning accuracy.

As depicted in Figure 7b, the improved top-down pathway and lateral connection employ concatenation instead of addition, as seen in the original approach shown in Figure 7a. Channel adjustment facilitates merging the feature map from the PWC with the upsampled output of the coarser-resolution feature map containing stronger semantic information. A DP block is also utilized to substitute the standard 3×3 convolution. Ultimately, the neck layer will generate a feature map of size $128 \times 128 \times 64$, which will subsequently be input into the head layer.

Additionally, within the neck layer, the network links a proposed attention module to the output of $32\times$ downsampled feature map, aiming to enhance the model's robustness against noise, as detailed in Section 3.2.

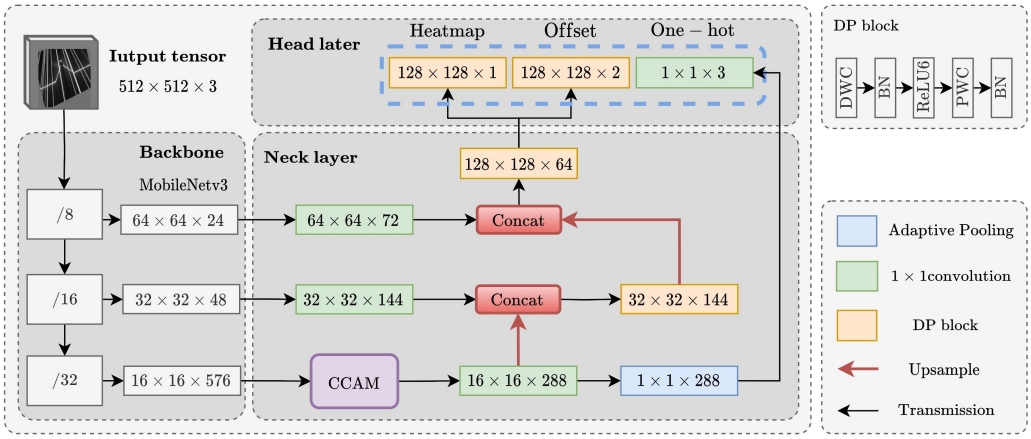

Figure 6. Overall layout of the proposed network.

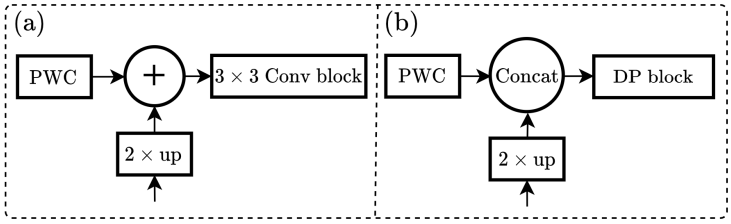

Figure 7. Architecture of top-down pathway and lateral connection. (**a**) is the original version of FPN; (**b**) is the proposed version.

In the head layer, it is necessary to predict the position of feature points, which is realized by a heatmap tensor and an offset tensor. The task also requires classifying the laser stripes to determine the weld type, which requires the head layer to output a one-hot tensor. The standard 3×3 convolution block for decoupling the input tensor of the head layer is also replaced by a DP block. At this stage, the PWC not only fuses features across channels but also adjusts the number of channels. Since WeldLight does not classify each feature point separately but instead predicts a category for the entire laser stripe image, the categorical output of the head layer mainly derives from the lower-resolution feature map in the deepest backbone layer, enabling large-scale perception and producing a global category prediction rather than point-specific classification.

3.2. Cascaded Channel Attention Module (CCAM)

The deepest output shape from the backbone network obtained by FPN is $H_{in} \times W_{in} \times C_{in}$, specifically $16 \times 16 \times 512$. It should be noted that a significant portion of the features are related to noise. These features will be merged in the neck layer and finally affect outputs from the head layer. In order to weaken the influence of noise, CCAM is proposed to filter the noise.

As shown in Figure 8, the core of CCAM is a channel attention (CA) block, which serves as a feature selector. Initially, a global average pooling layer is used to reduce the input feature map to a single vector, which is then fed to the first fully connected (FC) layer to reduce its dimensionality. A second FC layer restores the feature dimension. The resulting vector is passed through a softmax function to map it to the range of [0, 1], which enables channel-wise weighting of the input feature map through element-wise multiplication. Channels with higher weights are considered more significant.

In the CCAM design, we adopt softmax rather than sigmoid for channel weighting. While sigmoid is commonly used, it treats channels independently and may assign similar weights across them, limiting discrimination. Softmax instead normalizes the weights to sum to one, emphasizing the relative differences among channels and thereby supporting more effective suppression of noise and highlighting informative features. This choice is also consistent with established deep learning practices, where softmax is widely employed for normalized weighting in multi-channel scenarios [35]. To summarize, the refined feature map F_{CA} by CA block is computed as

$$F_{CA}(\mathrm{z}, \mathrm{W}) = \mathrm{Softmax}(\mathrm{W}_1(\mathrm{W}_0 AvgPool(\mathrm{z}))) * \mathrm{z}, \tag{5}$$

where z denotes the input feature, and $\mathbf{W}_0 \in \mathbb{R}^{\frac{C}{g \times r} \times \frac{C}{g}}$ and $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{g} \times \frac{C}{g \times r}}$ denote the parameters of two FC stages, in which r, as shown in Figure 8, stands for the reduction ratio with 16 in WeldLight, and C denotes the input tensor number of CCAM, while g stands for the number of groups (splits) for the input tensor.

In addition, CCAM adopts a cascaded group structure to further reduce the computational overhead. As illustrated in Figures 6 and 8, the deepest output feature map of the backbone is split into g parts, which are then fed into the CCAM. In the CCAM, the output from each CA block is merged with the next split by element-wise addition to enhance the information fusion between channels. The process is repeated across all g groups.

For CCAM, the computation generated by the FC layer is $\frac{2C^2}{g \times r}$. Instead of relying on the cascaded group structure, the computation generated by the FC layer through direct utilization of channel attention leads to a calculation amount of $\frac{2C^2}{r}$. The computation of the CA module, when employing a cascaded group structure, is approximately $\frac{1}{g}$ of the unused capacity. However, this structure also enables multi-layer attention effects, allowing the model to refine its feature selection at each stage.

