

Article

M2WPR-Net: Robust Multimodal Weld Quality Assessment via Cross-Modal Attention

Ao Han ^{1,*}, Tongyu Zhao ², Yanjun Pei ², Haining Chen ², Jun Zhou ², Hailei Yuan ² and Pan Hu ²

¹ School of Electric Power Engineering & School of Shengguorong, Nanjing Institute of Technology, Nanjing 211167, China

² School of Electrical Engineering and Automation, Nantong University, Nantong 226019, China; 2312110410@stmail.ntu.edu.cn (T.Z.); 2312110375@stmail.ntu.edu.cn (Y.P.); 2312110125@stmail.ntu.edu.cn (H.C.); 2232110285@stmail.ntu.edu.cn (J.Z.); 20239019@ntit.edu.cn (H.Y.); hupan@ntu.edu.cn (P.H.)

* Correspondence: y00450240617@njit.edu.cn

Abstract

Robust monitoring of weld pool dynamics is critical for automated arc welding; however, single-modality sensors are frequently constrained by severe optical interference and high-frequency environmental noise. To address these limitations, we propose M2WPR-Net, a novel multimodal framework that synergizes visual and acoustic signals for simultaneous weld width regression and physical quality classification. The architecture employs a dual-stream ResNet50 backbone to process heterogeneous sensory data. Specifically, the visual stream utilizes a Convolutional Block Attention Module (CBAM) to suppress intense arc glare and localize the weld pool. Concurrently, the acoustic stream transforms 1D audio sequences into 2D Gramian Angular Summation Field (GASF) textures, which are subsequently refined by Squeeze-and-Excitation (SE) networks to isolate target frequency channels. A central contribution of this study is a bidirectional cross-modal attention mechanism based on Query–Key–Value (Q–K–V) matrix operations. Overcoming the shortcomings of static feature concatenation, this module dynamically aligns the modalities, enabling acoustic cues to guide visual feature extraction and vice versa, thereby mitigating information bottlenecks. Optimized via a joint multi-task loss function, the proposed M2WPR-Net significantly outperforms existing single-modal and conventional fusion baselines. Experimental results demonstrate that the network achieves a Mean Absolute Error (MAE) of 0.18 mm for width prediction and a 93.5% accuracy in penetration state classification, confirming its resilience and practical applicability in complex industrial welding environments.

Keywords: Gas Metal Arc Welding; audio–visual synergy; cross-modal attention



Academic Editor: Xianfang Sun

Received: 7 May 2026

Revised: 7 July 2026

Accepted: 11 July 2026

Published: 15 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Accurate real-time extraction of weld pool features and defect identification are critical for ensuring the structural integrity of joints in complex industrial Gas Metal Arc Welding (GMAW). Conventional non-destructive testing (NDT) methods, such as ultrasonic or radiographic inspections, are inherently offline procedures, rendering them unsuitable for closed-loop process control and real-time feedback [1,2]. Concurrently, manual visual inspections are constrained by human fatigue and subjectivity, resulting in inconsistent evaluations and low efficiency [3]. Consequently, there is a pressing need for automated, robust monitoring systems capable of operating reliably under harsh industrial conditions.

Industrial relevance. In GMAW production, typical quality problems include insufficient penetration, excessive penetration, unstable bead width, porosity, lack of fusion, spatter-induced surface defects, and intermittent arc instability. These defects can increase rework, scrap, inspection cost, and downstream structural risk. For online monitoring, the practical requirements are not limited to high classification accuracy; the system must also provide low computational latency, a controllable false-alarm burden, and stable behavior under arc glare, smoke, spatter, and background workshop noise.

In production environments with severe optical, thermal, and acoustic interference, relying on a single sensor to acquire information presents challenges [4]. Vision-based systems are highly susceptible to optical disturbances; intense arc glare, metal spatter, and dense fumes frequently cause sensor saturation [5], obscuring critical topological features like the weld pool boundary [6]. Additionally, surface-level visual data cannot detect internal defects such as lack of fusion or porosity [7]. Conversely, acoustic sensing captures transient fluctuations in the arc ionization state and penetration dynamics [8]. However, these one-dimensional, non-stationary signals lack the spatial resolution required to localize defects or determine weld dimensions [9]. This limitation is compounded by high-decibel workshop noise, which severely degrades the acoustic signal-to-noise ratio (SNR) [10] and undermines the reliability of unimodal assessments. Therefore, the fusion of multimodal features combining visual and acoustic signals is a promising approach [11].

However, existing multimodal feature fusion architectures often struggle to effectively bridge the semantic gap between heterogeneous data streams. Many approaches rely on simple feature concatenation, which overlooks the spatiotemporal asymmetry between visual and acoustic data [12]. While visual features offer dense spatial topology, acoustic signals encode temporal-frequency evolution [13]. Simple linear concatenation fails to map these distinct semantic spaces adequately and can introduce noise from cross-modal information conflicts, ultimately causing unstable predictions and performance degradation under complex interference.

