Engineering Applications of Artificial Intelligence 165 (2026) 113473

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Research paper

Deep multimodal fusion of spectral and visual data for laser welding defect classification

Qin Zhang [a], Zhongyou Zhao [a,c], Zhenmin Wang [b], Zixuan Wan [c], Hui-ping Wang [c], Guangze Li [c,*]

[a] School of Computer Science and Engineering, South China University of Technology, Guangzhou, 510006, China
[b] School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou, 510641, China
[c] Manufacturing Systems and Controls Research Lab, General Motors, Warren, 48092, USA

ARTICLE INFO

Keywords:
Laser welding
Artificial intelligence
Multimodal fusion
Image processing
Spectrum
Cross attention

ABSTRACT

Laser welding defect detection requires accurate interpretation of heterogeneous signals, in which weld images and spectral time-series data provide complementary information. However, effectively integrating these two types of data remains challenging. In this study, we construct a multimodal dataset for automotive battery busbar welding and propose a fusion framework based on cross-attention. Weld seams are first segmented using a convolutional network to suppress background interference, and informative spectral channels are selected through correlation analysis. Visual and spectral features are then jointly modeled by means of an inverted spectral embedding module and a vision-to-spectrum cross-attention mechanism, enabling fine-grained multimodal interaction. The proposed artificial intelligence method achieves an overall accuracy rate of 99.2%, which further improves to 100.0% with an increased spectral embedding dimension, outperforming all single-modality and baseline fusion approaches. Extensive ablation studies validate the benefits of segmentation, channel selection, and embedding design. Moreover, experiments on publicly available industrial defect datasets confirm the generalizability and robustness of our approach across diverse industrial defect inspection scenarios. To the best of our knowledge, this is the first work to apply cross-attention for fusing image and spectral data in laser welding, offering a novel and practical solution for multimodal industrial inspection.

1. Introduction

Laser welding technology is widely utilized in industries such as automotive, shipbuilding, and aerospace. However, the welding process is influenced by factors including environmental conditions, material properties, and machine performance. As a result, defects such as defocusing, gaps (Pan et al., 2016), and insufficient power may occur, compromising weld quality. These defects can significantly impact product reliability, lead to economic losses, and pose safety risks. Therefore, effective detection and mitigation of weld defects are essential to ensuring product integrity, reliability, and cost efficiency.

Laser welding inspection utilizes various data styles to ensure the quality and integrity of welds. High-speed imaging captures dynamic phenomena, such as the behavior of the welding pool and keyhole, while laser power and energy data ensures consistent energy delivery. Spectroscopic data, through emission spectrum analysis, helps monitor material composition and process stability. Visual and laser scanning data provide surface and dimensional accuracy, aiding in the detection of surface defects and ensuring weld precision. Process control data monitors parameters like focus, speed, and gas flow to maintain stable conditions, while post-weld inspection using methods like X-ray or ultrasonic testing ensures internal weld quality. Integrating different data styles for laser welding inspection can pose several challenges due to the complexity of handling multiple types of data and the need for precise synchronization. Existing fusion approaches (Zhang et al., 2019; She et al., 2024; He et al., 2025) for laser welding inspection often face challenges in effectively combining visual and spectral data due to their differing characteristics and inherent noise. Traditional methods (Medak et al., 2021; Hwang and Lee, 2024; Liang et al., 2024) typically focus on unimodal data or apply basic fusion techniques that fail to capture complex, cross-modal relationships. These methods struggle to simultaneously integrate time-series visual data and spectral signals, limiting their ability to detect subtle defects that require deep temporal and feature alignment.

In this work, we collected two typical data types from the battery busbar laser welding process: 2D weld seam images and welding spectral process signals, as shown in Fig. 1. Visual data captures surface

* Corresponding author.
E-mail address: guangze.li@gm.com (G. Li).

https://doi.org/10.1016/j.engappai.2025.113473
Received 21 May 2025; Received in revised form 5 November 2025; Accepted 4 December 2025
Available online 6 December 2025

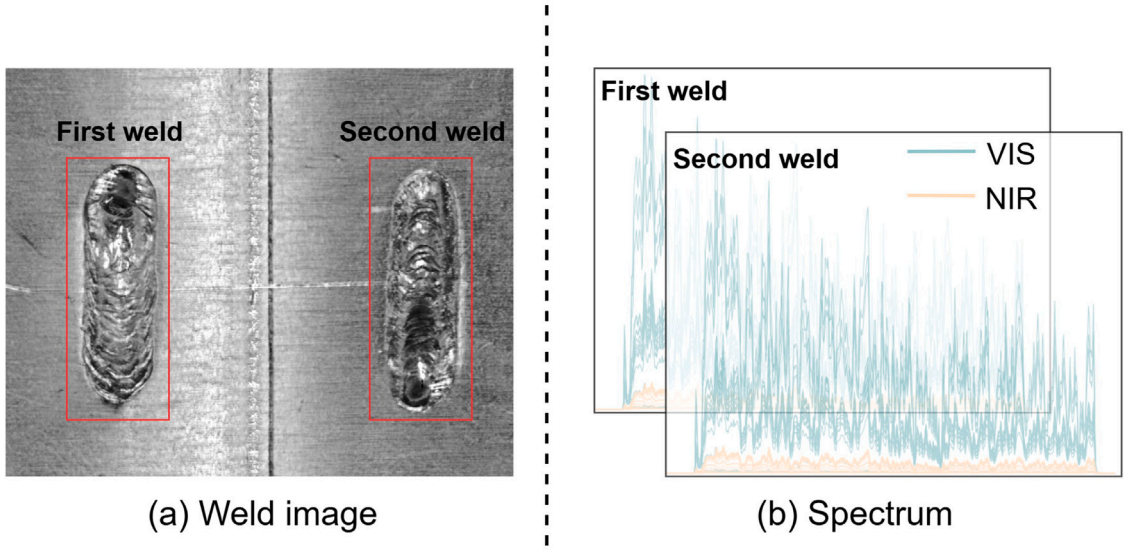


Fig. 1. Laser-welded battery busbar weld image and corresponding spectral signals. Each sample contains two weld seams, each associated with a segment of spectral data. The visual images provide geometric information of the welds, while the spectral signals, including visible (VIS) and near-infrared (NIR) bands, capture real-time plasma variations, temperature changes, and laser reflection fluctuations during the welding process.

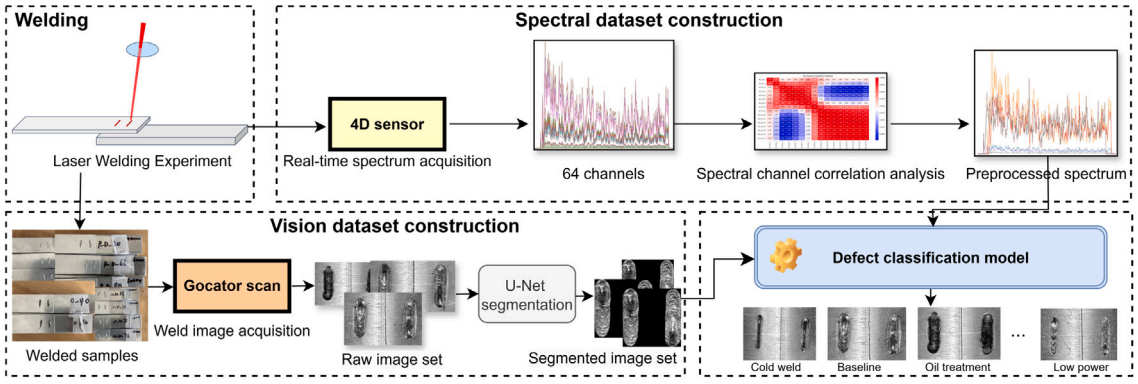


Fig. 2. Overall pipeline design. The experimental setup first performs laser welding under controlled parameters, during which a 4D sensor captures the real-time spectral signals and a Gocator scanner acquires corresponding visual images of the weld seam. The collected spectral data are preprocessed through correlation analysis for channel selection, while the weld regions in images are segmented using a U-Net model. Finally, a cross-modal fusion network integrates spectral and visual representations to classify different types of welding defects.

quality, seam geometry, spatter distribution, and keyhole dynamics, providing critical insights into weld integrity and process stability. It helps identify visible defects, irregularities, and cooling patterns that impact weld performance. Spectral data, including infrared, hyperspectral imaging, and optical emission spectroscopy, reveals thermal distribution, plasma characteristics, and material composition. This information aids in detecting hidden defects, assessing penetration depth, and monitoring real-time process stability at a molecular level.

Unlike traditional methods that rely on simple concatenation (He et al., 2025) or shallow fusion strategies (Zhang et al., 2019; She et al., 2024), our approach introduces a novel cross-attention mechanism to facilitate dynamic interaction between the visual and spectral modalities. This mechanism allows the model to selectively attend to relevant features across the modalities, mitigating the challenge of aligning heterogeneous data sources. By leveraging this attention-based (Vaswani et al., 2017) fusion, our method captures both fine-grained spatial information from images and deep temporal patterns from spectral signals, which significantly improves the accuracy of defect detection in complex welding scenarios.

In this study, we first perform a correlation analysis on the spectral signals to obtain a correlation matrix, based on which suitable spectral channels are selected as model inputs. To reduce noise in the welding images, we segment them to an appropriate size. The preprocessed spectral and image data are then fed into a deep learning model for defect detection, as shown in Fig. 2. To tackle multimodal data fusion challenges in laser welding inspection, we propose deep fusion method, leveraging a cross-attention mechanism (Chen et al., 2021) to integrate weld images and spectral signals. Cross-attention facilitates inter-modal interactions by enabling one modality to attend to another, proving highly effective for image time-series fusion. It dynamically weighs temporal information, aligns features across timestamps, and mitigates inconsistencies from varying conditions and has been widely used in tasks like vision-audio (Mo and Morgado, 2024) and vision-language (Radford et al., 2021) modeling. Thus, we design vision-spectrum cross-attention to fuse weld images and spectral signals for laser welding quality inspection. We conducted extensive experiments on the NEU (Bao et al., 2021) and DAGM (Wieler et al., 2007) datasets, which are widely recognized in the industrial inspection community for their challenging defect types and real-world relevance. Compared to traditional unimodal models, our cross-attention fusion strategy consistently outperforms these models across various evaluation metrics, confirming the superiority of our method in handling complex multimodal data for industrial defect detection.

The key contributions of this work are:

(1) A novel multimodal fusion method based on cross-attention is proposed for laser welding defect detection, enabling dynamic interaction between visual and spectral data for more accurate defect identification.
(2) A series of preprocessing techniques, including U-Net segmentation and Pearson correlation analysis, are applied to optimize the input data, ensuring effective model performance.
(3) Our cross-attention fusion model significantly outperforms both vision-only and spectrum-only models in terms of accuracy, demonstrating its capability in complex industrial inspection tasks.
(4) Extensive experiments on the NEU and DAGM datasets show that our fusion method is not only effective in welding defect detection but also generalizable to other industrial inspection domains.
(5) The proposed method is computationally efficient and suitable for real-time defect detection, making it practical for deployment in industrial environments.

The remainder of this article is structured as follows: Section 2 reviews related work. Section 3 introduces the dataset. Section 4 presents the proposed methodology. Section 5 describes the experimental setup and results. Finally, Section 6 concludes the study.

2. Related work

2.1. Machine learning-based methods

Machine Learning (ML) requires manually extracted features such as geometric, texture, and frequency domain characteristics. Engineers can clearly understand which features contribute to welding quality assessment. Malarvel and Singh (2021) classified welding defects by extracting features from X-ray images and applying an SVM. Khanzadeh et al. (2018) employed a dual-wavelength pyrometer and an infrared camera to capture the temperature distribution and morphological characteristics of the molten pool. Chen et al. (2018) captured molten pool images during welding using a high-speed camera and applied computer vision algorithms to extract geometric features. Linear Regression and Bagging Trees were then employed to predict the welding penetration based on the extracted features. Traditional machine learning approaches require the manual design of feature extraction modules, a process that is both time-consuming and labor-intensive. Additionally, determining the appropriate features to extract and selecting the most relevant ones after extraction remain significant challenges. While these ML-based methods provide interpretable results through handcrafted features, they rely heavily on domain expertise and cannot effectively exploit the joint dynamics of heterogeneous data sources such as visual and spectral signals. Consequently, their scalability and adaptability in complex welding scenarios are limited, highlighting the need for more advanced fusion strategies.

2.2. Deep learning-based methods

Deep learning utilizes multi-layer neural networks to automatically extract features, minimizing manual intervention and enhancing feature representation. Its ability to capture complex patterns enables effective adaptation to the dynamic nature of laser welding. Supported by large-scale data, deep neural networks achieve superior accuracy and robustness, ensuring reliable performance across varying welding conditions. Fan et al. (2021) proposed an intelligent laser welding defect detection method based on ACGAN-SVM-CNN, integrating data augmentation and multi-model fusion to enhance real-time detection accuracy. Knaak et al. (2021) employed k-Nearest Neighbors (kNN) and SVM in combination with CNN-GRU to enhance the model's welding classification performance. Deng et al. (2021) developed a deep learning network using CNN for feature extraction to detect welding defects. Gonzalez-Val et al. (2020) proposed a model based on CNN called ConvLBM, which utilizes high-speed Medium Wavelength Infrared (MWIR) to assess laser welding quality. Medak et al. (2021) introduced a deep learning-based defect detection method for ultrasonic scan image sequences, where features are extracted from each scan image and subsequently fused using standard convolutional layers (Conv2D) and Convolutional Long Short-Term Memory Networks (ConvLSTM) to enhance detection accuracy. Liang et al. (2024) proposed LAD-Net, a lightweight, attention-based non-destructive surface defect detection algorithm for welding. Designed to address the complexities of defect detection in ultrasonic welding caused by variations in welding parameters, equipment conditions, and operational techniques, LAD-Net integrates deformable convolution, stepwise attention, and multi-scale attention mechanisms to significantly enhance detection accuracy and efficiency. Hwang and Lee (2024) enhanced the feature representation of the welding region through edge detection and color filling and employed EfficientNet-B6 as the base model for transfer learning, significantly improving the classification accuracy of laser welding defects. Deep learning enables end-to-end learning, reducing process complexity and error propagation. With extensive data training, it enhances generalization, adapting to diverse welding scenarios. Although deep learning significantly reduces manual intervention and improves feature learning, most existing methods are unimodal, focusing either on visual images or on thermal/spectral data.

2.3. Multi-sensor methods

Multi-sensor fusion enhances welding quality inspection by integrating complementary data from visual, acoustic, thermal, and electrical signals, significantly improving defect detection accuracy and robustness. She et al. (2024) conducted real-time acquisition of keyhole and molten pool images along with spectral data during welding, using keyhole area and spectral intensity features to detect penetration depth. He et al. (2025) leveraged a magneto-optical infrared bi-imaging system, integrating multimodal fusion of magneto-optical and infrared image data with an attention-enhanced deep learning framework (AETMC-FCVT). Zhang et al. (2019) presented a deep learning-based approach for detecting welding defects during high-power disk laser welding of thick plates, utilizing a multi-sensor fusion system to extract key visual and spectral features. Multi-sensor fusion combined with deep learning significantly enhances welding defect detection by integrating visual, spectral, and thermal data for improved accuracy and robustness. These multi-sensor approaches demonstrate the benefits of integrating complementary modalities. However, their fusion is typically based on heuristic feature engineering or simple concatenation, which may not adequately align heterogeneous data representations. Such approaches cannot fully exploit temporal dynamics or selectively attend to the most informative inter-modal features, motivating the exploration of advanced attention-based fusion mechanisms.