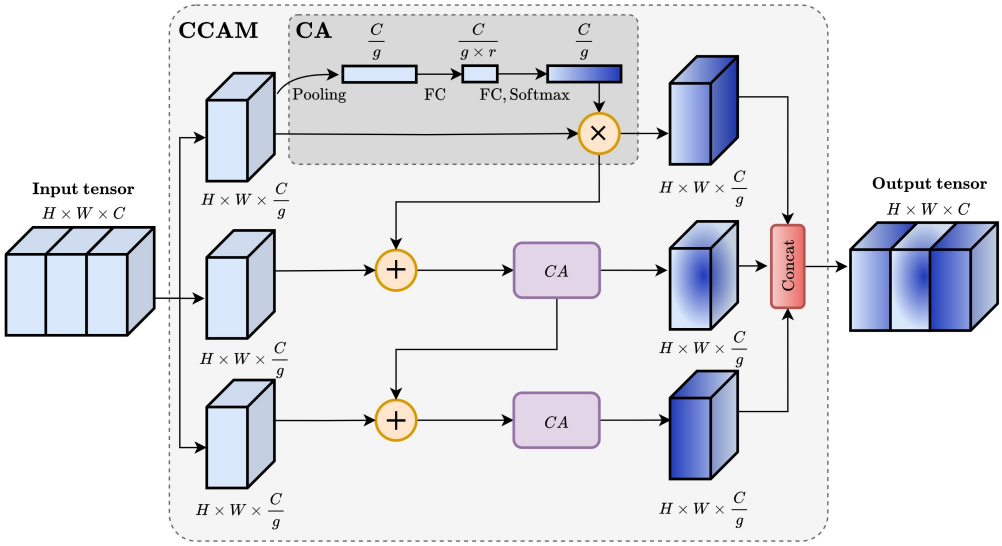

Figure 8. Specific structure of cascade channel attention module.

3.3. Loss Function

The head layer produces three outputs, a heatmap tensor $\hat{M} \in [0,1]^{\frac{W}{R} \times \frac{H}{R} \times 1}$, an offset tensor $\hat{O} \in [0,1]^{\frac{W}{R} \times \frac{H}{R} \times 2}$, and a classification tensor (i.e., one-hot encoding) $\hat{K} \in [0,1]^{1 \times 1 \times c}$, respectively, where W and H are the input image dimensions of the model, and c is the number of categories. Throughout the training process, the loss is determined by comparing the predictions generated by the head layer with the ground truth label values, emphasizing that the shapes of the label values should match those of the predictions. This contrast aims to optimize the neural network parameters to achieve an optimal score or regression value prediction for the model's output. Consequently, it is essential to address the generation of labels in this context. The coordinate of the feature points location, following their mapping to the corresponding bin in the label of heatmap tensor, is $(\tilde{G}_x, \tilde{G}_y) = (\lfloor \frac{x_k}{R} \rfloor, \lfloor \frac{y_k}{R} \rfloor)$, where (x_k, y_k) is the coordinate of the k-th feature point with respect to the input image size of the model (i.e., 512 × 512 for WeldLight), and $R = 4$ denotes the downsampling stride factor. On the label heatmap, positive sample points are represented using Gaussian circles $M(x,y) = \exp\left(-\frac{(x-\tilde{G}_x)^2+(y-\tilde{G}_y)^2}{2\sigma_p^2}\right)$ to accelerate the convergence of the model [36,37], where σ_p is a factor that varies with the size of the object [37], and $M(\tilde{G}_x, \tilde{G}_y) = 1$ stands for the center of Gaussian circles, which is the ideal positive location of feature points mapping on the heatmap.

In light of the inherent rounding effect of floor division, relying solely on the heatmap to restore position information for feature points can result in accuracy loss. Offsets are introduced to address this issue. The label value of offsets of every feature point is denoted by

$$O_k = \left(\frac{x_k}{R} - \tilde{G}_x, \frac{y_k}{R} - \tilde{G}_y\right), \tag{6}$$

where O_k is a 2-D coordinate representing the relative position proportion of a feature point in a corresponding bin of the label heatmap. Therefore, the offset is represented by a two-channel tensor. Additionally, the offset tensor bins are activated only at the corresponding positive locations within the heatmap tensor.

Let $\hat{M}_{ij}$ be the predicted score at location (i,j) of the heatmap produced by WeldLight, and let M_{ij} be the value at the same location on the label heatmap with Gaussian circles; a variant of focal loss [38] is used to optimize the heatmap prediction, denoted by

$$L_{\text{hm}} = \frac{-1}{N} \sum_{i=1}^{H} \sum_{j=1}^{W} \begin{cases} (1-\hat{M}_{ij})^{\alpha} \log(\hat{M}_{ij}) & \text{if } M_{ij} = 1 \\ (1-M_{ij})^{\beta} (\hat{M}_{ij})^{\alpha} \log(1-\hat{M}_{ij}) & \text{otherwise} \end{cases}, \tag{7}$$

where N is the number of feature points in a structured light stripe image, H and W stand for the spatial size of the heatmap, α is utilized to fine-tune the weights for challenging and easily locatable points, while β governs the weighting of non-central values within the Gaussian circle. Importantly, the loss weight assigned to a predicted score in $\hat{M}$ increases, as the distance from the Gaussian center in the label grows.

The essence of offset prediction is rooted in a regression task. In order to realize the accurate offset prediction, a smooth L1 loss function [39] is utilized as follows:

$$L_{off} = \frac{1}{N} \sum_{i=1}^{N} \text{SmoothL1Loss}(\hat{O}_k, O_k), \tag{8}$$

where N is the number of feature points in a structured light stripe image, too. The smooth L1 loss leverages the advantages of both the L1 and L2 loss functions. This method prevents gradient explosion in the initial stages of training and promotes the acquisition of a more tempered gradient during back-propagation at the end of training. This contributes significantly to improved model convergence.

Cross-entropy loss (CE loss) can be applied to the weld classification task. Here, let $\hat{K}_i \in [0, 1]$ represent the predicted score of structured light stripe image for class i, predicted by WeldLight, and boolean variable K_i denotes the label value of class i in the one-hot encoding K; hence, the loss produced by classification is denoted by

$$L_{cls} = \text{CEloss} = -\sum_{i=1}^{c} K_i \log(\hat{K}_i), \tag{9}$$

where c stands for the total number of welding seam types. Finally, the total loss L_{total} is the overall training objective, which is derived as follows:

$$L_{total} = L_{hm} + \lambda_{off} L_{off} + \lambda_{cls} L_{cls}, \tag{10}$$

where λ_{off} and λ_{cls} represent the constant weights assigned to the offset and classification loss, configuring to a value of 1.