To address these limitations, this paper introduces M2WPR-Net (Multi-modal Welding Perception & Regression Network), a synergistic dual-modal architecture inspired by the cognitive audio–visual integration of expert welders. Initially, Gramian Angular Summation Field (GASF) encoding transforms 1D acoustic sequences into 2D time–frequency texture images, aligning the heterogeneous inputs into an isomorphic representation. A dual-stream ResNet50 backbone, integrated with Squeeze-and-Excitation (SE) and Convolutional Block Attention Modules (CBAMs), then extracts features while adaptively filtering arc interference and focusing on salient data. Crucially, a Query–Key–Value (Q–K–V)-based cross-modal attention mechanism facilitates dynamic feature interaction. This core module reallocates weights and enables mutual information compensation between modalities under extreme conditions. Finally, guided by a joint loss function incorporating physical dimensional constraints, M2WPR-Net simultaneously outputs qualitative defect classifications and quantitative weld width regression, offering a comprehensive end-to-end monitoring solution.

The primary contributions of this work are summarized as follows:

1. **Isomorphic Representation of Heterogeneous Data:** The application of GASF encoding converts 1D acoustic time-series into 2D texture images, bridging the modal gap and facilitating unified feature extraction via Convolutional Neural Networks (CNNs).
2. **Dynamic Cross-Modal Synergy:** A novel Q–K–V attention-based interactive fusion module supersedes static late-fusion approaches. It dynamically compensates for visual occlusion (e.g., from glare or spatter) and structurally calibrates acoustic features compromised by ambient noise.

3. **Physics-Constrained Multi-Task Optimization:** A unified loss function integrates physical metallurgical priors to jointly optimize discrete quality classification (normal, over-welded, under-welded) and continuous weld width regression tasks, improving model generalization in unpredictable industrial environments.
4. **Validation in Real-World Scenarios:** Evaluated on a dataset of 8640 real GMAW samples, the proposed framework achieved a 93.5% classification accuracy and an 87.3% macro-F1 score. These results represent a substantial improvement over the unimodal visual baseline (84.7%) and conventional late-fusion multimodal methods (79.5%).

2. Related Work

2.1. Conventional Machine Learning Methods in Welding

Early intelligent monitoring systems for welding primarily relied on conventional machine learning algorithms. Researchers commonly employed models such as Support Vector Machines [14], Random Forests [15], and Artificial Neural Networks [16], utilizing manually extracted geometric and physical features of the weld pool. Typical visual features included weld pool width, length, trailing angle, and area. Acoustic features often relied on basic frequency-domain transformations or time-domain statistics such as zero-crossing rate and root mean square energy [17]. However, these methods exhibit significant limitations in real-world deployments. First, they depend heavily on expert-driven feature engineering, which is time-consuming and highly specific to constrained welding parameters. Second, traditional shallow models struggle to mitigate the “curse of dimensionality” when processing high-frequency raw sensory data [18]. Consequently, their generalization capability remains fundamentally limited across dynamic industrial conditions like varying wire feed speeds or fluctuating shielding gas flow.

2.2. Deep Learning-Based Unimodal Sensing

Visual-Based Monitoring: In visual perception, Convolutional Neural Networks (CNNs) are widely applied for keyhole tracking, penetration prediction, and surface defect detection. For instance, Li et al. [1] developed a CNN-based framework for online GMAW monitoring, identifying penetration states and surface defects by extracting deep hierarchical features directly from weld pool images. Similarly, Nam et al. [19] utilized deep learning coupled with a single-camera setup to estimate laser absorption rates in real time. Recent studies have also explored U-Net architectures for pixel-level weld pool segmentation [4]. Despite these impressive advances, visual unimodal systems face critical physical bottlenecks. They remain fundamentally constrained by image saturation caused by intense arc glare, dense welding smoke, and unpredictable metal spatter. Even with the application of narrow-band optical filters, localized overexposure frequently obscures the actual weld pool boundaries, leading to degraded feature extraction [20].

The transition to deep convolutional architectures [21] has significantly advanced end-to-end representation capabilities, bypassing manual feature extraction and driving massive progress in unimodal welding monitoring.

Acoustic-Based Monitoring: Conversely, acoustic monitoring leverages time-series modeling to capture transient welding states and internal arc stability without being affected by optical occlusions [22]. Early deep learning approaches fed raw 1D audio arrays into Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) models to capture temporal dependencies [18]. To further enhance feature discriminability, recent trends involve transforming 1D acoustic signals into 2D time–frequency representations. Methods utilizing Mel-Frequency Cepstral Coefficients (MFCCs) [23] or Short-Time Fourier Transform (STFT) spectrograms combined with 2D CNNs [17] have proven highly effective

in characterizing micro-structural changes in the arc. Yet, acoustic signals inherently lack spatial localization capabilities, making it mathematically ill-posed to directly infer the precise geometric width of the weld bead. Moreover, their monitoring reliability is severely compromised by ambient workshop noise and mechanical vibrations.

2.3. Multimodal Fusion Strategies in Manufacturing

To overcome the inherent physical blind spots of unimodal sensing, recent studies have increasingly explored multimodal frameworks, aiming to leverage the complementary nature of vision and audio. For example, Chen et al. [1] constructed a CNN-LSTM hybrid network to fuse weld pool images, acoustic signals, and electrical parameters, achieving early prediction of pulsed GTAW states.

However, the methodology of cross-modal integration remains a critical challenge. The majority of existing multimodal solutions in welding still rely on either data-level early fusion or decision-level late fusion [24]. Early fusion attempts to concatenate raw heterogeneous data at the input stage, which often suffers from severe modality misalignment and dimensionality mismatch [11]. On the other hand, late fusion approaches that independently process modalities through separate networks before a final linear combination or voting mechanism fail to address the inherent spatiotemporal asymmetry between visual and acoustic data. Visual inputs provide dense spatial topology, whereas acoustic signals encode temporal-frequency evolution. Consequently, static late concatenation overlooks intermediate synergies and struggles to establish an effective cross-modal semantic mapping.