2.4. Cross-attention methods

To the best of our knowledge, no prior study has applied cross-attention mechanism to fuse weld seam images and spectral signals in laser welding inspection. However, in the forefront of multimodal fusion research, cross-attention has been widely adopted. Mo and Morgado (2024) use cross-attention to update fusion tokens by attending to factorized audio-visual interactions. This mechanism allows the model to efficiently capture localized interactions between audio and visual modalities, enabling early fusion and improving performance on various audio-visual tasks. Shan et al. (2024) fuse quality-aware features from multiple views of a point cloud based on cross-attention mechanism. The semantic feature acts as the query, while the quality-aware features from different views act as the keys and values. Yi et al. (2024) proposed Text-IF model, cross-attention mechanism works by exchanging the queries of the two modalities and computing attention scores between the visible and infrared features. This allows the model to

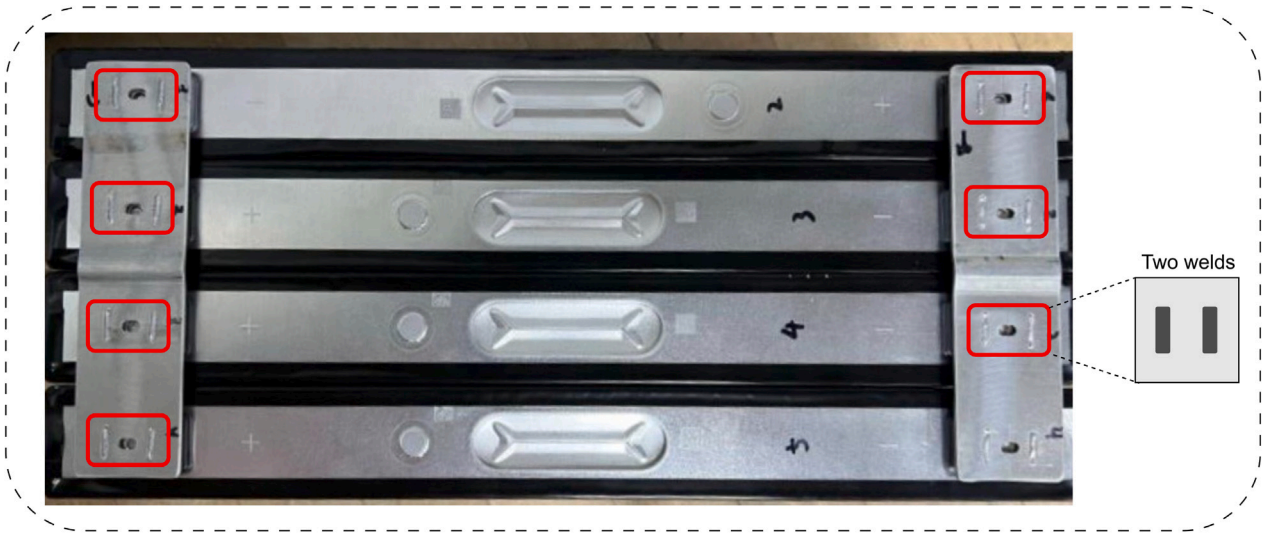


Fig. 3. Busbar welds of prismatic cells in a real welding scenario. Each welded connection consists of two individual welds, illustrating the typical geometry and arrangement encountered during laser welding of battery busbars.

effectively combine complementary information from both modalities, leading to high-quality fused images.

While cross-attention has been successfully applied in audio-visual, text-infrared, and 3D vision tasks, it has not yet been explored in the context of laser welding inspection. Our work is the first to design a vision-spectrum cross-attention framework tailored for welding defect detection. Unlike prior methods, our approach dynamically aligns heterogeneous modalities at multiple scales, incorporates correlation-based spectral channel selection, and demonstrates real-time feasibility, thereby addressing both accuracy and practicality in industrial inspection scenarios.

3. Data acquisition

In practical production, the welding target is the busbar weld of prismatic cells, as illustrated in Fig. 3. We designed welding coupons based on real application scenarios, as shown in Fig. 4. The experiments used 1.2 mm AI1100 as the busbar and AI3003 as the terminal material. Each coupon was welded twice, producing two individual welds. To investigate the impact of various factors on weld quality, we conducted controlled experiments under different conditions: defocus distances of ±4 mm and ±6 mm to induce defocus defects, gap sizes from 0 mm to 0.5 mm to produce gap-related defects, variations in laser power to simulate low-power defects, and different surface treatments (oil and water) to study their effects on welding quality. Moreover, when the gap or defocus exceeds a certain critical threshold, or when the laser power falls below a defined limit, cold weld defects are likely to occur. Based on the schedule, we have developed an in-house welding dataset comprising both spectral and visual image data. The dataset includes seven welding conditions: Baseline, Low Power, Low Gap, Defocus, Water Treatment, Oil Treatment, and Cold Weld, with their definitions detailed in Table 2. It is important to clarify the rationale for selecting these seven categories. The considered conditions—Baseline, Low Power, Low Gap, Defocus, Water Treatment, Oil Treatment, and Cold Weld—were deliberately chosen based on their prevalence and criticality in industrial battery busbar laser welding. These categories represent the most typical process-induced variations that directly influence joint formation and electrical/mechanical reliability. Other defect types, such as cracks and undercuts, are indeed relevant in general welding contexts, but they are rarely observed in prismatic cell busbar welding and often arise as secondary effects rather than primary process-induced flaws. Thus, our focus on these seven classes ensures both industrial relevance and feasibility of controlled data acquisition. Future work will extend the dataset to incorporate more complex defects such as cracks and undercuts. The dataset consists of 603 samples, which are split into training and testing sets at an 8:2 ratio.

Table 1
Configuration of the Gocator scanner. The operational parameters of the Gocator scanner were set according to the manufacturer's standard specifications to ensure accurate and consistent acquisition of weld surface profiles.

Parameters	Values
Resolution X (μm)	13.0–17.0
Z Linearity (μm)	1.5
Scan frequency (Hz)	1700
Field of View (FOV) (mm)	25.0–32.5

Table 2
Description of laser welding statuses for battery busbars. The dataset comprises seven representative welding categories, each corresponding to a specific process condition or defect type observed during laser welding.

Welding status	Definition
Baseline	The laser equipment operates at default setting, and the workpieces remains untreated.
Low Power	The welding power is set lower than the baseline.
Low Gap	There is a gap of less than 0.5 mm between the workpieces.
Defocus	The welding laser has a defocus of ±4 mm and ±6 mm.
Water treatment	The workpieces are cleaned with water before welding.
Oil treatment	The workpieces are cleaned with oil before welding .
Cold weld	The workpieces are not properly connected after welding.

3.1. Weld image acquisition

Weld images are captured using a Gocator visual scanner. The scanner settings are detailed in Table 1: the X-axis resolution is set to 13–17 μm, Z-axis linearity is 1.5 μm, scan frequency is 1.7 kHz, and the field of view is 25.0–32.5 mm.

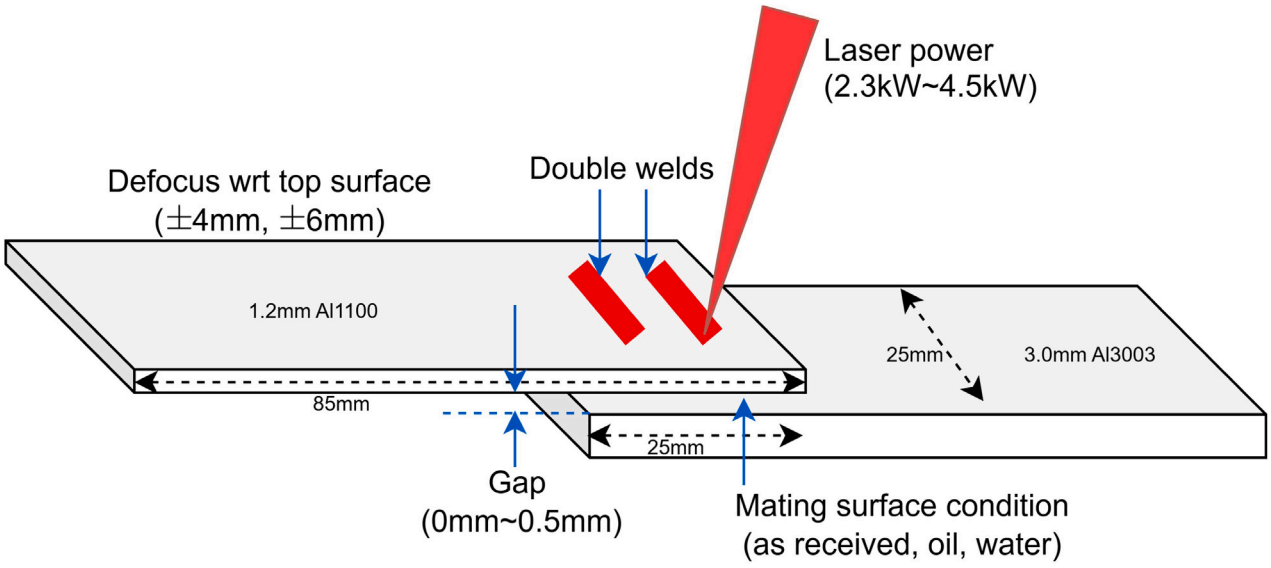


Fig. 4. Design of welding coupons and experimental parameters. The welding experiments were designed by systematically varying key process parameters, including defocus distance, laser power, inter-plate gap, and surface condition of the workpieces. These controlled variations ensure a representative dataset covering multiple defect types and welding conditions.

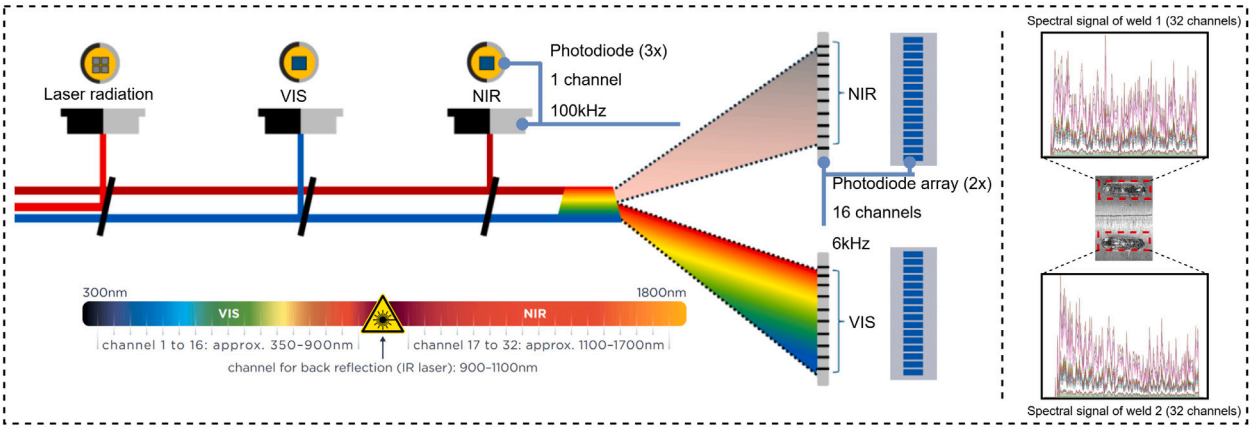


Fig. 5. Laser welding process spectral acquisition system. We employed a 4D sensor to capture multi-channel spectral signals. For each weld, a total of 32 spectral channels were recorded, including 16 visible and 16 near-infrared bands.

3.2. Spectrum acquisition

As shown in Fig. 5, the welding signal spectral acquisition system we use is the 4D.TWO. It captures full spectra simultaneously in both the visible (VIS) and near-infrared (NIR) ranges, utilizing 16 channels for VIS and 16 channels for NIR. In addition to these spectrally resolved channels, it includes 0th-order channels for VIS, NIR, and back reflection. This setup enables comprehensive spectral acquisition at sampling rates of up to 100 kHz. For each welding sample, we can obtain 32 channels of spectral signals of a single weld through the spectrometer, so for one welding sample we have a total of 64 channels of spectral information.

4. Method

In the forthcoming section, we analyze the spectral signals and 2D visual signals of laser welding, summarizing how different defect types affect the spectrum and weld images differently. Spectral correlation analysis is used to reduce the dimensionality of spectral inputs, thereby lowering the model's computational complexity. For weld images, we segment and crop the two weld seams, which are then separately paired with their corresponding spectral data as model inputs. Finally, we introduce a cross-attention-based deep learning mechanism to effectively fuse features from the two modalities. The following Algorithm 1 provides a structured overview of these steps, which are elaborated in detail in the subsequent sections.

4.1. Data preprocess

4.1.1. Channel dimensionality reduction

As shown in Fig. 5, after spectral acquisition, each weld seam yields 32 spectral channels: 16 visible light channels $C_v{}^i, i \in \{1, \ldots, 16\}$ and 16 near-infrared channels $C_n{}^j, j \in \{1, \ldots, 16\}$. These channels provide complementary information regarding the welding process, where visible light channels capture surface-level features such as weld geometry and spatter, while near-infrared channels reveal deeper material properties and thermal characteristics.

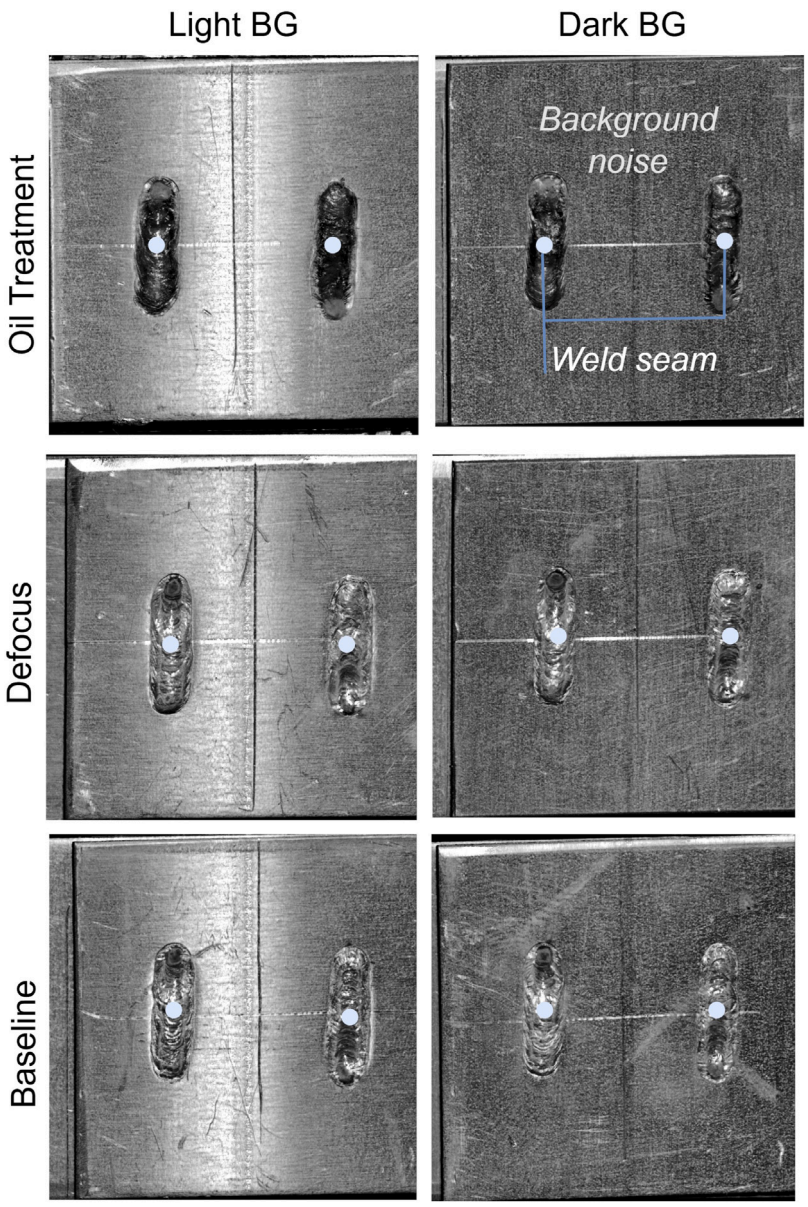


Fig. 6. Raw image data. The raw images exhibit inconsistent background noise such as scratches and surface marks, along with illumination variations that introduce additional noise (e.g., the same weld category appears brighter in the first column and darker in the second). The weld seams marked with blue dots indicate the target regions of interest (ROIs).