3.4. Post-Processing

A correlation exists between the prediction of point numbers and laser stripe classification, as discussed in [11]. Specifically, if the classification result is predicted as "lap-seam," the model identifies the first two bins on the heatmap tensor with the highest predicted scores. These are then combined with the corresponding bins on the offset tensor, leading to the identification of the final predicted positions. In this scenario, there is no need to consider the suppression by predicted score thresholds. However, as depicted in Figure 9, the heatmap tensor undergoes resizing to match the uniform shape of the original image, covering it entirely. The colors in the heatmap indicate the predicted scores assigned to the potential feature points.

Notably, the figure illustrates a broken laser stripe, where both ends are confidently identified as potential feature points by WeldLight, despite their proximity. An empirical solution to this issue involves applying a 5×5 max-pooling operation to the heatmap. Bins with unchanged values correspond to effective potential feature points by comparing the heatmap tensor before and after pooling. In cases of closely located potential feature points, only the one with the highest predicted score is retained, effectively resolving the issue.

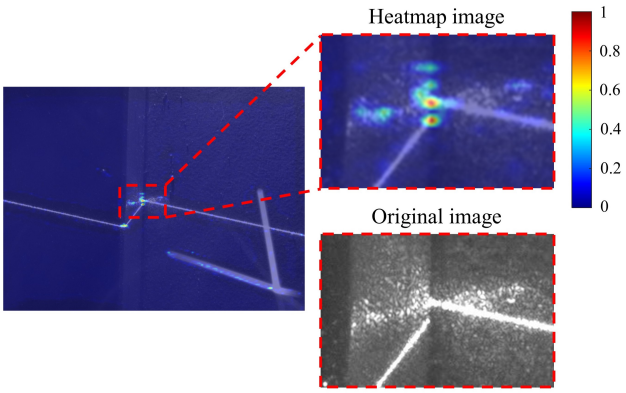

Figure 9. A case in which some candidate feature points may not be filtered out when the pool filter is small.

4. Verification

4.1. Test Environment

The network was constructed using the PyTorch 2.0.0 deep learning framework based on Python 3.8. The dataset utilized for training and testing the model was obtained from the welding seam positioning and tracking sensor mentioned earlier. The model was trained on a computer with an Nvidia 2080Ti GPU and an Intel Xeon (R) E5-2683 CPU. The model was converted to the ONNX format and used OpenVINO for DNN inference on an EIC with an Intel Core i7-8650U CPU. The network initializes its backbone with the weights of MobileNetv3 pretrained on ImageNet to expedite training and achieve rapid convergence. The FC layer was implemented using a standard 1×1 convolution, and the convolution layer parameters were initialized using the He initialization method. As for the BN layer parameters, the weights were initialized to one, and the biases were initialized to zero.

The initial learning rate was set to 5×10^{-4} using Adam as the optimizer (momentum = 0.9, weight decay = 0), and a cosine annealing strategy was adopted for learning rate scheduling. The batch size of the training process was set to 32. The maximum number of training epochs was set to 500, and an early stopping strategy was applied such that the training was terminated at the 50th epoch once the validation loss had stabilized. These hyperparameter values were determined empirically to balance convergence speed and training stability.

Upon observing convergence in the loss function of the validation set during the training stage, the model was utilized to evaluate both the positioning performance of the feature points and the classification performance of the welding seams. The comparative model opted for YOLOv5n and DETR, trained on the COCO dataset, to obtain initial weights, which were subsequently fine-tuned using the default training configurations provided by the respective open-source projects.

4.2. Weld Classification Performance

For welding tasks, the baseline network used in WeldLight is CenterNet, which, similar to DETR and YOLOv5n employed for comparison, belongs to the family of object detection networks. Since a welding image should contain only a single class, previous studies have commonly adopted the class of the point with the highest classification confidence as the image-level label. WeldLight, however, incorporates a dedicated prediction branch to directly infer the class of the entire image rather than distinguishing the classes of individual feature points, thereby better adapting to the requirements of weld seam classification in welding operations. In contrast, DETR, through its end-to-end prediction mechanism, is able to model the relationship between the global image context and the classification of feature points, thereby reducing the risk of assigning multiple classes to different points

within the same image. By design, YOLOv5n assigns a class label to each detected point, which necessitates determining the overall image category based on the point with the highest classification confidence.

Notably, all three models accurately classified 150 images in the test set. In pursuit of a more granular comparison, the images underwent adjustments, including random cropping, translation, deformation, and brightness modification. For a visual representation, a selection of the test set images utilized in the classification performance evaluation is presented in Figure 10.

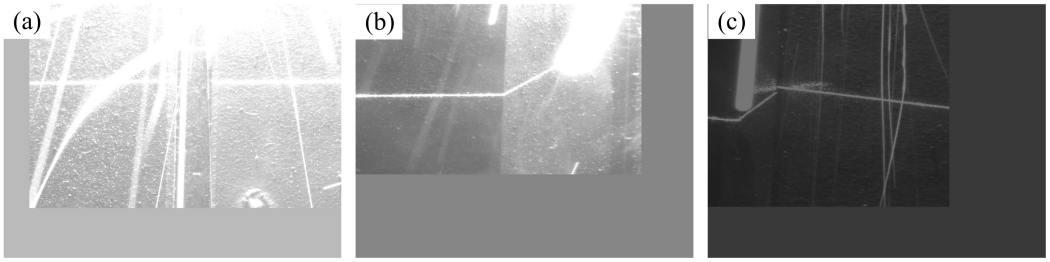

Figure 10. Part of the pictures in the adjusted test set used in the performance test of welding classification. (**a**–**c**) are examples of the adjusted Y-seam, lap-seam, and butt-seam images in the classification performance test.

The adjusted test set was used to evaluate the three models, respectively. Finally, the classification performance was evaluated by the confusion matrix obtained by each model, presented in Figure 11.