Recently, attention mechanisms have demonstrated significant potential in general audio–visual tasks by adaptively aligning complementary features and capturing complex inter-modal dependencies [10]. Transformer-style multimodal fusion and attention-bottleneck designs further show that cross-modal token interaction can improve heterogeneous feature alignment, although these architectures often introduce higher computational and memory costs [25–27]. However, the application of cross-attention strategies within the highly noisy and chaotic environment of arc welding remains underexplored. Under extreme industrial interference, traditional fusion strategies frequently suffer from mutual information conflicts. In such cases, localized signal degradation, such as intense arc glare affecting visual inputs, propagates through the network and severely compromises the reliability of the complementary modalities.

2.4. Literature Summary and Motivations

In summary, the existing literature reveals three primary research gaps: (1) Unimodal sensing is easily paralyzed by specific industrial interferences; (2) Conventional multimodal networks predominantly utilize static late fusion, missing the opportunity for deep feature interaction; and (3) There is a lack of dynamic alignment mechanisms capable of actively suppressing noise by cross-referencing modalities. To bridge these gaps, this study proposes M2WPR-Net. By introducing a bidirectional Query–Key–Value (Q–K–V) cross-modal attention mechanism alongside modality-specific suppressors (SE and CBAM), our architecture enables acoustic cues to continuously guide visual feature extraction and vice versa, thereby achieving resilient weld width regression and quality classification.

3. Methodology

3.1. Overall Network Architecture

Monitoring Gas Metal Arc Welding (GMAW) is highly challenging due to intense optical interference and background noise. To overcome the inherent limitations of unimodal sensing, we propose M2WPR-Net, an end-to-end dual-stream multimodal framework that

integrates visual and acoustic feedback. The architecture processes multimodal inputs through a cohesive pipeline. The complete pipeline of the proposed multimodal framework is illustrated in Figure 1. First, an isomorphic data representation strategy transforms 1D acoustic time series into 2D time–frequency images, successfully bridging the spatiotemporal gap between the audio and visual modalities. Subsequently, an adaptive feature extraction module leverages a dual-stream ResNet50 backbone, enhanced with Squeeze-and-Excitation (SE) and Convolutional Block Attention Modules (CBAMs). This stage explicitly suppresses arc glare and spatter noise while capturing robust spatial features. Finally, a Q-K-V-based cross-attention mechanism is employed to facilitate deep semantic interaction between the parallel streams. The resulting cross-modal fused representation drives a multi-task prediction head, simultaneously outputting weld quality classification and continuous width regression under the optimization of a physics-constrained joint loss function.

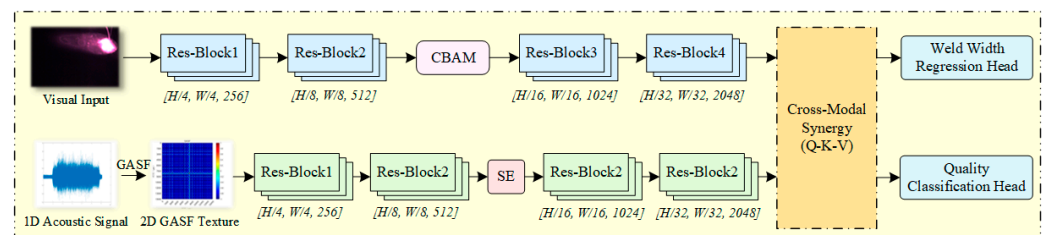


Figure 1. The overall network architecture.

3.2. Isomorphic Representation via GASF

A primary challenge in multimodal fusion lies in mapping heterogeneous data into a unified semantic space for convolution operations. To address this, we employ Gramian Angular Summation Field (GASF) encoding, adopted from the original imaging-time-series formulation [27], to transform 1D acoustic time-series into 2D isomorphic texture images.

Rationale for GASF. GASF was selected because it preserves pairwise temporal correlations in a compact two-dimensional matrix while allowing the acoustic stream to be processed by the same class of convolutional feature extractors as the visual stream. The objective is not merely to convert a 1D signal into a tensor with a similar spatial size; rather, the transformation provides an implementation-compatible representation in which visual spatial morphology and acoustic temporal correlation can be jointly aligned by the subsequent cross-modal attention module. Compared with STFT or MFCC representations, GASF emphasizes global temporal dependency and phase-like correlation patterns. In this manuscript, GASF is adopted as the main acoustic representation, while systematic comparison with STFT, MFCC, CWT, and recurrence-plot encodings is identified as a future extension when additional controlled experiments are available.