To reduce the dimensionality and select the most informative spectral channels, we apply *pearson correlation analysis* on the spectral data. This approach quantifies the linear relationship between pairs of spectral channels, helping identify redundant or highly correlated channels. The Pearson correlation coefficient between two channels x and y is calculated as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

where x_i and y_i represent the values of two spectral channels across n samples, and $\bar{x}$ and $\bar{y}$ are their respective means. The correlation coefficient r_{xy} ranges from -1 to 1, where values closer to 1 indicate strong positive correlation (i.e., redundancy), and values closer to 0 indicate weaker or no linear correlation. Channels with high correlation are considered redundant and removed, ensuring that only the most distinct and informative features are retained for further processing.

This dimensionality reduction strategy lowers computational cost and improves model efficiency, as redundant information is discarded while preserving the most representative spectral data. After computing the correlation matrix across all channels, we select a subset of channels that maximize the diversity of information. Specifically, we choose two visible light channels ${C_v}^i$, $i \in \{1, 3\}$, and two near-infrared channels ${C_n}^j$, $j \in \{1, 11\}$ from each weld seam, resulting in a total of four channels per seam. The corresponding wavelengths are:

$${C_v}^1 : 317 \text{ nm}, \quad {C_v}^3 : 393 \text{ nm}, \quad {C_n}^1 : 1017 \text{ nm}, \quad {C_n}^{11} : 1590 \text{ nm}$$

These selected channels represent a diverse set of wavelengths that capture both surface and deeper material characteristics, which are crucial for detecting various welding defects, such as surface irregularities, penetration depth variations, and thermal behavior. Finally, we set the length of the spectral time series to 560 time steps, ensuring that temporal variations in the spectral signals are adequately captured for subsequent analysis. This selection strategy helps reduce the feature space while maintaining the richness of the spectral data needed for accurate defect detection.

Algorithm 1: Image and Spectrum Preprocessing and Feature Extraction

Input: Image input $\mathbf{X}^v \in \mathbb{R}^{H\times W\times 3}$, Spectral data $\mathbf{X}^s \in \mathbb{R}^{N\times T}$
Output: Fused features $\mathbf{F}$
Data: Pre-trained vision encoder (e.g., MobileNetV2, ResNet50), Inverted embedding model
Step 1: Image Preprocessing
1. Resize input image to 512×512.
2. Normalize the image pixel values: $\mathbf{X}^v \rightarrow \text{Normalize}(\mathbf{X}^v)$.
Step 2: Vision Feature Extraction
3. Extract features from the preprocessed image using the vision encoder:
$\mathbf{V} = VisionEncoder(\mathbf{X}^v)$
where $\mathbf{V} \in \mathbb{R}^{L\times L\times C}$ is the feature map.
Step 3: Spectral Data Preprocessing
4. Normalize spectral data: $\mathbf{X}^s \rightarrow \text{Normalize}(\mathbf{X}^s)$.
5. Select relevant spectral channels based on correlation analysis:
$\mathbf{X}^s_{selected} \in \mathbb{R}^{N\times T}$.
Step 4: Spectrum Feature Extraction
6. Apply inverted embedding to the spectral data:
$\mathbf{S} = InvertedEmbedding(\mathbf{X}^s_{selected})$
where $\mathbf{S} \in \mathbb{R}^{N\times D_s}$ is the embedded feature.
Step 5: Feature Fusion
7. Fuse the visual and spectral features using cross-attention:
$\mathbf{F} = CrossAttention(\mathbf{V}, \mathbf{S})$.
Output: Fused feature representation $\mathbf{F}$ for further classification.

4.1.2. Weld image segmentation

As shown in Fig. 6, the raw scanned weld images contain not only the two target weld seams but also background information from the material surface. We observed that images of the same defect category may exhibit varying background brightness—for example, the first column shows brighter backgrounds, while the second column appears darker. Such variations can negatively affect the model's feature extraction. Additionally, surface noise such as scratches and other marks introduces irrelevant visual information, which may interfere with the model's judgment and degrade its overall performance. To address this issue, we introduce the U-Net (Ronneberger et al., 2015) deep learning model to segment the images and extract the weld seams, effectively removing background noise. This allows the model to focus more on feature extraction from the weld region. The architecture of the U-Net model is shown in Fig. 7. Each Conv Block ⟨channels, kernel size, stride, padding⟩ includes a BatchNorm layer and a ReLU activation function. The original weld image is processed by the U-Net to generate a weld seam mask, which is then used to segment the image and extract the region of interest, effectively removing background noise.

In Fig. 8, we present examples of different defect types. For defects such as (a), (b), and (c), the weld textures show clear differences, and their spectral signals exhibit unstable fluctuations with intensities that significantly deviate from the baseline. These defects can be easily distinguished from the baseline case (e) using either visual or spectral signals. However, for cases like (d) (defocus of 4 mm), the visual differences from the baseline are subtle, and the spectral fluctuations are also similar. Still, we observe a noticeable difference in spectral signal intensity compared to the baseline. This highlights that relying on a single modality may lead to difficulties in classifying certain defect types. Therefore, we use both spectral and image data for weld defect classification and explore three deep learning-based fusion methods: cross-attention, element-wise addition, and channel-wise concatenation.

4.2. Defect classification model

Our defect classification model is illustrated in Fig. 9. To ensure proper alignment between each weld seam and its corresponding spectral signal, we crop the input image into two separate welds during preprocessing and vertically flip the first weld seam. The processed weld images are then individually paired with their corresponding spectra and fed into deep learning feature extractors to obtain high-dimensional representations. For each weld seam, the extracted visual features and spectral sequence features are fused during the feature fusion stage. Finally, a self-attention projection layer is applied to the fused features to output the predicted defect class of the welding state.

4.2.1. Vision encoder

We employ deep learning-based feature extractors to obtain hierarchical representations of weld seams. Feature extraction is a critical step for capturing spatial information, such as edges, textures, and intensity variations that are essential for identifying weld defects. We investigate several commonly used convolutional neural network (CNN) architectures for this task, including VGG16 (Simonyan and Zisserman, 2014), MobileNetV2 (Sandler et al., 2018), GoogleNet (Szegedy et al., 2015), ResNet50 (He et al., 2016), and ViT-B/16 (Dosovitskiy et al., 2020). A comparison of these networks, based on their architecture and performance with a 512 × 512 input image resolution, is provided in Table 3. Among these, MobileNetV2 has the fewest parameters, making it more efficient for real-time applications, while VGG16 has the most parameters, offering more detailed feature extraction at the cost of computational efficiency.

All selected models are pre-trained on ImageNet-1K to capture general spatial features, enabling them to leverage pre-learned representations for weld seam detection. The input to the vision encoder is a raw image, $\mathbf{X}^v \in \mathbb{R}^{H\times W\times 3}$, where H and W are the height and width of the input image, and 3 represents the RGB color channels. The vision encoder transforms this input into a feature map, $\mathbf{V} \in \mathbb{R}^{L\times L\times C}$, where L is the spatial dimension of the output feature map, and C represents the number of channels in the feature map.

The encoding process can be mathematically described as:

$$\begin{aligned} &\mathbf{V} = VisionEncoder(\mathbf{X}^v) \\ &\text{where } VisionEncoder : \mathbf{X}^v \in \mathbb{R}^{H\times W\times 3} \rightarrow \mathbf{V} \in \mathbb{R}^{L\times L\times C} \end{aligned} \tag{2}$$

Here, (H, W) corresponds to the resolution of the original input image, L represents the output spatial resolution of the feature map, and C denotes the number of channels in the encoded representation.

In our experiments, we selected MobileNetV2 as the feature extractor for weld images. This choice is based on its balance between computational efficiency and feature extraction performance, making it suitable for real-time industrial applications where latency is a key consideration. MobileNetV2 is particularly advantageous due to its lightweight architecture, which provides fast inference times while maintaining sufficient accuracy for weld defect detection. The image encoder captures hierarchical representations at multiple scales, allowing it to identify fine-grained details such as weld seam edges, surface textures, and intensity variations. These features are crucial for identifying subtle weld defects, which might otherwise be missed using simpler or non-deep learning-based methods. Additionally, the encoded visual features are further processed in subsequent stages, where they are fused with spectral features for a comprehensive multimodal representation.

4.2.2. Spectrum embedding

During the laser welding process, real-time spectral data capture provides crucial insights into the dynamics of plasma, temperature fluctuations, and variations in laser reflection intensity, all of which are essential for evaluating weld quality. Spectral data offer valuable auxiliary information in our fusion framework, and we use them to enhance the detection of subtle defects in the weld seam. The feature

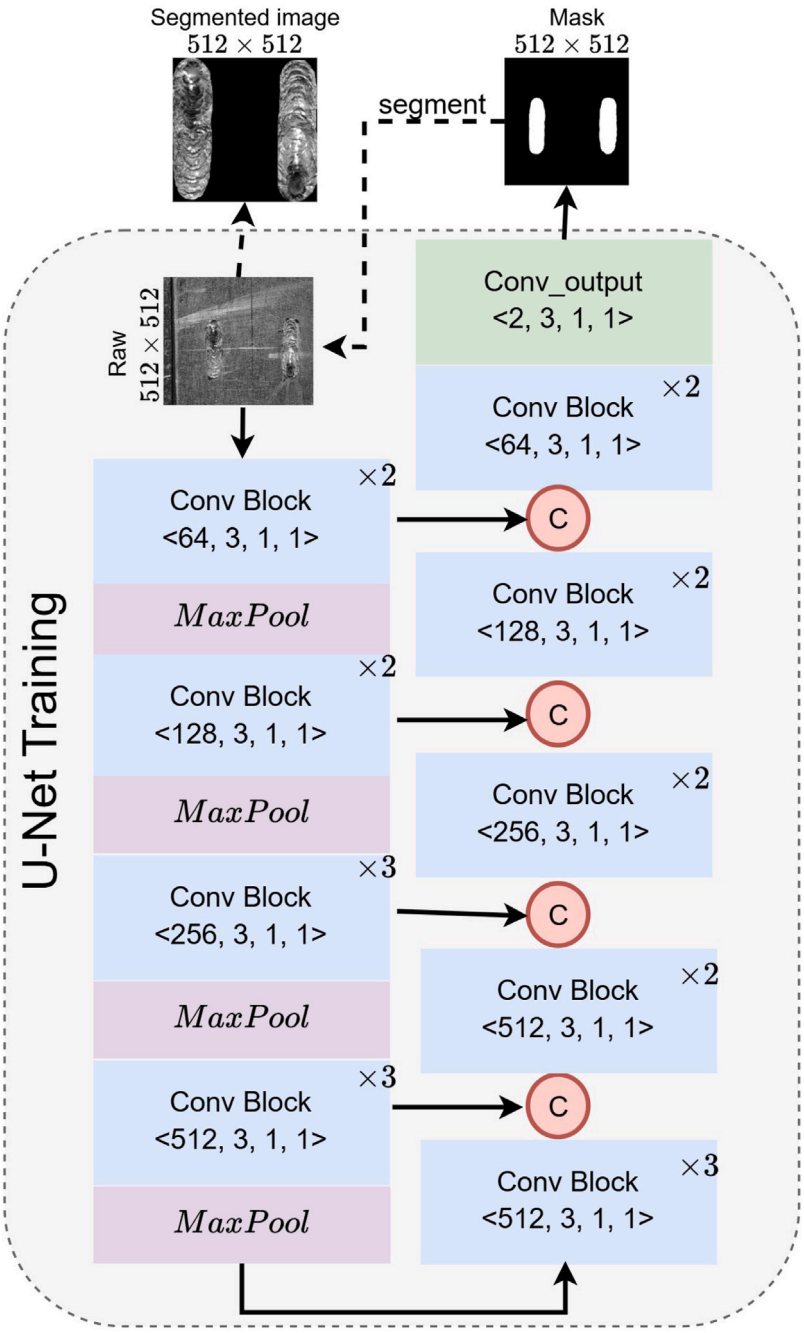


Fig. 7. U-Net framework for weld seam segmentation. A clear and compact U-Net architecture was designed to segment and extract the weld seams from the input images, facilitating subsequent multimodal fusion and defect classification.

Table 3
Network architectures comparison (512 × 512 Input).

Model	Key structural features	Output size	Params	FLOPs
VGG16	13 × [Conv3-64/128/256/512] 3 × FC Layers 5 × MaxPool (stride=2)	16 × 16 × 512 1000 Halves resolution	138M (FC dominates)	124B
MobileNetV2	53 Layers (Inverted Residuals) Depthwise Conv + Expansion	16 × 16 × 320	3.4M	6.5B
GoogleNet	9 × Inception Modules Parallel 1 × 1/3 × 3/5 × 5 Convs	16 × 16 × 832	6.8M	8.4B
ResNet50	50 Layers (Bottleneck Blocks) Skip Connections	16 × 16 × 2048	25.5M	38B
ViT-B/16	12 Transformer Encoders 32 × 32 Patches (vs 16 × 16@224)	1024 × 768 (Sequence Length)	86M	110B

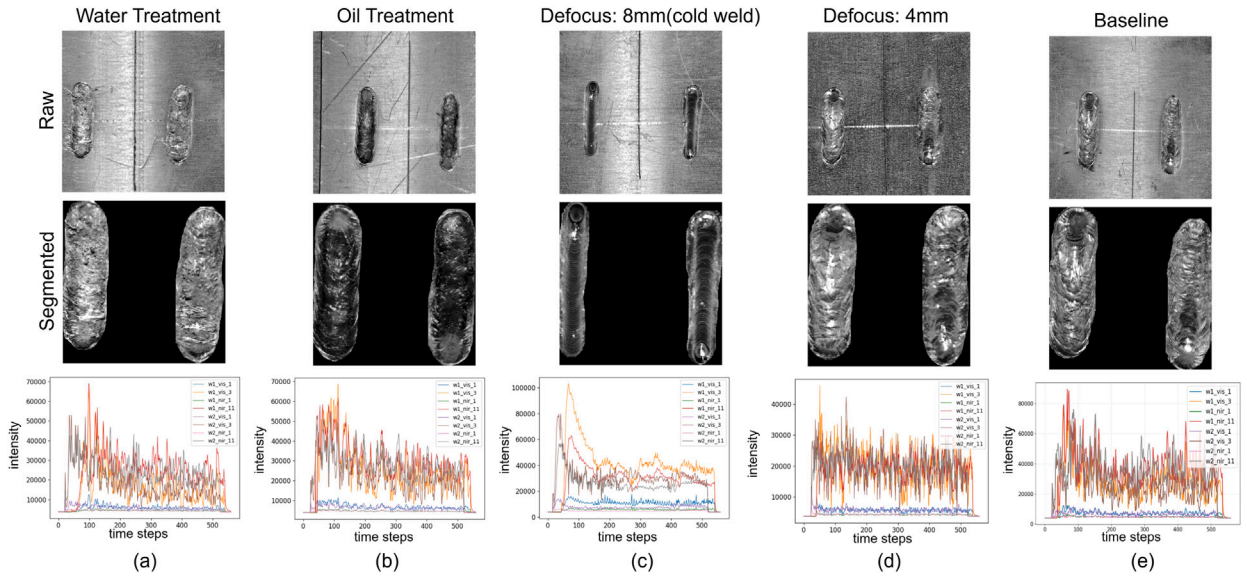


Fig. 8. Raw weld image, segmented weld image, and corresponding preprocessed spectral signal. After weld seam segmentation, the influence of background noise is removed. For different welding conditions, the spectral and visual data provide complementary feature information, enabling more accurate defect characterization.