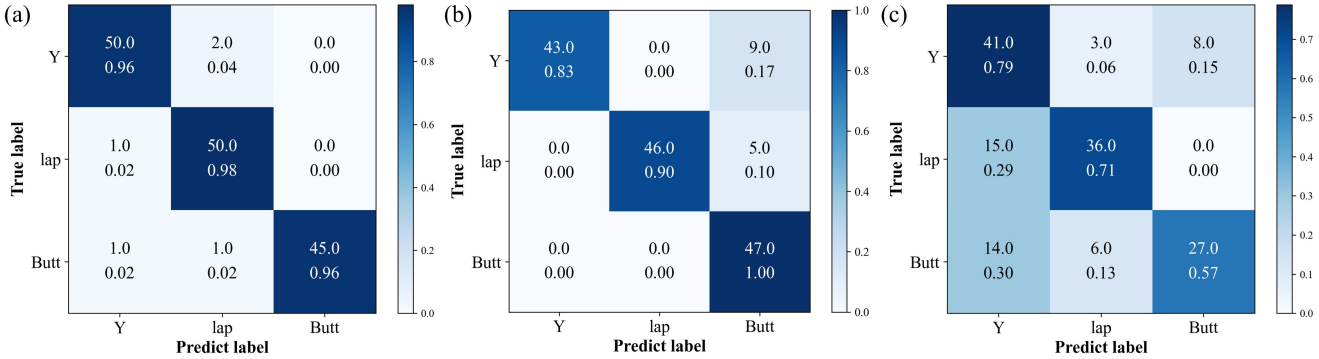

Figure 11. Confusion matrix obtained by different models on the adjusted test set. (**a**–**c**) correspond to the prediction results of WeldLight, YOLOv5n, and DETR, respectively.

The core of the weld classification involves a multi-classification task. To streamline comparisons and address discrepancies in sample numbers, the evaluation of the three models in welding seam classification employed the following three metrics:

$$P_w = \sum_{i=1}^{N} w_i P_i = \sum_{i=1}^{N} w_i \frac{TP_i}{TP_i + FP_i} \tag{11}$$

$$R_w = \sum_{i=1}^{N} w_i R_i = \sum_{i=1}^{N} w_i \frac{TP_i}{TP_i + FN_i} \tag{12}$$

$$F1_w = \sum_{i=1}^{N} w_i \frac{2P_i \cdot R_i}{P_i + R_i} \tag{13}$$

where P_w R_w $F1_w$ stand for the weighted-average precision, recall, and F1-score, respectively. P_i, R_i, and $F1_i$ are the precision, recall, and F1-score of class i, respectively. w_i

represents the ratio of the sample size of category i to the total sample size. True positive (TP), false negative (FN), and false positive (FP) are values that can be determined from the confusion matrix [40]. Table 2 displays the metric values of the three models on the adjusted test set. Notably, WeldLight exhibited the highest values across the three key metrics assessing classification performance, with a weighted-average precision of 0.9674, a weighted-average recall of 0.9666, and a weighted-average F1-score of 0.9668.

Table 2. Test overall metric scores of different models on adjusted test set.

	WeldLight	**YOLOv5n**	**DETR**
P_w	**0.9674**	0.9281	0.9336
R_w	**0.9666**	0.9067	0.9333
$F1_w$	**0.9668**	0.9090	0.9334

4.3. Feature Point Positioning Performance

Welding-related noise, such as splash, reflection, and smog, can compromise the quality of images captured by industrial cameras. Figure 12 demonstrates that all three models reliably distinguish laser stripes from noise and precisely localize feature points, maintaining robust performance even under noise conditions.

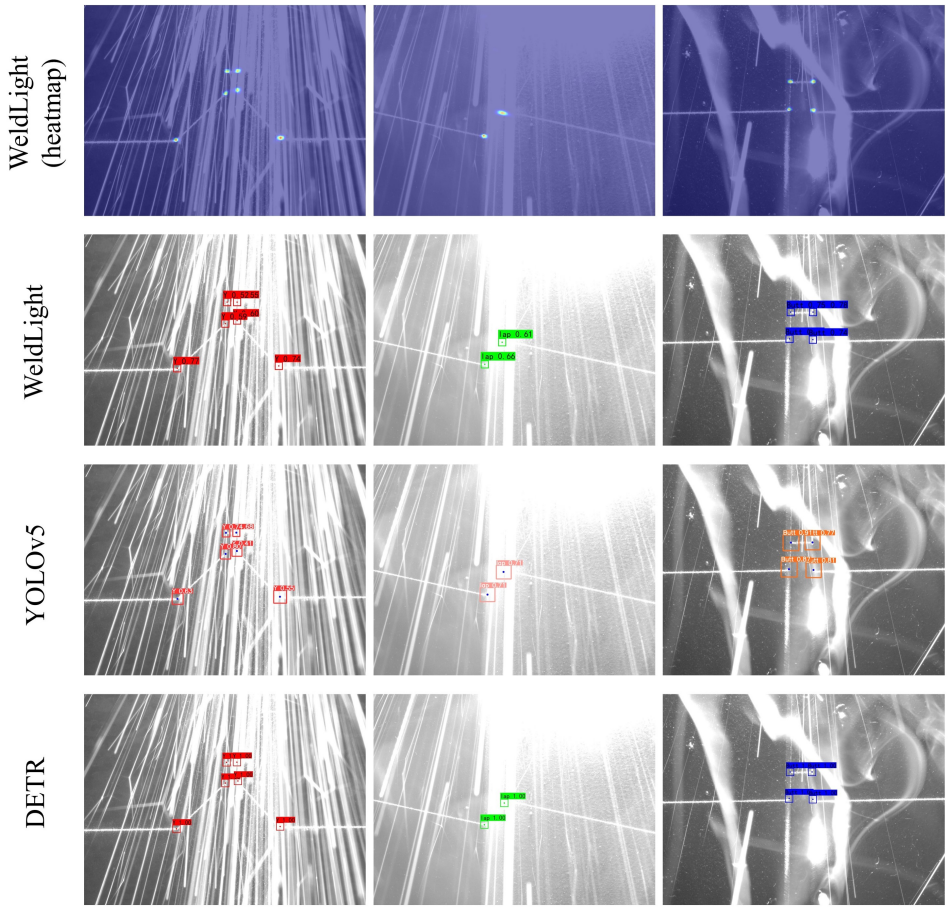

Figure 12. Part of the regression results obtained by different models on the test set.

To evaluate the robustness under different noise levels, the test dataset was divided into high-noise and low-noise subsets, each accounting for 50% of the samples. After post-processing, the three models were evaluated for their feature point positioning performance on both subsets. The predictions were generated for the coordinates of the feature points with respect to the original image resolution of 1440 × 1080. The purpose of

this evaluation is to assess the models' feature point positioning accuracy under varying noise conditions and to analyze their noise robustness. The performance metrics include the mean absolute error (MAE), root mean square error (RMSE), standard deviation (σ_f), and projecting the predicted coordinates and the label coordinates to compute the mean Euclidean distance (ρ_{mean}) with respect to the CCF. The positioning errors of the feature points were calculated along both the *X*-axis and *Y*-axis of the image, as follows:

$$e_i = \begin{cases} q_{xi} - \hat{q}_{xi} & \text{if axis } = x \\ q_{yi} - \hat{q}_{yi} & \text{if axis } = y \end{cases}, \tag{14}$$

where (q_{xi}, q_{yi}) and $(\hat{q}_{xi}, \hat{q}_{yi})$ stand for the coordinate of the label and the predicted coordinate of the feature points, respectively.