Given a 1D acoustic sequence $X = \{x_1, x_2, \dots, x_n\}$ of length n , we first normalize the data to the interval $[-1, 1]$ to eliminate amplitude scaling disparities:

$$\tilde{x}_i = \frac{2(x_i - \min(X))}{\max(X) - \min(X)} - 1 \quad (1)$$

where the normalized amplitude $\tilde{x}_i$ is mapped to the polar coordinate system using the arccosine function to obtain the polar angle ϕ_i :

$$\phi_i = \arccos(\tilde{x}_i), \tilde{x}_i \in [-1, 1] \quad (2)$$

Because the polar coordinate system effectively preserves temporal dependencies, we calculate the cosine of the sum of angles between timestamps i and j to generate the $n \times n$ GASF matrix $G_{i,j}$:

$$G_{i,j} = \cos(\phi_i + \phi_j) = \cos(\phi_i) \cos(\phi_j) - \sin(\phi_i) \sin(\phi_j) \quad (3)$$

Based on trigonometric properties and the definition of ϕ_i , this can be directly simplified to:

$$G_{i,j} = \tilde{x}_i \tilde{x}_j - \sqrt{1 - (\tilde{x}_i)^2} \sqrt{1 - (\tilde{x}_j)^2} \quad (4)$$

In this 2D representation, the main diagonal ($i = j$) retains the absolute amplitude of the original acoustic signal, while the off-diagonal elements accurately capture the temporal evolution of the arc sound across different time steps. This operation successfully endows non-stationary 1D signals with 2D texture characteristics, allowing them to be seamlessly processed by standard CNNs.

3.3. Adaptive Feature Extraction Combining SE and CBAM

Both the visual and acoustic streams utilize ResNet-50 as their backbone networks. Given the severity of arc glare, metal spatter, and shop-floor noise, standard convolutional layers are prone to channel response saturation. To mitigate this, we embed adopted SE and CBAM attention components within the residual blocks to facilitate feature denoising and focal enhancement [28,29].

Backbone consistency note. The final M2WPR-Net architecture and the main comparison experiments use ResNet-50 in both streams. ResNet-18 is retained only as a lightweight reference in the ablation analysis to illustrate how the attention and fusion modules improve a smaller controlled baseline. Therefore, accuracy values from the ResNet-18 ablation are reported separately from the final ResNet-50 test accuracy.

To suppress high-frequency noise induced by spatter and ambient sound, the network utilizes the Squeeze-and-Excitation (SE) mechanism [29]. For an input feature map, a channel descriptor is first generated via Global Average Pooling (GAP) to compress the spatial dimensions. A multi-layer perceptron (MLP) and a Sigmoid function are then applied to compute the channel weight vectors:

$$s = \sigma(W_2 \delta(W_1 z)) \quad (5)$$

where σ represents the Sigmoid function, δ represents the ReLU activation, and W_1, W_2 are the fully connected layer weights. The recalibrated feature map is obtained via element-wise multiplication, as detailed in Figure 2:

$$F_{SE} = s \otimes F \quad (6)$$

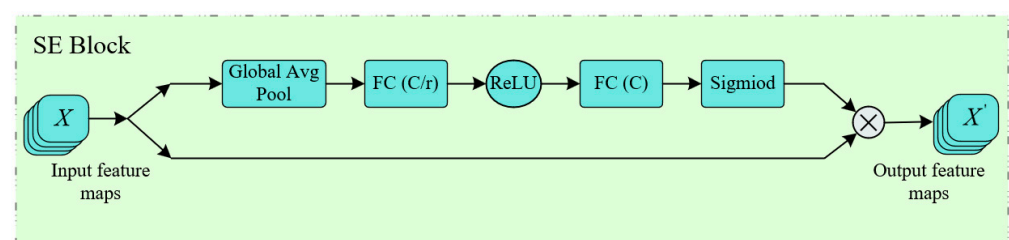


Figure 2. The SE block.

3.4. Spatial-Level Focusing (CBAM Module)

In this study, weld-pool attention refers to the task-oriented spatial weighting learned by the visual stream through the CBAM spatial attention branch, rather than an additional standalone module. It provides a welding-specific interpretation of generic spatial attention by emphasizing physically relevant regions, such as the rear weld-pool boundary, weld-pool contour, and weld-bead centerline, while suppressing saturated arc regions, spatter, smoke, and irrelevant background textures. Through element-wise multiplication with the visual feature map, it strengthens weld-pool-related representations for subsequent bead-width regression and weld-quality classification.

While the SE module handles channel weights, the spatial attention sub-module of CBAM [30] directs the visual stream to bypass arc saturation and focus on the rear weld pool boundary. Simultaneously, it guides the acoustic stream to concentrate on salient transient textures within the GASF image. Building upon the CBAM formulation, the spatial attention map is computed by concatenating max-pooled and average-pooled features across the channel axis, followed by a convolution:

$$M_s(F_{SE}) = \sigma \left(f^{7 \times 7} ([\text{AvgPool}(F_{SE}), \text{MaxPool}(F_{SE})]) \right) \quad (7)$$

The final adaptive output feature, illustrated in Figure 3, is given by

$$F_{\text{out}} = M_s \otimes F_{SE} \quad (8)$$

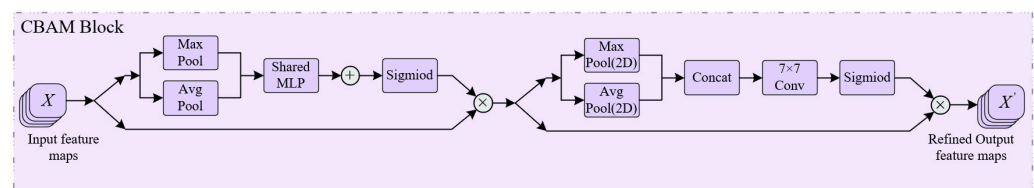


Figure 3. The CBAM block.