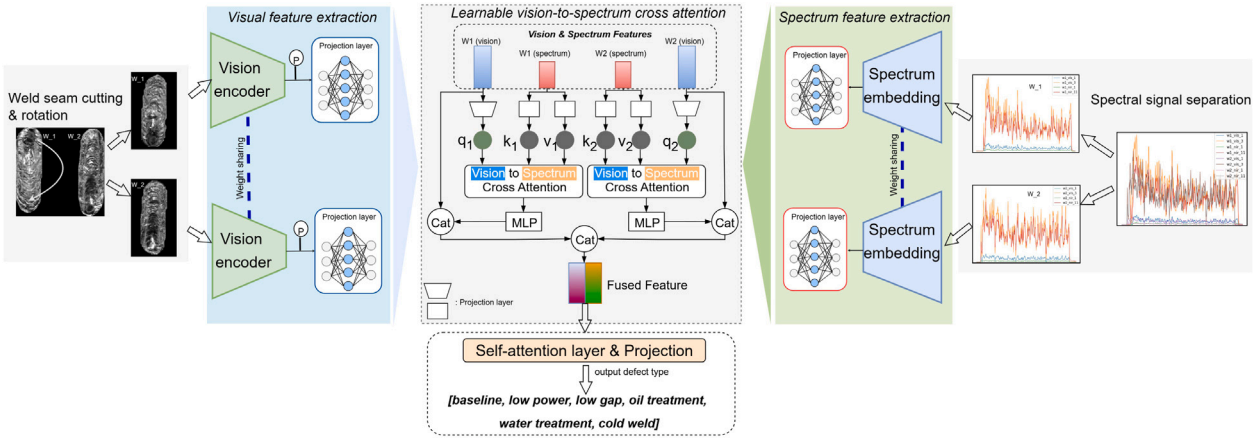


Fig. 9. Architecture of our fusion model. The weld images and spectral signals are first decoupled, and features are extracted separately. The spectral information corresponding to each weld is then integrated with the image features through a cross-attention module. Finally, a self-attention-based detection head outputs the defect classification results.

extraction process for spectral data involves a combination of temporal and channel-wise embeddings, which allows the model to capture both the time-dependent and multi-channel correlations.

In our system, we use a single embedding layer to encode spectral features, which efficiently integrates the spectral time series into a compact representation. Two common temporal embedding methods are typically considered for such tasks: traditional temporal embedding and inverted embedding. In traditional embedding, temporal tokens are used, where each token corresponds to a multivariate representation at a specific time step. On the other hand, Liu et al. (2023) introduce an inverted embedding strategy, where the time series data for each channel are independently embedded into channel tokens. This approach allows the attention module to capture the multivariate correlations across different spectral channels, while the subsequent feed-forward network (FFN) encodes these series representations.

As is shown in Fig. 10, we adopt the inverted embedding approach as it enables more effective capture of channel-wise dependencies. The spectrum embedding process is described as follows:

$$\begin{aligned} &\mathbf{s}_n = Inverted\,Embedding(\mathbf{X}^s_{:,n}) \quad \text{for each channel } n, \\ &\mathbf{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_N\} \in \mathbb{R}^{N \times D_s}, \\ &Inverted\,Embedding : \mathbb{R}^{N \times T} \rightarrow \mathbb{R}^{N \times D_s} \end{aligned} \tag{3}$$

where $\mathbf{X}^s_{:,n}$ denotes the spectral time-series data for the nth spectral channel, and $\mathbf{s}_n$ is the corresponding embedded token. The embedding process maps each time-series from T time steps into a D_s-dimensional space, resulting in N embedded tokens that are used to capture both temporal and spectral dependencies. The dimension D_s represents the size of the output embedding for each channel.

The spectral feature extraction is performed using a multi-layer perceptron (MLP), which processes the raw spectral time series data and projects it into a higher-dimensional space. This helps preserve the most important temporal and spectral features, enabling the subsequent fusion step to leverage both spatial (visual) and spectral information effectively. This spectrum embedding method enhances the model's

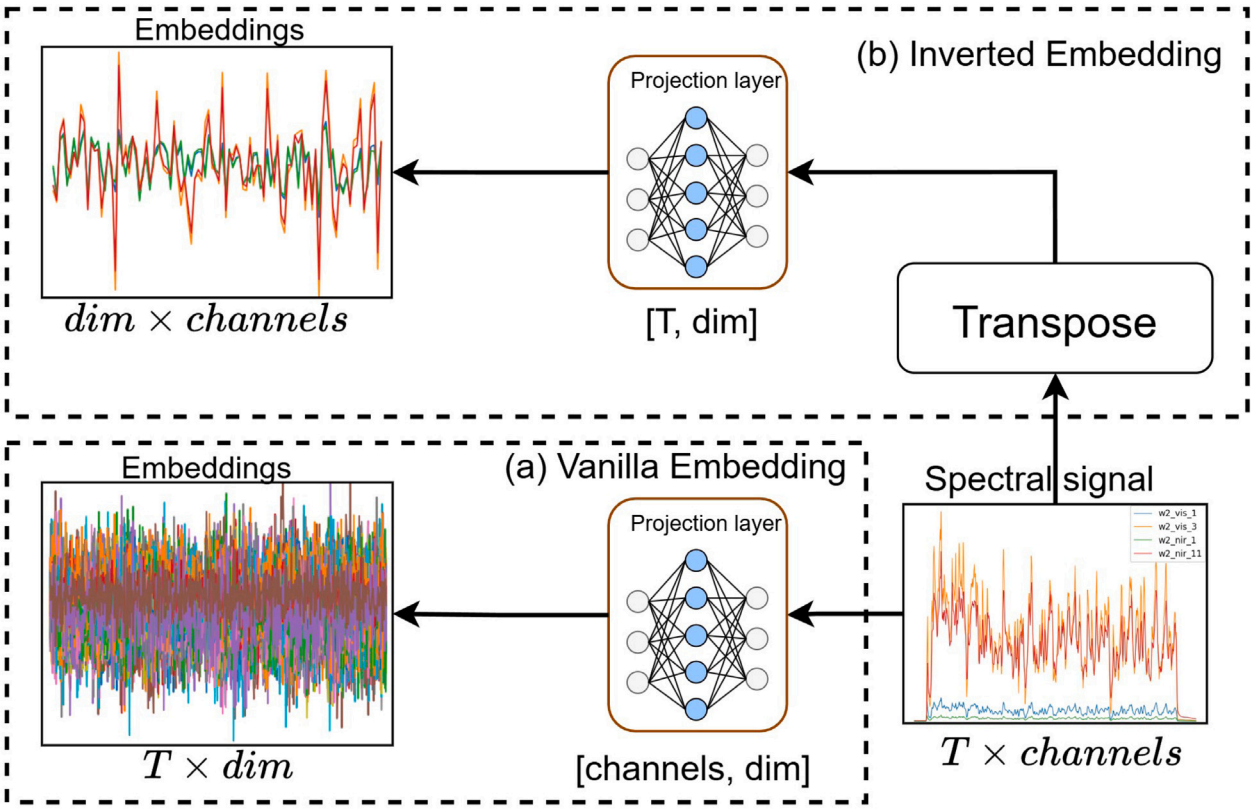


Fig. 10. Spectrum embedding methods. We employ inverted embedding to extract spectral features, treating each variable sequence as a token. This approach prevents forcibly correlating visible and near-infrared channels of different scales, thereby preserving the independence of each spectral channel.

ability to capture temporal dynamics and multi-channel relationships, which are essential for detecting hidden defects in the welding process. The encoded spectral features are then fused with visual features from the weld seam images, forming the foundation for the cross-attention-based multimodal fusion mechanism.

4.2.3. Fusion methods

In Fig. 9, we propose a learnable cross-attention-based fusion mechanism that enables effective interaction from visual features to spectral representations, allowing the model to adaptively align and integrate multimodal information. Unlike simple feature-level fusion, this design explicitly models modality-specific dependencies through attention.

Feature representation. Let the visual features be denoted as $\mathbf{X}_{\text{img}} \in \mathbb{R}^{C\times L\times L}$ and the spectral features as $\mathbf{X}_{\text{spec}} \in \mathbb{R}^{D\times C}$, where C is the channel dimension, $L\times L$ is the spatial resolution, and D is the spectral sequence length. We flatten the spatial dimensions and obtain:

$$\mathbf{X}_{\text{img}} \in \mathbb{R}^{L^2\times C}, \quad \mathbf{X}_{\text{spec}} \in \mathbb{R}^{D\times C}. \tag{4}$$

Multi-head cross-attention. To enable diverse representation learning, we employ multi-head cross-attention with h heads. Each head projects the inputs into lower-dimensional subspaces:

$$\mathbf{Q}_i = \mathbf{X}_{\text{img}}\mathbf{W}_Q^{(i)}, \quad \mathbf{K}_i = \mathbf{X}_{\text{spec}}\mathbf{W}_K^{(i)}, \quad \mathbf{V}_i = \mathbf{X}_{\text{spec}}\mathbf{W}_V^{(i)}, \tag{5}$$

where $\mathbf{W}_Q^{(i)}, \mathbf{W}_K^{(i)}, \mathbf{W}_V^{(i)} \in \mathbb{R}^{C\times d_h}$ and $d_h = d/h$ is the sub-dimension per head. The attention of head i is then computed as:

$$\mathbf{A}_i = \text{softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^\top}{\sqrt{d_h}}\right)\mathbf{V}_i \in \mathbb{R}^{L^2\times d_h}. \tag{6}$$

The outputs of all heads are concatenated and projected:

$$\mathbf{F}_{\text{attn}} = \text{Concat}(\mathbf{A}_1, \ldots, \mathbf{A}_h)\mathbf{W}_O, \quad \mathbf{W}_O \in \mathbb{R}^{d\times C}. \tag{7}$$

Residual connection and layer normalization. The fused feature is stabilized with residual learning and normalization:

$$\mathbf{F}_{\text{fused}} = \text{LayerNorm}(\mathbf{F}_{\text{attn}} + \mathbf{X}_{\text{img}}). \tag{8}$$

Why vision-to-spectrum cross-attention? In our implementation, the query represents the spatially encoded visual features of the weld region, while the keys and values correspond to spectral embeddings that carry information about plasma emission, temperature fluctuation, and laser reflection intensity. Therefore, the attention weights describe how visual regions correlate with their spectral responses during the welding process. We use image features as queries and spectral features as keys/values. This design allows each image pixel to dynamically attend to the full spectral sequence, enabling the model to exploit material-sensitive cues from spectra to refine spatial defect localization. In contrast: - **Spectrum-to-vision cross-attention:** Queries from spectra focus on attending to spatial locations, which is less intuitive because spectra lack strong spatial priors. - **Bidirectional cross-attention:** While possible, it doubles computational cost and may cause redundancy, offering marginal performance gains in our experiments. Comparison of cross-attention variants:

$$\begin{aligned}
&\text{Vision-to-Spectrum:} \quad \mathbf{Q} = \mathbf{X}_{\text{img}}, \ \mathbf{K}, \mathbf{V} = \mathbf{X}_{\text{spec}}.\\
&\text{Spectrum-to-Vision:} \quad \mathbf{Q} = \mathbf{X}_{\text{spec}}, \ \mathbf{K}, \mathbf{V} = \mathbf{X}_{\text{img}}.\\
&\text{Bidirectional:} \quad \mathbf{F}_{\text{fused}} = \text{CrossAttn}(\mathbf{X}_{\text{img}} \leftarrow \mathbf{X}_{\text{spec}}) + \text{CrossAttn}(\mathbf{X}_{\text{spec}} \leftarrow \mathbf{X}_{\text{img}}).
\end{aligned} \tag{9}$$

For comparison, we also implemented two commonly used fusion baselines:

Element-wise addition:

$$\mathbf{F}_{\text{fused}} = \mathbf{X}_{\text{img}} + \mathbf{X}_{\text{spec}}. \tag{10}$$

Channel-wise concatenation:

$$\mathbf{F}_{\text{fused}} = \text{Concat}(\mathbf{X}_{\text{img}}, \mathbf{X}_{\text{spec}}) \in \mathbb{R}^{L^2 \times 2C}. \tag{11}$$

These methods are computationally simpler but lack the adaptive modeling ability of cross-attention.

Algorithmic pseudocode. To further clarify, the following pseudocode provides implementation sketches of three fusion strategies:

Algorithm 2: Vision-to-Spectrum Cross-Attention Fusion

Input: Image features $\mathbf{X}_{\text{img}}$, Spectral features $\mathbf{X}_{\text{spec}}$
Output: Fused features $\mathbf{F}_{\text{fused}}$
Project $\mathbf{X}_{\text{img}}$ into queries $\mathbf{Q}$;
Project $\mathbf{X}_{\text{spec}}$ into keys $\mathbf{K}$ and values $\mathbf{V}$;
Compute attention weights: $\mathbf{P} = \text{softmax}(\mathbf{QK}^\top/\sqrt{d})$;
Compute output: $\mathbf{F}_{\text{attn}} = \mathbf{PV}$;
Apply residual + normalization:
$\mathbf{F}_{\text{fused}} = \text{LayerNorm}(\mathbf{F}_{\text{attn}} + \mathbf{X}_{\text{img}})$;

Algorithm 3: Element-wise Addition Fusion

Input: $\mathbf{X}_{\text{img}}, \mathbf{X}_{\text{spec}}$ with equal shape
Output: $\mathbf{F}_{\text{fused}}$
for *each position* i **do**
  $\mathbf{F}_{\text{fused}}[i] = \mathbf{X}_{\text{img}}[i] + \mathbf{X}_{\text{spec}}[i]$
end

Algorithm 4: Channel-wise Concatenation Fusion

Input: $\mathbf{X}_{\text{img}}, \mathbf{X}_{\text{spec}}$
Output: $\mathbf{F}_{\text{fused}}$
Stack features along channel dimension:;
$\mathbf{F}_{\text{fused}} = \text{Concat}(\mathbf{X}_{\text{img}}, \mathbf{X}_{\text{spec}})$;

In summary, the proposed vision-to-spectrum cross-attention enables fine-grained alignment by letting each pixel attend to the full spectral sequence. Compared with addition or concatenation, this approach adaptively integrates complementary multimodal information, enhancing defect classification performance.