The MAE was used to evaluate the static precision of the positioning, which is derived as follows:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} \|\mathrm{e}_i\|, \tag{15}$$

where N denotes the total number of feature points in the designated test set and keeps the same definition in the following formulas.

The RMSE would highlight the impact of predicted outliers, which is derived as follows:

$$\text{RMSE} = \sqrt{\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{e}_{\mathrm{i}}^2}. \tag{16}$$

σ_f was used to evaluate the stability/robustness of the model [7], which is derived as follows:

$$\sigma_f = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (e_i - \bar{e}_i)^2}, \tag{17}$$

where $\bar{e}_i$ is the average value of e_i at all feature points.

The metric ρ_{mean} was employed to assess the error level of three models after projecting 2-D positioning results into the 3-D space. It is derived as follows:

$$\rho_{mean} = \frac{1}{N} \sum_{i=1}^{N} \sqrt{(p_{xi} - \hat{p}_{xi})^2 + (p_{yi} - \hat{p}_{yi})^2 + (p_{zi} - \hat{p}_{zi})^2}, \tag{18}$$

where (p_{xi}, p_{yi}, p_{zi}) and $(\hat{p}_{xi}, \hat{p}_{yi}, \hat{p}_{zi})$, which stand for the label and predicted coordinates of feature points with respect to CCF, can be calculated by Equation (3), based on (q_{xi}, q_{yi}) and $(\hat{q}_{xi}, \hat{q}_{yi})$. An a priori assumption is made that the calibration result of the line structured light plane equation and the intrinsic matrix of the camera is ideal, attributing any error solely to inaccuracies in the feature points' locations within the PCF, as predicted by the positioning models.

The absolute error curve, which represents the test outcomes using three models, is depicted in Figure 13, Figure 14 and Figure 15, respectively. The green curve corresponds to the absolute error measured from the low-noise test set, the blue curve corresponds to absolute error measurements from the high-noise test set, and the red horizontal line denotes the MAE. All metrics related to the positioning errors of the feature points were calculated and arranged in Tables 3 and 4; the MAE, RMSE, and σ_f were assessed along the X and Y directions of the image, and the average values along the two directions were calculated, represented as the average-MAE, average-RMSE, and average-σ_f. Simultaneously, the outcomes of WeldLight without CCAM are included in Tables 3 and 4, too. Notably, there was no specialized positioning performance analysis for different weld types in the test set due to the inherent uncertainty in weld types during welding processes.

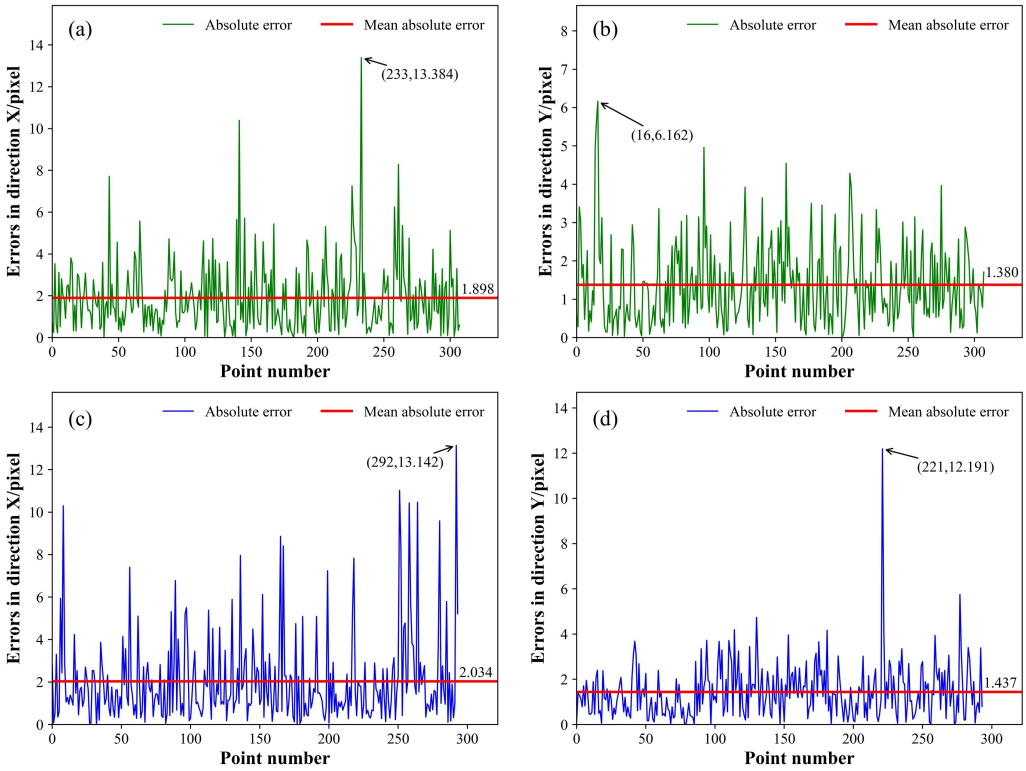

Figure 13. The absolute error curve of the WeldLight. (**a**,**b**) illustrate the feature points regression absolute errors of the model in the X and Y directions, respectively, evaluated on the low-noise test set. (**c**,**d**) showcase the feature points regression absolute errors of the WeldLight in the X and Y directions, respectively, tested on the high-noise test set. The peak value, the peak position of the absolute error, and the MAE are annotated.

Table 3. Statistical characteristics of the positioning performance of the three models for light stripe feature points of the low-noise weld structure. †: WeldLight omited the utilization of CCAM. The units for MAE, RMSE, and σ_f were all expressed in pixels, while the ρ_{mean} was measured in mm.

	Ours	**Ours †**	**YOLOv5n**	**DETR**
X-MAE	1.898	1.859	**1.813**	1.917
Y-MAE	**1.380**	1.502	1.545	1.402
Average-MAE	**1.639**	1.681	1.679	1.660
X-RMSE	2.544	**2.326**	2.401	2.535
Y-RMSE	**1.740**	1.885	1.926	**1.740**
Average-RMSE	2.142	**2.106**	2.164	2.138
X-σ_f	2.258	**2.183**	2.392	2.422
Y-σ_f	**1.627**	1.660	1.911	1.730
Average-σ_f	1.943	**1.922**	2.152	2.076
ρ_{mean}	**0.197**	0.216	0.232	0.209

Note: The best results are highlighted in bold.