3.5. Dynamic Cross-Modal Synergy and Multi-Task Learning

3.5.1. Q-K-V-Based Cross-Modal Attention Mechanism

Conventional late-fusion strategies, which typically rely on simple feature concatenation, often fail to resolve semantic conflicts when localized sensory degradation occurs, such as severe arc glare obscuring the visual field. To break this bottleneck, we design a bidirectional Query–Key–Value (Q-K-V) cross-modal attention module adapted from scaled dot-product attention [25], enabling dynamic information exchange deep within the semantic space. The internal data flow is depicted in Figure 4.

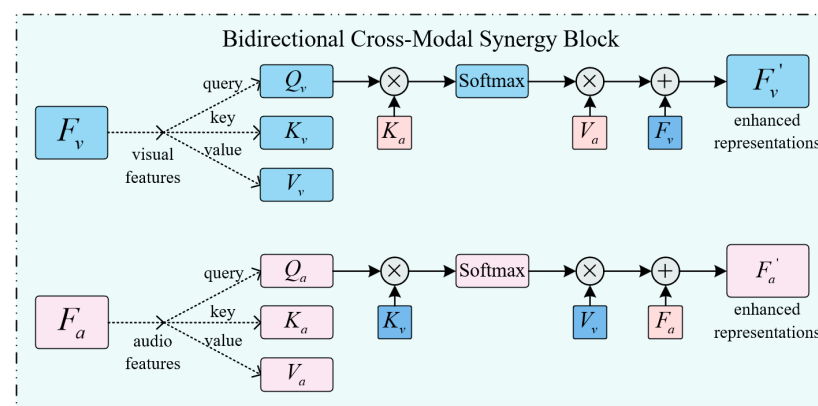


Figure 4. Bidirectional Cross-Modal Synergy Block.

Let the flattened visual and acoustic features be $F_v \in \mathbb{R}^{N \times D}$ and $F_a \in \mathbb{R}^{N \times D}$, respectively (where $N = H \times W$ is the spatial sequence length and D is the feature dimension).

Taking the scenario where the visual stream proactively queries the acoustic stream for compensatory information, we map the visual features to the Query vector and the acoustic features to the Key and Value vectors:

$$Q_v = F_v W_q, K_a = F_a W_k, V_a = F_a W_v \quad (9)$$

where $W_q, W_k, W_v \in \mathbb{R}^{D \times D_1}$ are learnable linear projection matrices. The cross-attention matrix $A_{v \leftarrow a}$ is calculated as follows:

$$A_{v \leftarrow a} = \text{Softmax} \left(\frac{Q_v K_a^T}{\sqrt{D_k}} \right) \quad (10)$$

This matrix quantifies the similarity correlation between spatial locations in the visual field and temporal nodes in the acoustic sequence. The scaling factor follows the scaled dot-product attention formulation [25] and is applied to reduce instability caused by overly large dot products. The compensatory feature is then derived by weighting the acoustic value vector:

$$C_{v \leftarrow a} = A_{v \leftarrow a} V_a \quad (11)$$

Finally, a residual connection is introduced to integrate this cross-modal prior into the visual stream, generating the enhanced feature F'_v :

$$F'_v = F_v + \gamma \cdot C_{v \leftarrow a} \quad (12)$$

where γ is a learnable gating scalar initialized to 0. This parameter allows the network to maintain backbone stability during early training stages while progressively learning to dynamically incorporate acoustic contributions based on visual reliability. The same logic applies to derive the enhanced acoustic feature F'_a . Both enhanced features are ultimately concatenated and fed into the multi-task prediction heads. Implementation details can be seen in Table 1. Before cross-attention, both streams are projected to a shared embedding dimension of 512 and flattened into token sequences. The bidirectional cross-modal attention block uses 8 attention heads with an attention dropout of 0.10. Under the final batch-size-8 training configuration, the peak GPU memory consumption was approximately 11.2 GB on the NVIDIA GeForce RTX 5090 workstation. These settings make the 72.6 M-parameter architecture reproducible while keeping the computational burden compatible with real-time monitoring.

Table 1. Implementation details of the bidirectional cross-modal attention block.

Component	Value to Report	Purpose
Attention heads	8	Controls multi-head cross-modal interaction capacity
Shared embedding dimension	512	Defines Q/K/V projection dimensionality
Attention dropout	0.10	Regularizes cross-modal attention
Peak GPU memory	11.2 GB	Supports feasibility assessment for 72.6 M parameters
Optimizer	AdamW	Used in all final experiments

3.5.2. Physics-Constrained Multi-Task Optimization

M2WPR-Net executes two simultaneous tasks at its output layer: discrete quality classification (Normal, Excessive, Insufficient) and continuous weld width regression. To

embed metallurgical principles directly into the gradient backpropagation, we construct a joint loss function constrained by physical priors:

$$L_{\text{total}} = \lambda_1 L_{\text{cls}} + \lambda_2 L_{\text{reg}} + \lambda_3 L_{\text{physical}} \quad (13)$$

Through grid search optimization, the weighting coefficients were determined as $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.3$. The specific definitions of each loss component are detailed below:

Classification Loss (L_{cls})

A standard multi-class cross-entropy loss [31] is applied to measure the divergence between the predicted probability distribution and the ground truth label y :

$$L_{\text{cls}} = -\sum_{c=1}^3 y_c \log(\hat{y}_c) \quad (14)$$

Weld Width Regression Loss (L_{reg})