4.3. Detection head

After the fusion stage, we obtain a unified feature representation $\mathbf{F}_{\text{fused}} \in \mathbb{R}^{B\times N\times d}$, where $N = L^2$ represents the number of visual tokens, and d is the feature dimension. To aggregate global contextual information across all fused tokens, we apply a self-attention layer over $\mathbf{F}_{\text{fused}}$. This is implemented as a standard multi-head self-attention (MHSA) mechanism:

$$\mathbf{Q} = \mathbf{F}_{\text{fused}}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{F}_{\text{fused}}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{F}_{\text{fused}}\mathbf{W}_V,$$
$$\mathbf{F}_{\text{attn}} = \text{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d}}\right)\mathbf{V} \tag{12}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d}$ are learnable projection matrices. The output $\mathbf{F}_{\text{attn}} \in \mathbb{R}^{B\times N\times d}$ is passed through a layer normalization and feedforward network to produce the final aggregated representation. To perform defect classification, we apply a global average pooling over the token dimension followed by a linear projection:

$$\mathbf{f}_{\text{global}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{F}_{\text{attn}}^{(i)}, \quad \hat{\mathbf{y}} = \text{softmax}(\mathbf{W}_{\text{cls}}\mathbf{f}_{\text{global}} + \mathbf{b}) \tag{13}$$

where $\mathbf{W}_{\text{cls}} \in \mathbb{R}^{d\times C}$ and $\mathbf{b} \in \mathbb{R}^C$ are the classification weights and bias, and C is the number of defect categories. This final projection maps the global fused representation to a probability distribution over possible defect types.

4.4. Loss function

Focal Loss is an extension of Cross-Entropy Loss designed to address class imbalance by reducing the impact of well-classified samples and focusing more on hard, misclassified examples. In our work, The focal loss $\mathcal{L}_{FL}$ is described as:

$$\mathcal{L}_{FL} = -\sum_{i=1}^{C}(1-\tilde{y}_i)^{\lambda} y_i log(\tilde{y}_i) \tag{14}$$

where λ is a tunable focusing parameter. This mechanism ensures that the model prioritizes learning from difficult cases, such as rare or subtle defect types in weld classification. The main advantage of Focal Loss is its ability to mitigate class imbalance without requiring explicit data re-sampling, thereby improving model robustness and generalization.

5. Experiment

In this section, we present extensive experiments to validate the effectiveness of the proposed defect classification model. First, we assess the advantages of our cross attention-based fusion strategy compared to two alternative fusion methods, as shown in Table 4. Next, we benchmark our approach against several vision-only and spectrum-only time-series models, with results summarized in Table 6. In addition, we calculate the number of parameters, FLOPs, and FPS of the proposed model to evaluate its real-time performance. And we conduct a series of ablation studies to evaluate (i) the impact of using segmented versus raw images, (ii) the attention localization performance, and (iii) the effectiveness of spectral preprocessing techniques. Finally, we extend our proposed cross-attention fusion strategy to two additional industrial defect datasets—NEU and DAGM—by performing multiscale feature fusion experiments to further validate its generalizability and effectiveness.

5.1. Implementation details

Our experiments were conducted using PyTorch on a Windows 10 system equipped with an NVIDIA RTX A5500 GPU. The welding images were resized to 512 × 512 pixels and normalized to the $(0, 1)$ range. Spectral signals were normalized using min–max normalization. The spectrum embedding approach is described in Section 4.2.2, and its associated hyperparameters are listed in Table 5. We employed the AdamW optimizer with an initial learning rate of 2×10^{-4}, a batch size of 24, and trained the model for 100 epochs.

5.2. Metrics

To rigorously evaluate the classification performance of our defect detection model, we adopt four standard metrics: precision, recall, F1-score, and area under the ROC curve (AUC). Precision is defined as the ratio of true positive predictions (TP) to the total number of positive predictions, capturing the model's ability to avoid false positives:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

Recall, also known as sensitivity, measures the proportion of actual positive samples correctly identified by the model:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

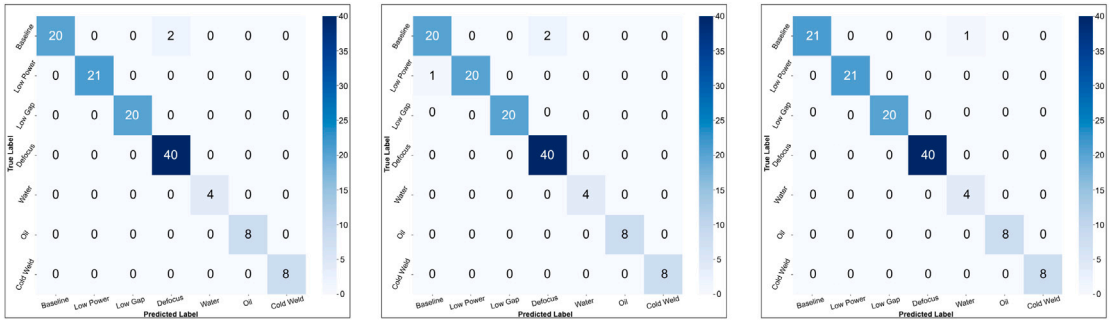
The F1-score is the harmonic mean of precision and recall, balancing both metrics in a single value:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

In addition, we compute the Area Under the Receiver Operating Characteristic Curve (AUC), which evaluates the model's ability to distinguish between classes across varying classification thresholds. AUC is particularly useful in assessing models under class imbalance and offers a global perspective on classification performance.

Table 4
Classification precision, recall, F1-score and AUC evaluation metrics using three feature fusion methods.

Defect type	Element-wise addition				Channel-wise concatenation				Cross attention (Ours)			
	Pre(%)	Rec(%)	F1(%)	AUC(%)	Pre(%)	Rec(%)	F1(%)	AUC(%)	Pre(%)	Rec(%)	F1(%)	AUC(%)
Baseline	100.0	91.0	95.2	95.6	95.2	90.9	93.0	95.0	100.0	95.6	**97.8**	**97.7**
Low power	100.0	100.0	**100.0**	**100.0**	100.0	95.2	97.6	97.6	100.0	100.0	**100.0**	**100.0**
Low gap	100.0	100.0	**100.0**	**100.0**	100.0	100.0	**100.0**	**100.0**	100.0	100.0	**100.0**	**100.0**
Defocus	95.2	100.0	97.6	98.8	95.2	100.0	97.6	98.8	100.0	100.0	**100.0**	**100.0**
Water treatment	100.0	100.0	**100.0**	**100.0**	100.0	100.0	**100.0**	**100.0**	80.0	100.0	88.9	99.6
Oil treatment	100.0	100.0	**100.0**	**100.0**	100.0	100.0	**100.0**	**100.0**	100.0	100.0	**100.0**	**100.0**
Cold weld	100.0	100.0	**100.0**	**100.0**	100.0	100.0	**100.0**	**100.0**	100.0	100.0	**100.0**	**100.0**



(a) Element-wise addition (b) Channel-wise concatenation (c) Cross attention

Fig. 11. Confusion matrix visualization of three feature fusion methods. (a) Element-wise addition and (b) Channel-wise concatenation both exhibit confusion between challenging classes such as baseline and defocus. In contrast, our proposed method (c) Cross-attention fusion achieves a clear distinction between these categories.

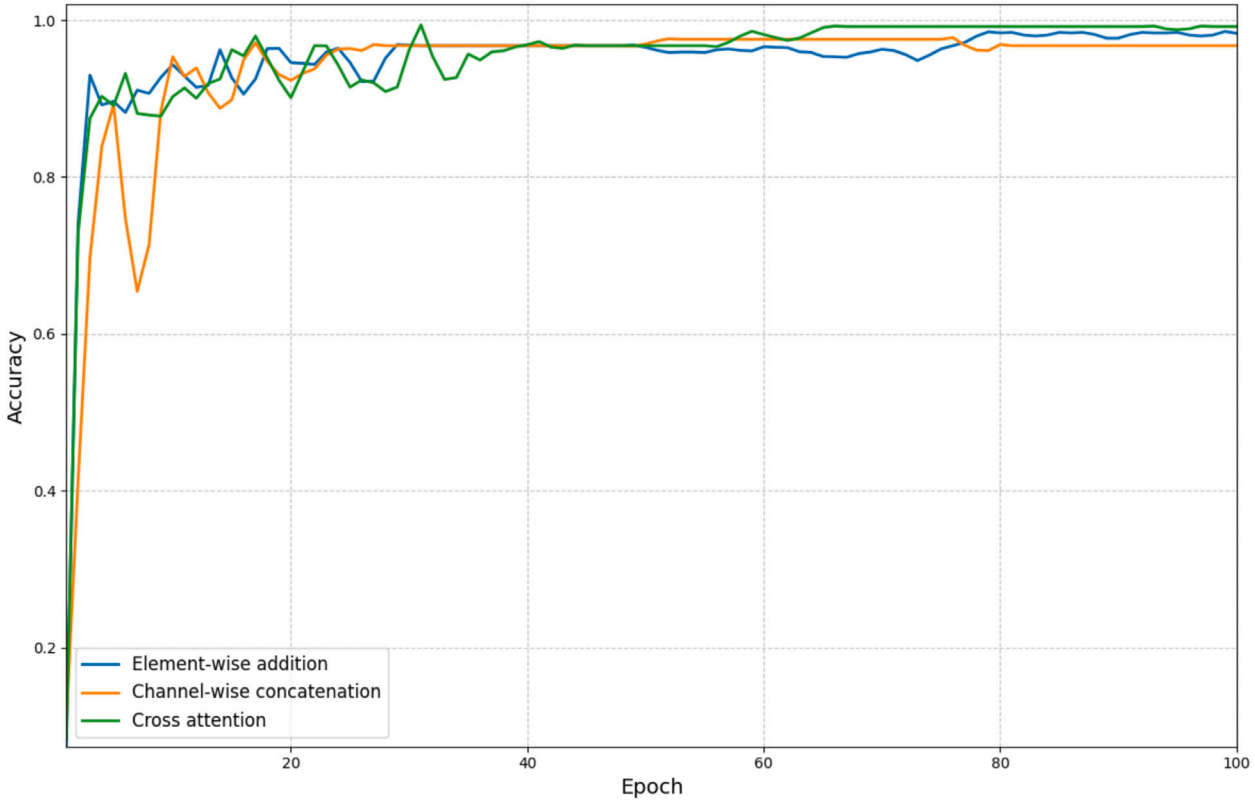


Fig. 12. Accuracy comparison of three feature fusion methods. Our proposed Cross-attention method (green curve) demonstrates more stable convergence during training and achieves the highest final classification accuracy among all compared methods.

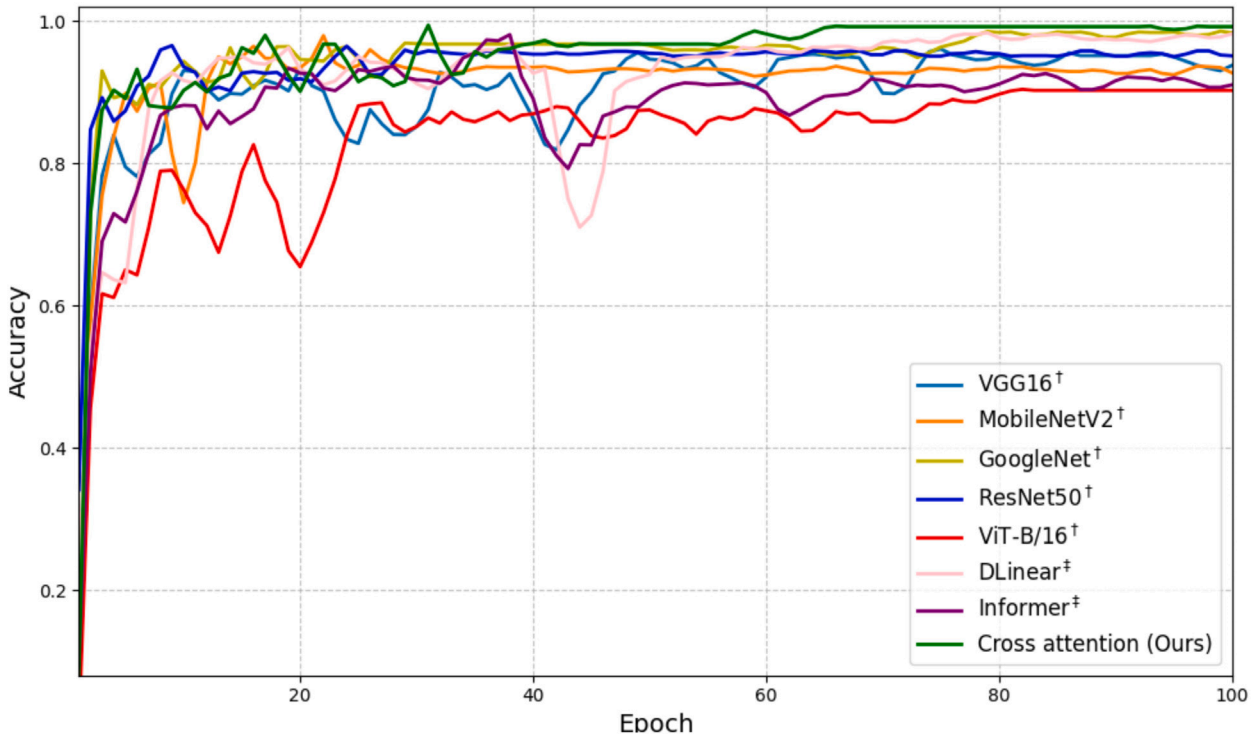


Fig. 13. Accuracy curves during training for the vision-only (†), spectrum-only (‡), and our proposed cross-attention fusion method. Our proposed cross-attention fusion approach (green curve) achieves the highest convergence accuracy, demonstrating the effectiveness of multimodal feature interaction.