The outcomes of the positioning performance evaluation for the weld feature points on a low-noise test dataset are listed in Table 3. Within the low-noise scenario, the three models are basically at the same level for the MAE and RMSE. Notably, regarding the average-σ_f, WeldLight with and without CCAM exhibited outstanding performance compared to the other two models, achieving values of 1.943 pixels and 1.922 pixels, respectively. In terms of ρ_{mean}, WeldLight showed the best performance, reaching 0.197 mm. The utilization of CCAM did not yield a significant difference in the low-noise test set condition.

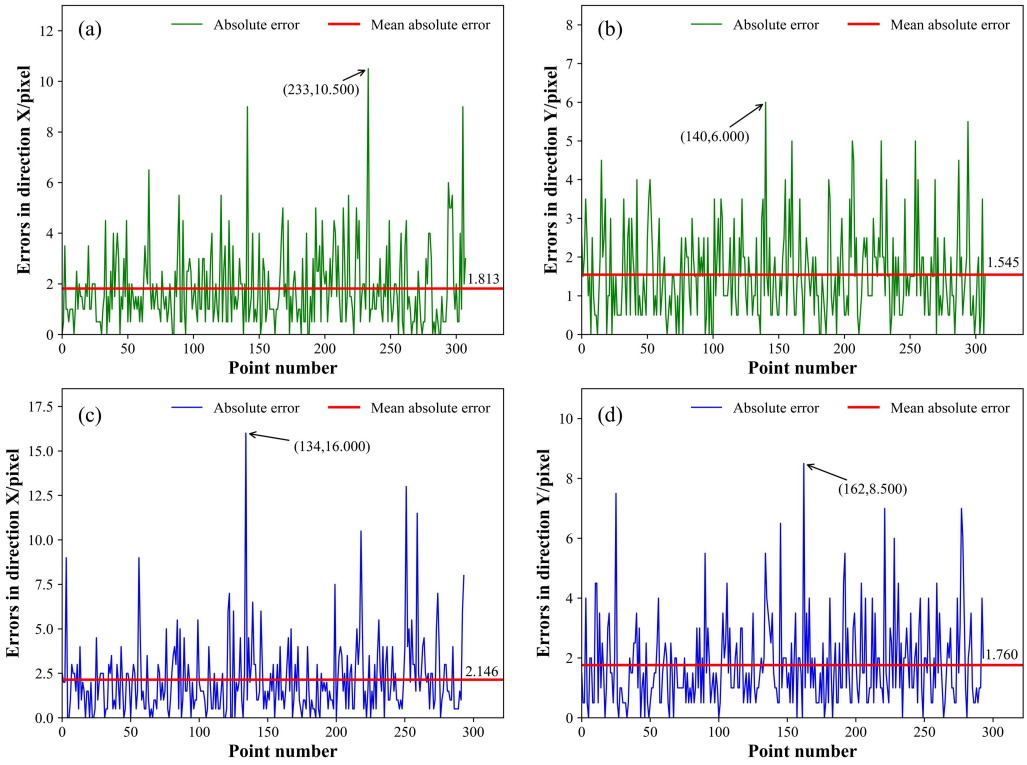

Figure 14. The absolute error curve of the YOLOv5. (**a**,**b**) illustrate the feature points regression absolute errors of the model in the X and Y directions, respectively, evaluated on the low-noise test set. (**c**,**d**) showcase the feature points regression absolute errors of the YOLOv5 in the X and Y directions, respectively, tested on the high-noise test set. The peak value, the peak position of the absolute error, and the MAE, are annotated.

Table 4. Statistical characteristics of the positioning performance of the three models for light stripe feature points of a high-noise weld structure. †: WeldLight omited the utilization of CCAM.

	Ours	**Ours †**	**YOLOv5n**	**DETR**
X-MAE	**2.034**	2.123	2.146	2.080
Y-MAE	**1.437**	1.646	1.760	1.668
Average-MAE	**1.736**	1.885	1.953	1.874
X-RMSE	2.931	**2.927**	3.023	3.043
Y-RMSE	**1.883**	2.206	2.275	2.172
Average-RMSE	**2.407**	2.567	2.649	2.608
X-σ_f	**2.642**	2.731	3.029	2.919
Y-σ_f	**1.792**	2.032	2.254	2.176
Average-σ_f	**2.217**	2.382	2.646	2.548
ρ_{mean}	**0.205**	0.231	0.253	0.238

Note: The best results are highlighted in bold.

The experimental findings from the positioning performance assessment of the welding seam feature points in a high-noise test set are illustrated in Table 4. Within the high-noise environment, WeldLight exhibited an average MAE, RMSE, and ρ_{mean} of 1.736 pixels, 2.407 pixels, and 0.205 mm, respectively. These metrics outperformed both YOLOv5 and DETR. Moreover, compared to WeldLight without CCAM, the model integrating CCAM showcased enhanced performance across nearly all the positioning metrics, reflecting more stable and accurate localization results. This improvement suggests that CCAM contributes to filtering out noise interference and enhancing the robustness of feature extraction, which is particularly beneficial in high-noise welding scenarios. This phenomenon contrasts the measurement results observed in the low-noise test set. WeldLight excels in average σ_f,

registering a value of 2.217 pixel, underscoring WeldLight's superior stability in accurately localizing weld feature points within high-noise welding images.