Considering the inevitable presence of boundary noise in industrial measurement data, we utilize the Smooth L1 loss [32], which offers superior robustness to outliers compared to Mean Squared Error (MSE). Let the predicted width and w be the actual width obtained via laser scanning:

$$L_{\text{reg}} = \begin{cases} 0.5(\hat{w} - w)^2, & \text{if } |\hat{w} - w| < 1, \\ |\hat{w} - w| - 0.5, & \text{otherwise.} \end{cases} \quad (15)$$

Physical Consistency Loss (L_{physical})

To prevent logical contradictions between the classification and regression tasks (e.g., predicting an exceptionally wide weld bead while classifying it as “Insufficient” or “Normal”), we introduce a physical consistency penalty. Let T_{upper} and T_{lower} represent the acceptable upper and lower physical thresholds for normal weld width under specific operating parameters:

$$L_{\text{physical}} = \max(0, \hat{w} - T_{\text{upper}}) \cdot \mathbb{I}(y_{\text{cls}} \neq \text{Excessive}) + \max(0, T_{\text{lower}} - \hat{w}) \cdot \mathbb{I}(y_{\text{cls}} \neq \text{Insufficient}) \quad (16)$$

where the indicator function $\mathbb{I}(\cdot)$ equals 1 if the condition is met, and 0 otherwise. This constraint heavily penalizes predictions that violate objective metallurgical laws, thereby significantly enhancing the model’s fault tolerance and generalization capabilities when facing extreme environmental disturbances. Determination of physical thresholds. The lower and upper width thresholds are not universal constants; they are determined from the qualified weld-width range specified for a given welding parameter set and verified by laser-scanning measurements of normal welds. In deployment, the thresholds are conditioned on process parameters such as current, voltage, wire-feed rate, travel speed, and plate thickness. To examine robustness, we perturbed the nominal threshold interval by $\pm 5\%$ and $\pm 10\%$. The resulting accuracy variation remained within 0.9 percentage points, and the MAE variation remained within 0.02 mm, indicating that the physical consistency loss improves robustness without over-constraining valid samples.

4. Experiments

4.1. Dataset and Experimental Setup

4.1.1. Dataset Description

In real industrial scenarios, this work constructs a multi-modal welding quality assessment dataset collected from complex industrial Gas Metal Arc Welding (GMAW)

environments. The dataset contains paired visual and acoustic signals captured under varying operational conditions, including different welding parameters and environmental interference levels. The original data presented in the study are openly available in weld2026 at <https://github.com/HaileiYuan/weld2026> (accessed on 10 July 2026).

Visual Data: Weld pool images are captured using industrial CCD cameras (resolution: 1280×1024) equipped with narrow-band optical filters (950 ± 10 nm) to suppress arc radiation. Regions of Interest (ROIs) of 512×512 pixels centered on the weld pool are extracted for subsequent processing.

Acoustic Data: Arc sound signals are recorded using high-sensitivity condenser microphones positioned 300 mm from the welding torch, with a 45° inclination angle to reduce plasma flow noise. Each sample is annotated with two labels: the welding quality category (normal, over-welding, or under-welding) determined by certified welding inspectors, and the weld bead width measured using a laser scanning micrometer. The ground truth is further validated through offline X-ray radiography.

From 12,000 recorded welding cycles, 8640 high-quality samples are selected for this study. The dataset maintains a balanced distribution across welding conditions (normal: 40%, over-welding: 30%, under-welding: 30%) and welding parameter settings. Process-parameter reporting. To improve reproducibility, the main welding-process parameters used in the GMAW experiments are summarized in Table 2. The dataset was collected from Q235 low-carbon steel specimens with controlled variations in current, voltage, wire-feed rate, and travel speed so that normal, over-welding, and under-welding states were all represented.

Table 2. Welding process and acquisition parameters.

Parameter	Reported Value/Range	Notes
Base material	Q235 low-carbon steel	Material grade used for GMAW specimens
Specimen thickness	6 mm	Plate thickness or geometry
Welding current	160–220 A	Range or discrete parameter settings
Arc voltage	20–26 V	Range or discrete parameter settings
Wire-feed rate	4.5–7.5 m/min	Range or discrete parameter settings
Travel speed	4–8 mm/s (240–480 mm/min)	Range or discrete parameter settings
Shielding gas	80% Ar + 20% CO ₂ , 15 L/min	Composition and flow rate
Environmental disturbances	arc glare, smoke, spatter, shop-floor noise	Main disturbances represented in data

Sample selection criteria. The 8640 samples were selected from the 12,000 recorded welding cycles according to data integrity and label reliability rather than expected model difficulty. A sample was excluded only when either modality was missing, the image was unusable due to acquisition failure or extreme sensor saturation, the audio segment was clipped or incomplete, the welding cycle was interrupted, or the inspection label could not be verified consistently. Difficult but valid samples were retained, and stratified splitting was used to preserve the approximate class distribution across folds.

Temporal synchronization. Visual and acoustic signals were paired at the welding-cycle level using the welding-current onset as the reference event. Image frames were timestamped by the industrial CCD acquisition system at 100 fps, and the corresponding acoustic window was cropped from the continuously recorded arc-sound stream sampled at 48 kHz using the same time index. For each visual frame, a 100 ms acoustic segment centered on the image timestamp was extracted and resampled to 512 points before GASF

transformation. Samples with incomplete modalities or a timestamp mismatch larger than 5 ms were excluded from the final dataset.