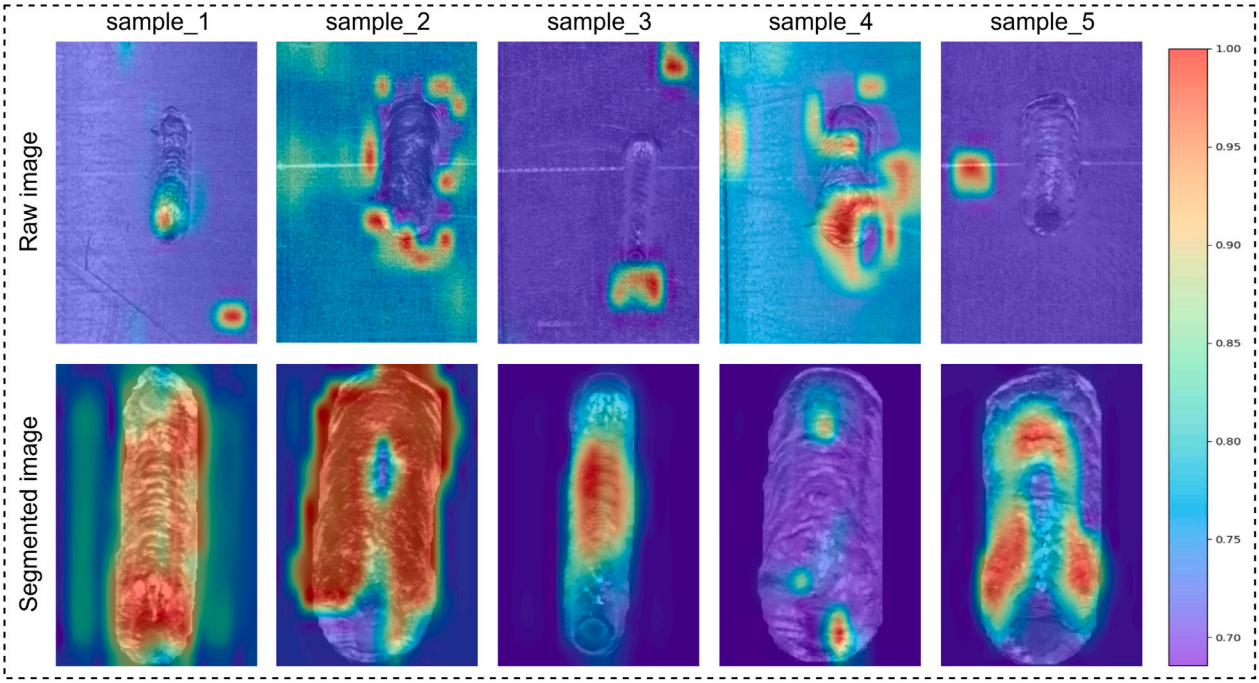


Fig. 14. Attention localization comparison between raw weld images and segmented weld images using the proposed cross-attention mechanism. As shown in the second row, the trained model effectively focuses on different weld seam regions corresponding to specific defect states. In contrast, the first row demonstrates that using unsegmented raw images causes the attention to shift away from the true defect regions, confirming the necessity of weld seam segmentation.

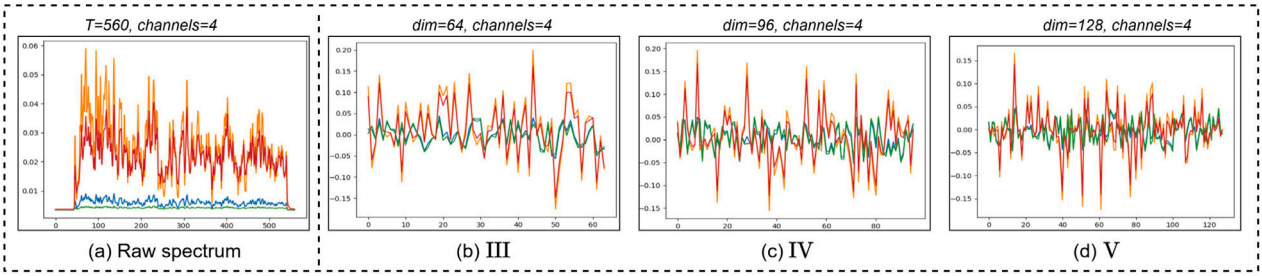


Fig. 15. Learned spectral patterns under different embedding dimensions, illustrating the model's ability to capture diverse spectral features as the projection dimension varies. As the embedding dimension increases, the spectral representations become more semantically enriched, enabling the model to capture finer and more distinctive spectral characteristics across different defect types.

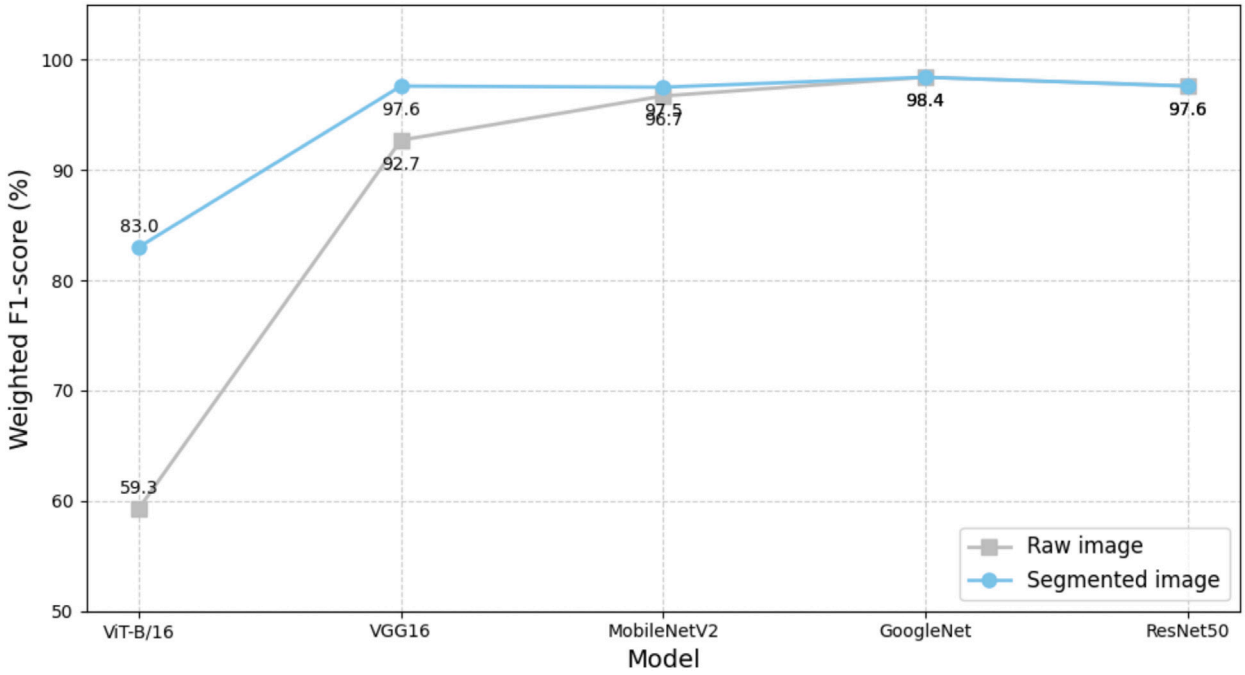


Fig. 16. Comparison of weighted F1-scores between segmented weld images and raw images across different vision-only models. After segmentation preprocessing, the performance of nearly all vision-only models shows a clear improvement, demonstrating the effectiveness of weld seam segmentation in reducing background noise and enhancing discriminative feature learning.

Table 5
Embedding hyperparameters of a weld's spectral signal.

Hyperparameter	Values
T	560
dim	96
channels	$vis_1, vis_3, nir_1, nir_{11}$

5.3. Comparisons with three fusion methods

We first employed the lightweight MobileNetV2 as the vision encoder to conduct comparative experiments across the three fusion methods. Table 4 presents quantitative metrics—including precision, recall, F1-score, and AUC—for each defect category. All three methods achieved 100% classification accuracy for low gap, oil treatment, and cold weld defects. As discussed in Section 4.1.2, baseline and defocus are considered challenging samples. In these categories, both element-wise addition and channel-wise concatenation underperformed compared to the cross-attention-based fusion. For the baseline category, the F1-score of the cross-attention method reached 97.8%, outperforming element-wise addition (95.2%) and channel-wise concatenation (93.0%) by 2.6% and 4.8%, respectively. Similarly, the AUC for cross-attention was 97.7%, exceeding the other two methods by 2.1% and 2.7%. For the defocus category, cross-attention achieved perfect scores with 100% in both F1 and AUC, surpassing the other two methods by 2.4% in F1 and 1.2% in AUC.

As shown in the confusion matrix in Fig. 11, the element-wise addition method misclassified two baseline samples as defocus, and the channel-wise concatenation method exhibited the same misclassification. In contrast, the cross-attention fusion method did not produce such errors. Across the entire test set, only one baseline sample was misclassified when using the cross-attention method. Additionally, Fig. 12 illustrates the overall training accuracy curves. It can be observed that after convergence, the cross-attention method (green curve) consistently outperforms the other two methods in terms of classification accuracy.

5.4. Comparisons with spectrum-only and vision-only models

To highlight the advantages of our multi-source signal fusion approach, we conducted comparative experiments against commonly used unimodal models based on visual and sequential backbones. For

Table 6
Comparison of weighted F1-score, weighted precision, and weighted recall between the proposed fusion-based method and spectrum-only and vision-only baselines.

Metrics	Vision-only models (†)					Spectrum-only models (‡)		Ours
	VGG16	MobileNetV2	GoogleNet	ResNet50	ViT-B/16	Informer	DLinear	Cross attention (Fusion)
Weighted Pre. (%)	97.7	97.5	98.5	97.9	83.2	93.8	93.1	**99.4**
Weighted Rec. (%)	97.6	97.6	98.4	97.6	82.9	93.5	92.7	**99.2**
Weighted F1. (%)	97.6	97.5	98.4	97.6	83.0	93.5	92.6	**99.2**

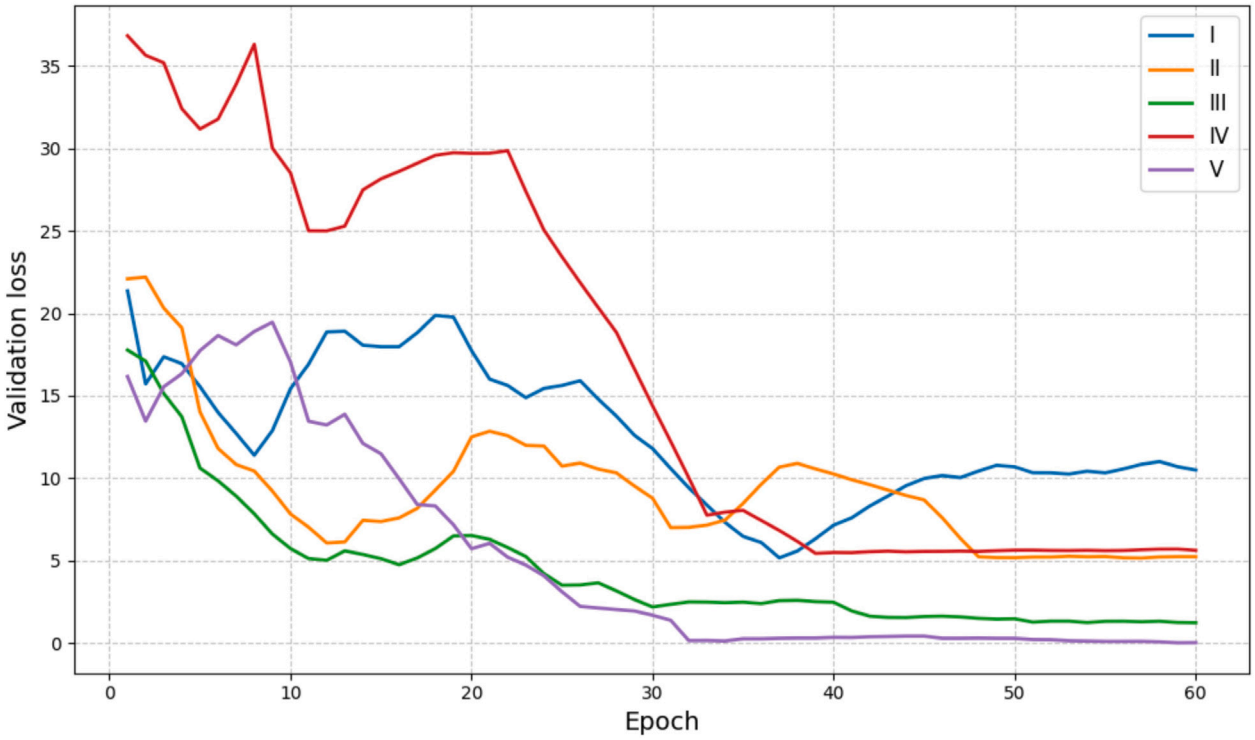
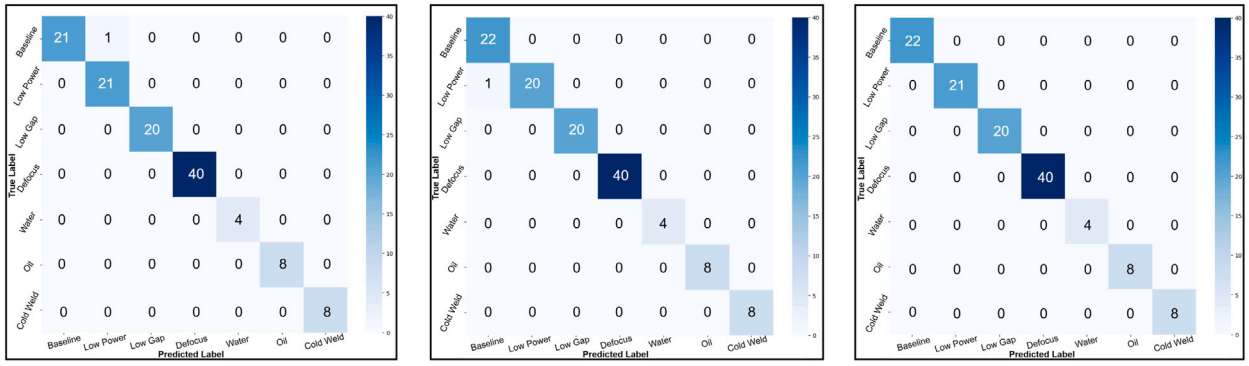


Fig. 17. Validation loss curves under different spectral processing configurations, highlighting the impact of channel selection, embedding strategy, and projection dimension on model performance. Method I, which uses vanilla embedding, yields the poorest performance. Method II employs inverted embedding but lacks spectral channel selection, resulting in limited improvement. Methods III to V explore the effect of varying embedding dimensions, with Method V — featuring the highest projection dimension — achieving the best overall performance.



(a) III (b) IV (c) V

Fig. 18. Confusion matrices of models with different embedding dimensions (64, 96, and 128), illustrating the variation in classification performance across spectral embedding capacities. As the embedding dimension increases, the model achieves progressively better classification performance, indicating that higher-dimensional spectral embeddings enable richer feature representation and improved defect discrimination.

vision-only defect classification using weld images, we employed CNN-based architectures including VGG16 (Simonyan and Zisserman, 2014), MobileNetV2 (Sandler et al., 2018), GoogLeNet (Szegedy et al., 2015), and ResNet50 (He et al., 2016), as well as the Transformer-based ViT-B/16 (Dosovitskiy et al., 2020). For spectrum-only classification based on sequential spectral features, we adopted Transformer-based Informer (Zhou et al., 2021) and the multilayer perceptron-based DLinear (Zeng et al., 2023). Compared to these unimodal vision-only and spectrum-only models, our proposed cross-attention-based fusion method achieved the highest defect classification accuracy.