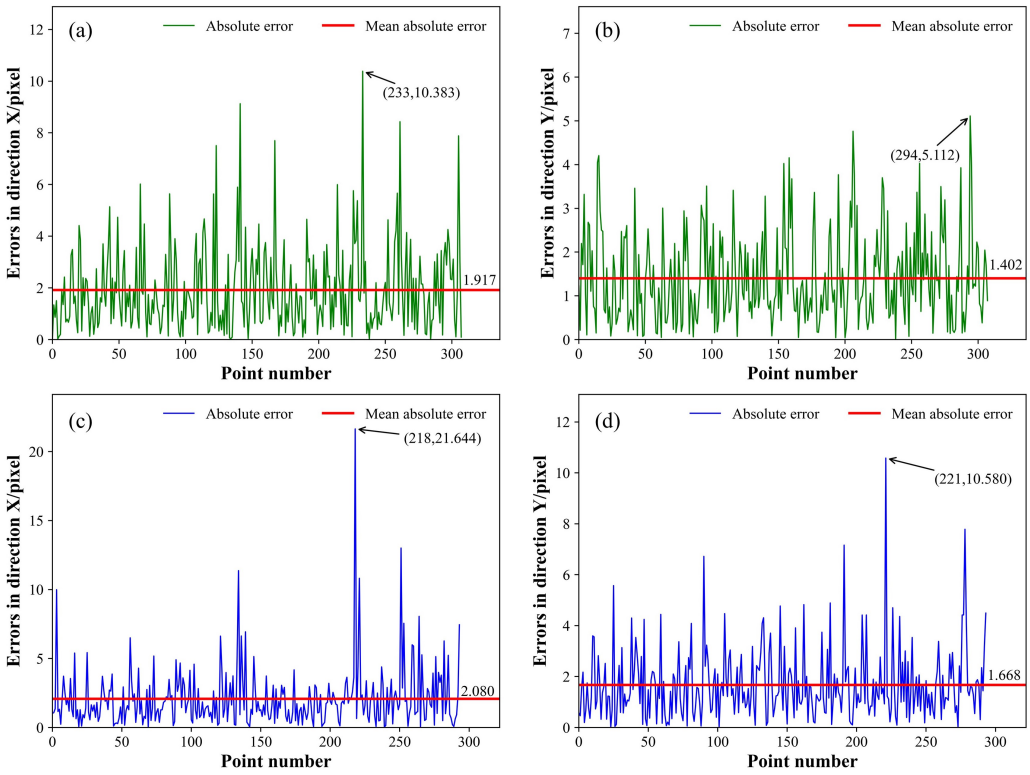

Figure 15. The absolute error curve of the DETR. (**a**,**b**) illustrate the feature points regression absolute errors of the model in the X and Y directions, respectively, evaluated on the low-noise test set. (**c**,**d**) showcase the feature points regression absolute errors of the DETR in the X and Y directions, respectively, tested on the high-noise test set. The peak value, the peak position of the absolute error, and the MAE are annotated.

4.4. Lightweight Level

The model's lightweight characteristics were evaluated based on multiple metrics, including the total number of model parameters (Params), floating point operations (FLOPs), mean latency, and frames per second (FPS). Params serve as an indirect measure of the computational complexity and memory utilization, while FLOPs represent the computational cost of the model [41]. A lower mean latency (or a higher FPS) implies reduced inference time, a critical factor for optimal real-time performance in the welding seam tracking system employing this model. This aspect is particularly crucial for complex seam tracking applications, welding speed, and welding process control. The mean latency of the three models was determined through 100 tests using a single image, and the frames per second (FPS) value was calculated by dividing 1000 ms by the mean latency.

It is worth noting that, despite a consistent evaluation criterion, DETR was not designed for real-time tasks and thus lacks advantages in inference speed. Moreover, Transformer-based architectures such as DETR face inherent challenges in learning low-level features from scratch on limited datasets, making large-scale pretraining essential to achieve competitive performance with CNNs [42]. In this study, these limitations were mitigated by pretraining on the COCO dataset, followed by local fine-tuning until full convergence. Nevertheless, such architectural characteristics remain important considerations when interpreting the results. Naturally, these factors were also taken into account during the initial design of WeldLight, to avoid similar limitations.

Table 5 provides an overview of the lightweight metrics for the three models. WeldLight outperformed the others across all metrics, with the Params being 1.3 M, translating to 72% of YOLOv5n, 3.5% of DETR, and 0.87 GFLOPs, representing 21% of YOLOv5n and 1.5% of DETR. With a mean latency of 29.32 ms and FPS at 34.11 Hz, WeldLight fulfills stringent requirements for real-time weld seam tracking.

Table 5. Comparison of lightweight metrics of three models on Intel Core i7-8650U CPU.

	WeldLight	**YOLOv5n**	**DETR**
Params (M)	**1.3**	1.8	37
GFLOPs	**0.87**	4.1	57
Mean latency (ms)	**29.32**	53.07	681.43
FPS (Hz)	**34.11**	18.83	1.47

Note: The best results are highlighted in bold.

5. Conclusions

1. An improved CenterNet framework was proposed for positioning welding light-stripe feature points. The average MAE in the X and Y directions was 1.639 and 1.736 pixels, respectively, on both low- and high-noise test sets, demonstrating high prediction accuracy. On a single-line structured light platform, the mean Euclidean distance (ρ_{mean}) of the 3-D feature point coordinates in the CCF was 0.197 mm and 0.205 mm for the two test sets, further confirming the robustness of the proposed method.
2. In the context of welding seam classification, an approach was proposed that directly derived one-hot encoding for welding seam classification predictions from coarse-resolution features selected by the attention module. This method differed from the conventional method of upsampling to the high-resolution features for prediction. WeldLight achieved an F1-score of 0.9668 on the adjusted test set.
3. An online data augmentation method, focused on randomly generated splash noise, was proposed to adapt the model to natural splash noise arising from welding processes, thereby reducing its dependence on the quality of the dataset. Furthermore, a lightweight attention module, CCAM, was introduced as a feature selector to enhance the model's robustness against welding noise. The results show that WeldLight exhibited improved performance, particularly in high-noise scenes, upon incorporating CCAM.
4. The model was deployed on an EIC, utilizing the CPU for inference. The average processing speed for a single image was 29.32 ms, translating to 34.11 Hz, thereby satisfying the real-time requirements of the seam tracking system.

This study, rooted in real-world applications, reduces the computational and dataset requirements, thereby providing manufacturers with practical guidelines for deploying weld-seam tracking on edge devices without the need for costly GPUs. This capability not only lowers the deployment costs but also enhances automation and reduces the reliance on skilled labor.

Nevertheless, some limitations remain. Future research will focus on evaluating the model in more diverse welding scenarios and employing statistical methods to strengthen the reliability of the results. Moreover, the current architecture still has room for refinement; integrating advanced lightweight designs and exploring Transformer- or Mamba-based frameworks may further improve the performance. Finally, because the present approach still relies on manual annotation, future studies will investigate unsupervised and semi-supervised strategies to better adapt to evolving welding environments.

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

Funding: Central Funds Guiding Local Science and Technology Development (Grant No. YDZX2024115), the Key Research and Development Program of Shandong Province (Grant No. 2020CXGC010206), and the Shandong Provincial Natural Science Foundation (Project Nos. ZR2024QE326 and ZR2024QE494).