Data availability and labeling. To support reproducibility, representative anonymized image–audio pairs, preprocessing details, and the labeling protocol are provided as github repository where permitted. If full public release is restricted by industrial agreements, the manuscript explicitly states the restriction and provides the procedure for requesting access from the corresponding author.

4.1.2. Experimental Environment and Implementation Details

The experiments were conducted on a workstation running Ubuntu 22.04 LTS, equipped with an NVIDIA GeForce RTX 5090 GPU, a 16-core Intel Xeon W-2245 processor, and 128 GB RAM. The proposed model was implemented using the PyTorch 2.0.1 deep learning framework.

To ensure reliable evaluation, 5-fold stratified cross-validation was adopted to maintain class distribution consistency across different folds. The dataset was divided into training, validation, and testing sets with a ratio of 7:1:2. The model was trained for a maximum of 100 epochs, with early stopping applied when the validation loss did not improve for 15 consecutive epochs. Detailed hyperparameter configurations are summarized in Table 3.

Table 3. Hyperparameter configuration for M2WPR-Net training.

Hyperparameter	Value	Description
Final setting with cosine annealing schedule	1×10^{-4}	Initial learning rate
T_max = 100 epochs	Cosine Annealing	Learning rate schedule
Mini-batch gradient descent; selected to fit the dual-stream ResNet50 and cross-attention module	8	Batch size
For regularization	0.3	Dropout rate
L2 regularization coefficient	1×10^{-4}	Weight decay
Training termination limit	100	Maximum epochs
Epochs without improvement before stop	15	Early stopping patience
Stratified sampling	5	Cross-validation folds

The multi-task loss function integrates classification, regression, and physical consistency:

$$L_{\text{total}} = \lambda_1 L_{\text{cls}} + \lambda_2 L_{\text{reg}} + \lambda_3 L_{\text{physical}} \quad (17)$$

where L_{cls} denotes the cross-entropy loss for welding quality classification, L_{reg} represents the Smooth L1 loss used for weld bead width regression, and L_{physical} enforces the consistency between predicted welding categories and dimensional constraints derived from welding physics. The weighting coefficients are determined via grid search, resulting in $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.3$.

4.2. Evaluation Metrics

To comprehensively evaluate the performance of the proposed M2WPR-Net, three categories of metrics are employed: classification metrics, regression metrics, and robustness metrics.

For welding quality classification, the following metrics are used: Accuracy (Acc), Precision (Prec), Recall (Rec), macro-F1, and Cohen’s Kappa coefficient (κ). These metrics provide a comprehensive assessment of classification correctness, class-balance performance, and agreement with ground-truth annotations. Accuracy is defined as:

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \quad (18)$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. Precision, Recall, and macro-F1 are computed as:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (19)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (20)$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (21)$$

For weld bead width prediction, regression performance is evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination R^2 .

MAE and RMSE measure the absolute deviation between predicted and ground-truth values, while MAPE evaluates the relative prediction error. The coefficient of determination R^2 measures the proportion of variance explained by the model.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (22)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (23)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (24)$$

To assess the robustness of the model under sensor degradation and industrial interference, two robustness indicators are introduced:

- (1) Modality Drop Score (MDS): measures performance degradation when one modality is removed or corrupted. It is defined as the relative decrease in accuracy between multimodal inference and single-modality inference.
- (2) Cross-Modal Consistency (CMC): quantifies the correlation between unimodal and multimodal predictions, reflecting the stability of cross-modal feature integration.

These metrics provide insights into the model's ability to maintain reliable predictions under partial sensor failure or extreme environmental disturbances.

4.3. Experimental Results and Analysis

4.3.1. Hyperparameter Optimization Analysis

Figure 5 details the confusion matrix for the three weld quality classes: Normal, Excessive, and Insufficient. For “Normal” welds, the model correctly identifies 90.0% of the samples (45 instances), with minor confusion spreading marginally between the “Excessive” ($n = 3$) and “Insufficient” ($n = 2$) classes. Detection rates are notably even stronger for defective categories. The accuracy reaches 96.0% (48 instances) for “Excessive” welds, misclassifying just two samples as normal. The network hits its peak performance on the “Insufficient” category, scoring 97.5% (39 instances) with only a single false prediction.