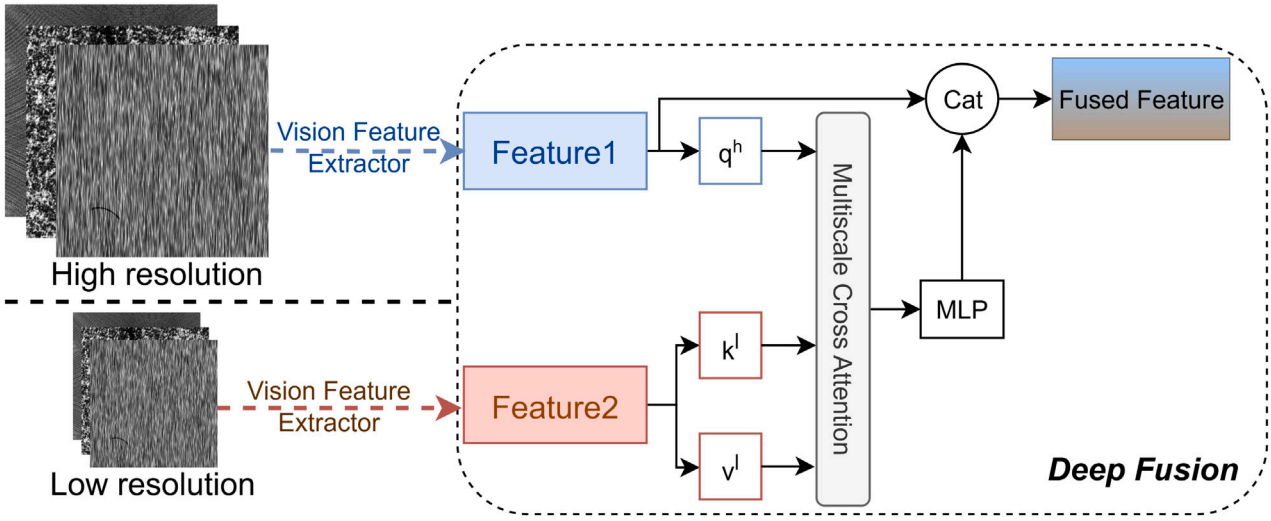


Fig. 19. Multiscale fusion based on cross-attention. We use high-resolution images as the query (Q) and low-resolution images as the key (K) and value (V) for cross-attention fusion, followed by defect classification detection. This setup demonstrates the transferability of our cross-attention method across different resolutions.

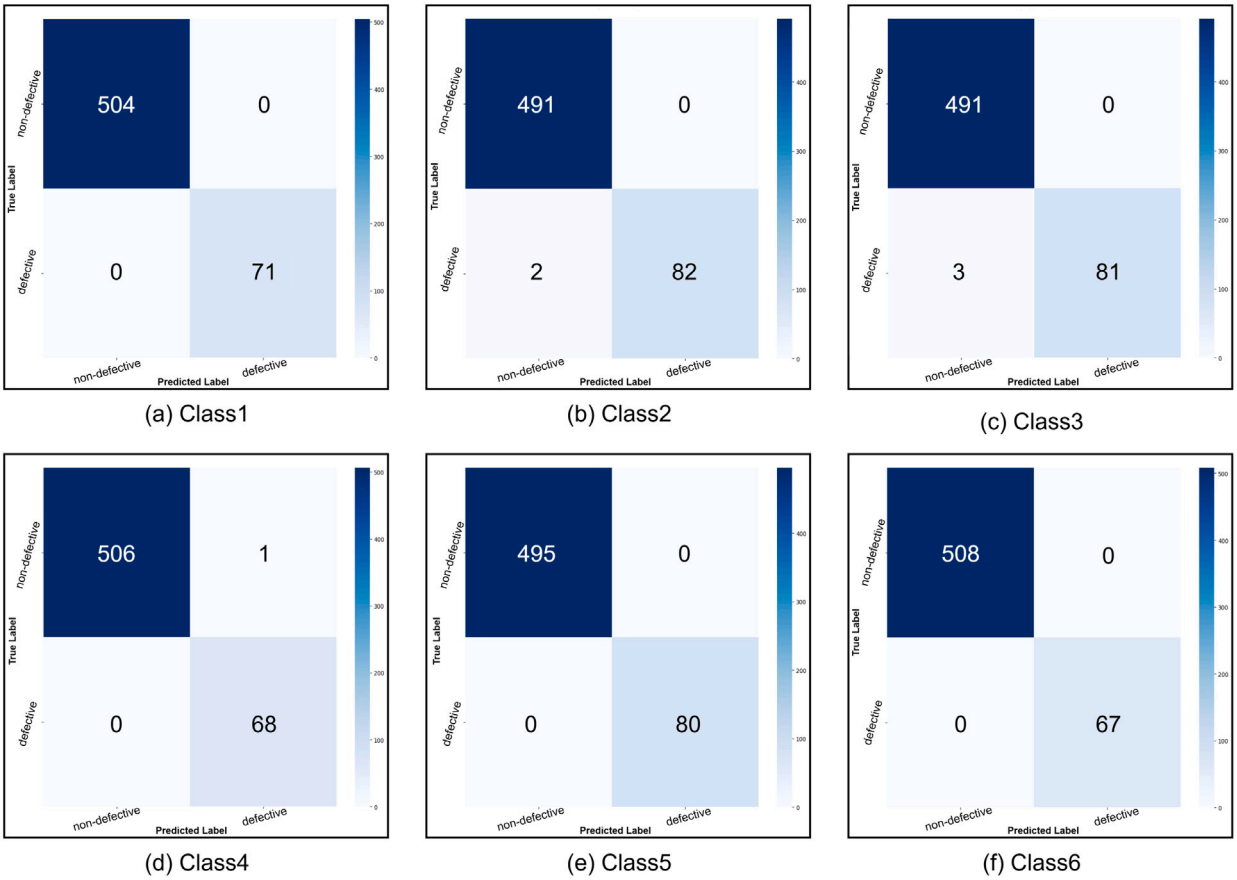


Fig. 20. Confusion matrices on the DAGM dataset. The results demonstrate that our cross-attention fusion method maintains strong performance when transferred to a multiscale defect detection scenario, further validating the effectiveness and generalization capability of the proposed approach.

We evaluate the defect classification performance of each model using weighted precision, weighted recall, and weighted F1-score. As shown in Table 6, our multi-signal fusion method outperforms the best-performing vision-only model, GoogLeNet, by 1.1%, 0.8%, and 0.8% on these three metrics, respectively. Compared to the spectrum-only model Informer, our method achieves improvements of 6.3%, 5.7%, and 5.7%, and exceeds the performance of DLinear by 6.3%, 6.5%, and 6.6%, respectively. As shown in the validation accuracy curves during training in Fig. 13, our method achieved the highest classification accuracy after approximately 60 epochs, outperforming all other unimodal models. During training, we observed that the loss and accuracy curves exhibited slight fluctuations. This behavior is

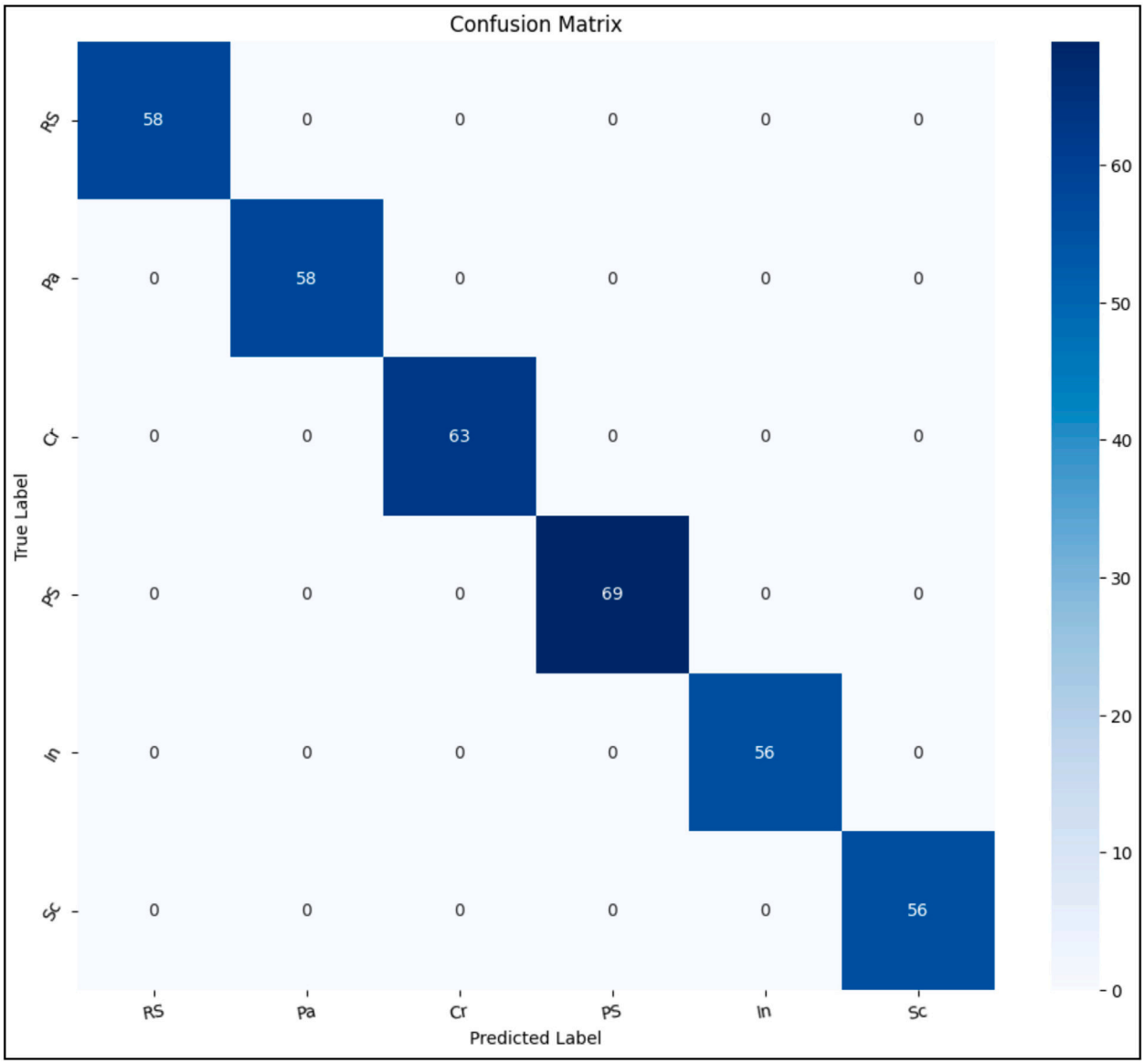


Fig. 21. Confusion matrix on the NEU dataset. The results further confirm the effectiveness and generalization capability of our cross-attention fusion method across different defect datasets.

Table 7
Performance comparison of model variants with different fusion configurations. The proposed model with the cross-attention fusion module achieves the highest classification accuracy, demonstrating the effectiveness of the designed multimodal interaction mechanism.

Model variant	Fusion module	Cross-attention	Params↓	FLOPs↓	Weighted F1.(%)↑
Image only	✗	✗	5.6M	8.5G	97.5
+Spectrum branch	✓	✗	15.9M	12.6G	98.4
+Cross-attention fusion (proposed)	✓	✓	25.9M	18.5G	**99.2**

mainly attributed to the use of large pretrained models on our limited-scale industrial dataset. While pretrained weights provide strong prior knowledge and significantly improve final accuracy and generalization, they may introduce instability in the early fine-tuning stage due to domain mismatch and the small dataset size. Such oscillations are common in transfer learning scenarios with industrial data, and they do not indicate dataset inconsistency. We further verified the dataset integrity and hyperparameters, ensuring that the observed variations reflect the intrinsic noise of welding signals rather than annotation errors.

While spectrum-only models rely solely on material composition and chemical properties, they lack spatial context, making it difficult to localize and differentiate defects with similar spectral signatures. The inclusion of spectral data provides a more comprehensive view of the weld, leading to more accurate defect detection and assessment compared to a vision-only approach.

5.5. Ablation study

To further verify the effectiveness of the proposed cross-attention fusion mechanism, we conducted an ablation study with three model variants, as summarized in Table 7. The Image-only model serves as the baseline, relying solely on visual features for defect classification. In the second variant (+ Spectrum branch), we introduce a spectral branch and perform channel-wise concatenation to fuse the image and spectral features. This simple fusion strategy leads to a moderate

Table 8
Ablation studies on the spectral part.

Methods	Channel	Embedding layer	Embedding dim	Params↓	FLOPs↓	FPS↑	Weighted F1-score(%)↑
I	$vis_1, vis_3, nir_1, nir_{11}$	vanilla	$dim = 96$	25.9M	19.2G	103	96.7
II	$32 channels$	inverted	$dim = 96$	26.0M	18.6G	101	98.4
III	$vis_1, vis_3, nir_1, nir_{11}$	inverted	$dim = 64$	25.9M	18.5G	107	99.2
IV	$vis_1, vis_3, nir_1, nir_{11}$	inverted	$dim = 96$	26.0M	18.5G	105	99.2
V	$vis_1, vis_3, nir_1, nir_{11}$	inverted	$dim = 128$	26.1M	18.5G	104	**100.0**

Table 9
Weighted F1-scores on the NEU and DAGM datasets using the proposed cross-attention fusion strategy.

Dataset	High resolution	Low resolution	Category	Weighted F1-score (%)	Params	FLOPs	FPS	Training time
DAGM	512 × 512	224 × 224	Class1	100.0	16.6M	9.8G	115	6min
			Class2	99.7				
			Class3	99.5				
			Class4	99.8				
			Class5	100.0				
			Class6	100.0				
NEU	200 × 200	128 × 128	–	100.0	16.7M	1.9G	116	4 min

improvement in performance, indicating that spectral information provides complementary cues to the visual modality. Finally, the proposed variant (+ Cross-attention fusion) replaces the naive concatenation with a cross-attention fusion module, which enables adaptive interaction and alignment between spectral and visual representations. As a result, the model achieves the highest weighted F1-score of 99.2%, surpassing other variants while maintaining a reasonable increase in parameters and computational cost. These results confirm that the cross-attention design effectively enhances multimodal feature synergy, thereby improving defect classification accuracy and validating the necessity of our proposed fusion strategy. To further verify the robustness and generalization ability of the proposed multimodal fusion model, we additionally conducted a five-fold cross-validation on the dataset. The entire dataset was randomly divided into five equal subsets, where four subsets were used for training and one for validation in each iteration. The process was repeated five times, and the final performance was averaged over all folds. The results showed that the proposed cross-attention fusion model achieved a mean weighted F1-score of 99.2% (±0.3) across the folds, indicating high stability and consistent performance. This confirms that the model does not rely on a particular data split and generalizes well despite the limited dataset size.

5.5.1. Ablation experiment of visual part

In Section 4.1.2, we introduced the use of U-Net to segment weld images, aiming to extract the weld seam morphology, eliminate background noise, and mitigate issues related to uneven brightness and contrast. In this section, we conduct comparative experiments using both the segmented weld images and the original images in vision-only models to validate the effectiveness of our image preprocessing. Specifically, we evaluate defect classification performance using VGG16, MobileNetV2, ResNet50, GoogleNet, and ViT-B/16, and report the weighted F1-score (%) for both original and segmented images. As shown in Fig. 16, ViT-B/16 achieved the lowest weighted F1-score of 59.3% when using raw images. However, after applying segmented weld images as input, the score improved significantly to 83.0%, marking a 23.7% increase. Similarly, VGG16 and MobileNetV2 also showed improvements of 4.9% and 0.8%, respectively, when using segmented images. These results demonstrate that our weld image segmentation preprocessing effectively enhances defect classification performance.

We applied the proposed cross-attention fusion mechanism on both raw and segmented weld images and visualized the attention maps (Q, K, V) after training, as shown in Fig. 14. The attention localization results extracted from models trained on raw and segmented images are compared. We observed that, without weld segmentation, the attention maps often assign significant weights to background noise. For instance, in sample_1 and sample_3, although attention is partly focused on the weld seam, considerable weights are also assigned to irrelevant regions such as the bottom-right and top-left corners. In sample_2 and sample_4, attention is misdirected outside the weld seam, focusing on surrounding pixels instead of the weld itself. In sample_5, attention is diverted to surface scratches on the material rather than the weld area. In contrast, when using segmented weld images as input, the attention maps from the cross-attention module precisely focus on the weld seams, effectively ignoring background noise. This focused localization significantly enhances the model's ability to extract structural and textural features of the welds.