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 thank all colleagues who contributed to the discussions and writing of this manuscript. We also sincerely appreciate the anonymous reviewers for their constructive comments and suggestions, which greatly improved the quality of this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lei, T.; Wang, W.; Rong, Y.; Xiong, P.; Huang, Y. Cross-lines laser aided machine vision in tube-to-tubesheet welding for welding height control. *Opt. Laser Technol.* **2020**, *121*, 105796. [CrossRef]
2. Lei, T.; Wu, C.; Yu, H. Electric arc length control of circular seam in welding robot based on arc voltage sensing. *IEEE Sens. J.* **2022**, *22*, 3326–3333. [CrossRef]
3. He, Y.; Chen, Y.; Xu, Y.; Huang, Y.; Chen, S. Autonomous detection of weld seam profiles via a model of saliency-based visual attention for robotic arc welding. *J. Intell. Robot. Syst.* **2016**, *81*, 395–406. [CrossRef]
4. Wong, X.I.; Majji, M. A structured light system for relative navigation applications. *IEEE Sens. J.* **2016**, *16*, 6662–6679. [CrossRef]
5. Xiao, R.; Xu, Y.; Xu, F.; Hou, Z.; Zhang, H.; Chen, S. LSFP-tracker: An autonomous laser stripe feature point extraction algorithm based on siamese network for robotic welding seam tracking. *IEEE Trans. Ind. Electron.* **2023**, *71*, 1037–1048. [CrossRef]
6. Wu, C.; Yang, P.; Lei, T.; Zhu, D.; Zhou, Q.; Zhao, S. A teaching-free welding position guidance method for fillet weld based on laser vision sensing and EGM technology. *Optik* **2022**, *262*, 169291. [CrossRef]
7. Li, X.; Li, X.; Ge, S.S.; Khyam, M.O.; Luo, C. Automatic welding seam tracking and identification. *IEEE Trans. Ind. Electron.* **2017**, *64*, 7261–7271. [CrossRef]
8. 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: A review. *IEEE Sens. J.* **2018**, *18*, 8631–8641. [CrossRef]
9. Singh, A.; Kalaichelvi, V.; Karthikeyan, R. Application of convolutional neural network for classification and tracking of weld seam shapes for TAL Brabo manipulator. *Mater. Today Proc.* **2020**, *28*, 491–497. [CrossRef]
10. 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]
11. Deng, L.; Lei, T.; Wu, C.; Liu, Y.; Cao, S.; Zhao, S. A weld seam feature real-time extraction method of three typical welds based on target detection. *Measurement* **2023**, *207*, 112424. [CrossRef]
12. Liu, C.; Shen, J.; Hu, S.; Wu, D.; Zhang, C.; Yang, H. Seam tracking system based on laser vision and CGAN for robotic multi-layer and multi-pass MAG welding. *Eng. Appl. Artif. Intell.* **2022**, *116*, 105377. [CrossRef]
13. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30, Proceedings of the NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing System, Red Hook, NY, USA, 4–9 December 2017*; Curran Associates Inc.: Red Hook, NY, USA, 2017.
15. 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. [CrossRef]

16. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
17. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
18. Qin, D.; Leichner, C.; Delakis, M.; Fornoni, M.; Luo, S.; Yang, F.; Wang, W.; Banbury, C.; Ye, C.; Akin, B.; et al. MobileNetV4-Universal Models for the Mobile Ecosystem. *arXiv* **2024**, arXiv:2404.10518.
19. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
20. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
21. Yun, S.; Ro, Y. Shvit: Single-head vision transformer with memory efficient macro design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 5756–5767.
22. Wang, A.; Chen, H.; Lin, Z.; Han, J.; Ding, G. LSNet: See Large, Focus Small. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 10–17 June 2025; pp. 9718–9729.
23. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; Huang, Z. Efficient multi-scale attention module with cross-spatial learning. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rodos, Greece, 4–10 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
24. Si, Y.; Xu, H.; Zhu, X.; Zhang, W.; Dong, Y.; Chen, Y.; Li, H. SCSA: Exploring the synergistic effects between spatial and channel attention. *Neurocomputing* **2025**, *634*, 129866. [CrossRef]
25. Rahman, M.M.; Munir, M.; Marculescu, R. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 11769–11779.
26. Liu, T.; Wang, J.; Huang, X.; Lu, Y.; Bao, J. 3DSMDA-Net: An improved 3DCNN with separable structure and multi-dimensional attention for welding status recognition. *J. Manuf. Syst.* **2022**, *62*, 811–822. [CrossRef]
27. Ma, Y.; Fan, J.; Zhou, Z.; Zhao, S.; Jing, F.; Tan, M. WeldNet: A deep learning based method for weld seam type identification and initial point guidance. *Expert Syst. Appl.* **2024**, *238*, 121700. [CrossRef]
28. Zou, Y.; Zeng, G. Light-weight segmentation network based on SOLOv2 for weld seam feature extraction. *Measurement* **2023**, *208*, 112492. [CrossRef]
29. Liu, X.; Peng, H.; Zheng, N.; Yang, Y.; Hu, H.; Yuan, Y. EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 14420–14430.
30. Zhang, Z. A Flexible New Technique for Camera Calibration. *Microsoft Research Technical Report*; Microsoft Corporation: Redmond, WA, USA, 1998.
31. Fan, J.; Jing, F.; Fang, Z.; Liang, Z. A simple calibration method of structured light plane parameters for welding robots. In Proceedings of the 2016 35th Chinese Control Conference (CCC), Chengdu, China, 27–29 July 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 6127–6132.
32. Wang, S. Automated non-PPE detection on construction sites using YOLOv10 and transformer architectures for surveillance and body worn cameras with benchmark datasets. *Sci. Rep.* **2025**, *15*, 27043. [CrossRef] [PubMed]
33. Sonwane, S.; Chiddarwar, S. Enhancing weld defect detection and classification with MDCBNet: A multi-scale dense cross block network for improved explainability. *NDT E Int.* **2024**, *142*, 103029. [CrossRef]
34. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
35. Nwankpa, C. Activation functions: Comparison of trends in practice and research for deep learning. *arXiv* **2018**, arXiv:1811.03378. [CrossRef]
36. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
37. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
38. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
39. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
40. Shmueli, B. *Multiclass Metrics Made Simple, Part I: Precision and Recall;* Medium: San Francisco, CA, USA, 2020.

41. Dehghani, M.; Arnab, A.; Beyer, L.; Vaswani, A.; Tay, Y. The efficiency misnomer. *arXiv* **2021**, arXiv:2110.12894.
42. Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; Dosovitskiy, A. Do vision transformers see like convolutional neural networks? *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12116–12128.

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.