5.5.2. Ablation experiment of spectral part

In Sections 4.1.1 and 4.2.2, we introduced the use of correlation analysis to select spectral channels—specifically visible channels 1 and 3 (vis_1, vis_3), and near-infrared channels 1 and 11 (nir_1, nir_{11})—as input for our spectral embedding layer. We employed inverted embedding with a projection dimension of 96. In this section, we conduct a series of experiments to evaluate model performance when using all 32 spectral channels versus the selected channels. Additionally, we compare the effects of vanilla embedding and inverted embedding, and investigate how varying the embedding projection dimension (dim) influences model performance. We varied the projection dimension of the spectral embedding layer and observed the results in Table 8. Comparing Methods III, IV, and V with dim = 64, 96, and 128, respectively, we found that when the embedding dimension was increased to 128, the model achieved a classification performance of 100.0% (see Fig. 18). As shown in the loss curve in Fig. 17, after training stabilized around 60 epochs, the model with dim = 128 reached the lowest loss, indicating the best performance on the validation set. In Fig. 15, we visualize the different spectral patterns learned by the embedding layer under various dimensions. These results suggest that increasing the embedding dimension enables the model to capture more expressive spectral representations, which in turn contributes to improved classification accuracy. Method I, which adopts vanilla embedding, shows a 2.5% performance drop. Method II, which uses all 32 spectral channels from a single weld, achieves 98.4% accuracy—a 0.8% decrease. These results demonstrate that channel redundancy negatively impacts detection performance and validate the effectiveness of our correlation-based channel selection strategy.

5.6. Real-time evaluation and deployment

Lightweight design and real-time performance are critical requirements for industrial deployment. To this end, we evaluate the proposed cross-attention fusion network in terms of parameters, FLOPs, inference speed (FPS), and training time. Table 8 reports the results under

different spectral embedding strategies. Two clear observations can be drawn:

(1) Parameter–accuracy trade-off. Increasing the spectral embedding dimension slightly raises the parameter count (from 25.9M to 26.1M) while leaving the FLOPs almost unchanged (≈18.5G). This stability is attributed to our inverted embedding strategy, which projects features along the temporal dimension of the spectral signal. Since cross-attention is computed in a vision-to-spectrum manner, the number of operations does not scale with the temporal length of the spectral input. By contrast, the vanilla embedding (Method I) projects along the channel dimension and consequently increases the FLOPs to 19.2G. Importantly, even with the larger embedding size ($dim = 128$), the model achieves the best performance (100.0% Weighted F1-score) without a noticeable increase in computational cost.

(2) Real-time feasibility. All model configurations achieve inference speeds beyond 100 FPS on an RTX A5500 GPU. Specifically, Method V reaches 104 FPS, while Method III achieves the highest speed of 107 FPS. These results confirm that our approach is highly efficient and meets the real-time requirements of online industrial inspection systems.

Table 9 further validates the efficiency of our model across different datasets and resolutions. On the DAGM dataset (512 × 512 input), the model achieves near-perfect classification across all six defect categories with only 16.6M parameters and 9.8G FLOPs, while sustaining 115 FPS. The training process converges in only 6 min, showing the practicality of fast deployment. On the NEU dataset, which uses lower-resolution inputs (200 × 200 and 128 × 128), the model attains 100.0% Weighted F1-score with merely 1.9G FLOPs and 16.7M parameters, reaching 116 FPS. This demonstrates that the proposed network adapts well to different resolutions and remains lightweight without compromising accuracy. The cross-attention fusion framework achieves a favorable balance between accuracy, computational cost, and speed. With fewer than 26M parameters and FLOPs ranging from 1.9G to 19.2G depending on resolution, our model remains substantially more efficient than many existing transformer-based detectors, while consistently delivering real-time inference (⩾100 FPS). These properties make it well-suited for real-world industrial scenarios, where computational efficiency and deployment feasibility are as important as detection accuracy.

5.7. Multiscale fusion evaluation on DAGM and NEU datasets

To further verify the generalizability and robustness of the proposed fusion strategy, we conducted extensive experiments on two publicly available industrial defect classification benchmarks: DAGM 2007 and the NEU Surface Defect Database. We selected the first six categories from the DAGM 2007 dataset, each containing 1000 non-defective and 150 defective samples. For the NEU Surface Defect Database, a total of 1800 grayscale BMP images were used, comprising six defect types: rolled-in scale (RS), patches (Pa), crazing (Cr), pitted surface (PS), inclusion (In), and scratches (Sc), with 300 images per category. Both datasets were split into training and testing sets using an 8:2 ratio.

As illustrated in Fig. 19, we implemented a multiscale cross-attention fusion framework, in which the original high-resolution image is used to generate the query embeddings (Q), while a rescaled lower-resolution version of the same image is used to derive the key (K) and value (V) representations. We employed the AdamW optimizer with an initial learning rate of 5×10^{-4}, a batch size of 16, and trained the model for 15 epochs.

We present the evaluation results in Table 9, where it can be observed that our proposed fusion model achieved over 99% weighted F1-score across all six defect categories on the DAGM dataset. For the NEU dataset, our model achieved perfect classification performance with a weighted F1-score of 100%. The confusion matrices for both datasets are shown in Figs. 20 and 21, further illustrating the effectiveness and robustness of our approach. In addition, the training time shows that our model converges rapidly within just 10 min, demonstrating the training efficiency of the proposed fusion strategy. Furthermore, as shown in Table 9, our model achieves a throughput of about 115 frames per second (FPS), confirming its capability for real-time deployment in industrial settings. This highlights both the practical value and application potential of our approach in manufacturing environments.

Our multiscale cross-attention fusion strategy enables the model to jointly leverage fine-grained local textures captured from high-resolution images and abstract global semantics from coarser-scale views. This multiscale design addresses the inherent trade-off between detail preservation and receptive field coverage, making the fused features more semantically informative and spatially aware. Empirical results from both datasets consistently demonstrate that our approach not only excels in laser welding inspection, but also generalizes well to broader industrial quality inspection tasks involving subtle surface defects. These successful applications confirm that our fusion framework is not limited to welding image–spectrum integration, but is also applicable to fusing other forms of multimodal industrial data. This provides a promising direction for future industrial AI solutions based on cross-modal learning.

6. Conclusion

In this study, we proposed a novel laser welding defect detection framework tailored for the battery busbar welding scenario, and we conducted extensive experiments to evaluate its effectiveness. First, we collected both 2D grayscale images of the weld seams and spectral signals during the welding process. The raw weld images often contained background noise such as scratches and uneven brightness, which degraded feature quality. By applying U-Net-based segmentation, we successfully isolated the weld seam from the background. This preprocessing step not only reduced irrelevant information but also improved the stability of feature extraction, as later confirmed by attention map visualizations (Fig. 14). The segmented inputs consistently led to higher accuracy compared with raw images, demonstrating the importance of noise reduction in industrial inspection tasks. For the spectral modality, Pearson correlation analysis was applied to select informative channels. This reduced redundancy and improved computational efficiency while preserving discriminative spectral characteristics. Our ablation study (Table 8) shows that using the correlation-selected channels achieved comparable or superior detection performance to using the full set, while significantly lowering model complexity. This confirms that redundancy in spectral inputs can be detrimental and that careful channel selection contributes to both robustness and efficiency. We further evaluated the proposed inverted embedding for spectral feature extraction. Compared with the conventional embedding approach, inverted embedding consistently produced higher F1-scores across different embedding dimensions (64, 96, 128). The improvement can be attributed to its better preservation of temporal relationships in the spectral signal, enabling more discriminative representations. Finally, we introduced a cross-attention-based fusion framework to integrate visual and spectral features. As shown in Table 4, our cross-attention design outperformed element-wise addition and channel-wise concatenation baselines. Visualization of the attention maps further confirmed that the fused model successfully learned to focus on defect-relevant regions of the weld seam, while suppressing background noise. To the best of our knowledge, this is the first application of cross-attention-based multimodal fusion in laser welding analysis. Furthermore, the multiscale fusion experiments conducted on the DAGM and NEU datasets demonstrate that our cross-attention fusion method can be extended to other industrial quality inspection scenarios.

Overall, these results highlight three key findings: (1) segmentation is crucial for noise reduction and effective feature learning; (2) correlation-based channel selection and inverted embedding improve the quality and efficiency of spectral representations; and (3) cross-attention fusion provides a principled way to leverage complementary

modalities, significantly boosting defect detection performance. Beyond busbar welding, our approach is generalizable to other multimodal or multiscale industrial inspection tasks, such as combining 2D images with 3D point clouds or integrating depth maps with grayscale images, offering a promising solution for multi-source data fusion in manufacturing.

CRediT authorship contribution statement

Qin Zhang: Writing – review & editing, Resources, Funding acquisition. **Zhongyou Zhao:** Writing – original draft, Visualization, Validation, Methodology, Formal analysis. **Zhenmin Wang:** Supervision, Project administration, Funding acquisition. **Zixuan Wan:** Data curation, Conceptualization. **Hui-ping Wang:** Supervision, Resources, Funding acquisition. **Guangze Li:** Writing – review & editing, Supervision, Resources, Project administration, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No. U23A20625, Grant No. U2141216 and Grant No. 52375334); Science and Technology Program of Shenzhen, China (Grant No. KJZD20230923114614029).

Data availability

Data will be made available on request.

References

Bao, Y., Song, K., Liu, J., Wang, Y., Yan, Y., Yu, H., Li, X., 2021. Triplet-graph reasoning network for few-shot metal generic surface defect segmentation. IEEE Trans. Instrum. Meas. 70, 1–11.

Chen, Z., Chen, J., Feng, Z., 2018. Welding penetration prediction with passive vision system. J. Manuf. Process. 36, 224–230. http://dx.doi.org/10.1016/j.jmapro.2018.10.009.

Chen, C.-F.R., Fan, Q., Panda, R., 2021. Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 357–366. http://dx.doi.org/10.48550/arXiv.2103.14899.

Deng, H., Cheng, Y., Feng, Y., Xiang, J., 2021. Industrial laser welding defect detection and image defect recognition based on deep learning model developed. Symmetry 13 (9), 1731. http://dx.doi.org/10.3390/sym13091731.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. http://dx.doi.org/10.48550/arXiv.2010.11929, arXiv preprint arXiv:2010.11929.

Fan, K., Peng, P., Zhou, H., Wang, L., Guo, Z., 2021. Real-time high-performance laser welding defect detection by combining ACGAN-based data enhancement and multi-model fusion. Sensors 21 (21), 7304. http://dx.doi.org/10.3390/s21217304.

Gonzalez-Val, C., Pallas, A., Panadeiro, V., Rodriguez, A., 2020. A convolutional approach to quality monitoring for laser manufacturing. J. Intell. Manuf. 31 (3), 789–795. http://dx.doi.org/10.1007/s10845-019-01495-8.

He, G., Gao, X., Yang, H., 2025. AETMC-FCVT: An end-to-end welding defect detection and classification method based on magneto-optical infrared bi-imaging system. Mech. Syst. Signal Process. 224, 112058. http://dx.doi.org/10.1016/j.ymssp.2024.112058.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778. http://dx.doi.org/10.48550/arXiv.1512.03385.

Hwang, S., Lee, J., 2024. Classification of battery laser welding defects via enhanced image preprocessing methods and explainable artificial intelligence-based verification. Eng. Appl. Artif. Intell. 133, 108311. http://dx.doi.org/10.1016/j.engappai.2024.108311.

Khanzadeh, M., Chowdhury, S., Marufuzzaman, M., Tschopp, M.A., Bian, L., 2018. Porosity prediction: Supervised-learning of thermal history for direct laser deposition. J. Manuf. Syst. 47, 69–82. http://dx.doi.org/10.1016/j.jmsy.2018.04.001.

Knaak, C., von Eßen, J., Kröger, M., Schulze, F., Abels, P., Gillner, A., 2021. A spatio-temporal ensemble deep learning architecture for real-time defect detection during laser welding on low power embedded computing boards. Sensors 21 (12), 4205. http://dx.doi.org/10.3390/s21124205.

Liang, F., Zhao, L., Ren, Y., Wang, S., To, S., Abbas, Z., Islam, M.S., 2024. LAD-net: A lightweight welding defect surface non-destructive detection algorithm based on the attention mechanism. Comput. Ind. 161, 104109. http://dx.doi.org/10.1016/j.compind.2024.104109.

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., Long, M., 2023. Itransformer: Inverted transformers are effective for time series forecasting. arXiv preprint arXiv:2310.06625.

Malarvel, M., Singh, H., 2021. An autonomous technique for weld defects detection and classification using multi-class support vector machine in X-radiography image. Optik 231, 166342. http://dx.doi.org/10.1016/j.ijleo.2021.166342.

Medak, D., Posilović, L., Subašić, M., Budimir, M., Lončarić, S., 2021. Deep learning-based defect detection from sequences of ultrasonic B-scans. Ieee Sensors J. 22 (3), 2456–2463. http://dx.doi.org/10.1109/JSEN.2021.3134452.

Mo, S., Morgado, P., 2024. Unveiling the power of audio-visual early fusion transformers with dense interactions through masked modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27186–27196. http://dx.doi.org/10.48550/arXiv.2312.01017.

Pan, Q., Mizutani, M., Kawahito, Y., Katayama, S., 2016. High power disk laser-metal active gas arc hybrid welding of thick high tensile strength steel plates. J. Laser Appl. 28 (1), http://dx.doi.org/10.2351/1.4934939.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PmLR, pp. 8748–8763.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, pp. 234–241.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4510–4520. http://dx.doi.org/10.48550/arXiv.1801.04381.

Shan, Z., Zhang, Y., Yang, Q., Yang, H., Xu, Y., Hwang, J.-N., Xu, X., Liu, S., 2024. Contrastive pre-training with multi-view fusion for no-reference point cloud quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25942–25951. http://dx.doi.org/10.48550/arXiv.2403.10066.

She, K., Li, D., Yang, K., Li, M., Wu, B., Yang, L., Huang, Y., 2024. Online detection of laser welding penetration depth based on multi-sensor features. Materials 17 (7), 1580. http://dx.doi.org/10.3390/ma17071580.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. http://dx.doi.org/10.48550/arXiv.1409.1556, arXiv preprint arXiv:1409.1556.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9. http://dx.doi.org/10.48550/arXiv.1409.4842.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wieler, M., Hahn, T., Hamprecht, F.A., 2007. Weakly supervised learning for industrial optical inspection [dataset].

Yi, X., Xu, H., Zhang, H., Tang, L., Ma, J., 2024. Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27026–27035. http://dx.doi.org/10.48550/arXiv.2403.16387.

Zeng, A., Chen, M., Zhang, L., Xu, Q., 2023. Are transformers effective for time series forecasting? In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, (9), pp. 11121–11128. http://dx.doi.org/10.1609/aaai.v37i9.26317.

Zhang, Y., You, D., Gao, X., Zhang, N., Gao, P.P., 2019. Welding defects detection based on deep learning with multiple optical sensors during disk laser welding of thick plates. J. Manuf. Syst. 51, 87–94. http://dx.doi.org/10.1016/j.jmsy.2019.02.004.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., Zhang, W., 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, (12), pp. 11106–11115. http://dx.doi.org/10.1609/aaai.v35i12.17325.