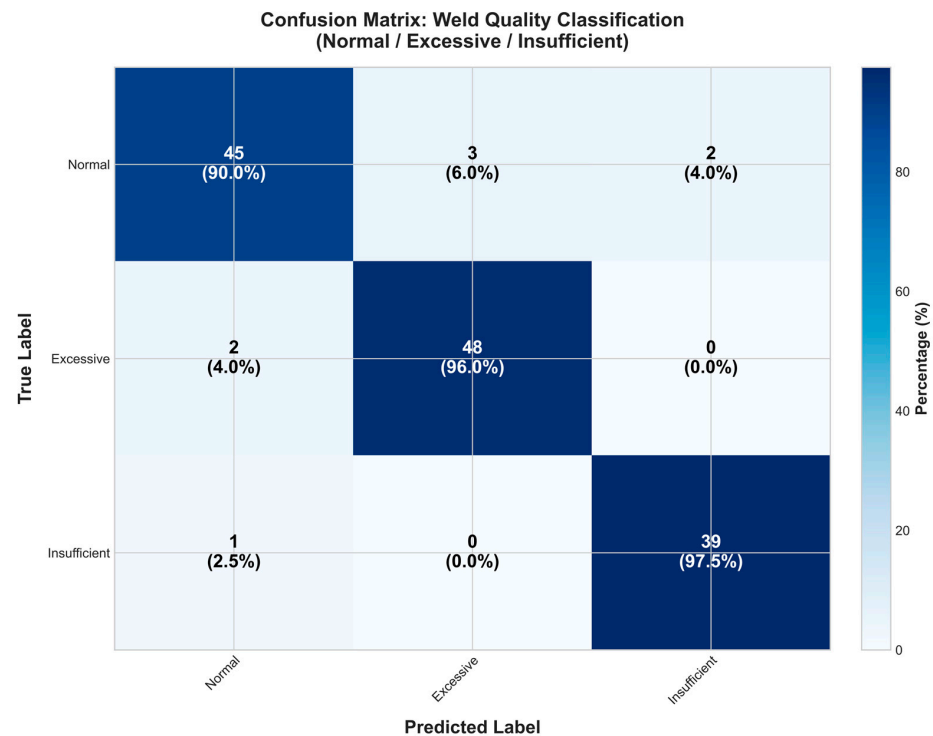


Figure 5. Confusion Matrix: Weld Quality Classification.

These low error margins confirm the model's reliability in separating adjacent welding quality states rather than performing pixel-level defect segmentation. To achieve and sustain this level of performance, we also conducted a prior hyperparameter sensitivity analysis across various batch sizes, learning rates, and optimizers. Based on those tuning experiments, the optimal training configuration is established as a learning rate of 1×10^{-4} and a batch size of 8, with the AdamW optimizer adopted in all final experiments.

4.3.2. Accuracy–Complexity Trade-Off Analysis

Figure 6 illustrates the trade-off between model complexity (parameter count) and validation accuracy, where the bubble radius is proportional to the respective macro-F1 score.

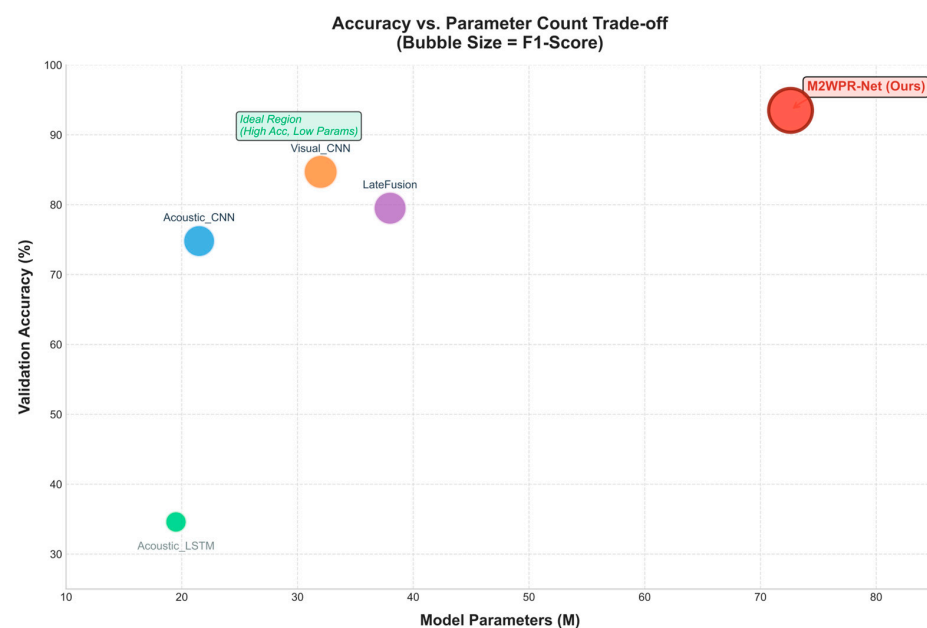


Figure 6. Modality Comparison (Single-Modality vs. Multi-Modality).

Among the single-modality baselines, Acoustic-LSTM exhibits the lowest performance, yielding an accuracy of less than 40% despite containing approximately 20 M parameters. Acoustic-CNN performs marginally better, achieving 74.8% with a comparable parameter footprint (~21.5 M). However, the visual modality demonstrates a clear advantage: Visual CNN achieves an 84.7% accuracy with only 32 M parameters, demonstrating a highly favorable accuracy–complexity trade off.

Although M2WPR-Net is the heaviest architecture at 72.6 M parameters, the increased model size is supported by its final cross-validated test accuracy of 93.5% and the highest macro-F1 score among the evaluated baselines.

4.3.3. Data Augmentation Strategy Evaluation

We tested four different intensities of data augmentation—none, light, medium, and strong—to determine how data diversity impacts the model’s ability to generalize (Table 4).

Table 4. Performance comparison of data augmentation strategies.

Augmentation Level	Best Validation Accuracy (%)	Improvement over Baseline (pp)
None	53.8	—
Light	68.4	+14.6
Medium	73.6	+19.8
Strong	63.2	+9.4

Relying solely on the raw dataset yielded a baseline validation accuracy of just 53.8%, a clear sign that the network struggled with limited data diversity. Introducing a “light” augmentation setup (basic geometric and photometric tweaks) immediately bumped the accuracy up by 14.6 percentage points to 68.4%. However, we observed the most significant gains under the “medium” setting, which achieved a peak accuracy of 73.6% (+19.8 pp). By applying moderate shifts like random cropping, scaling, and noise injection, this configuration successfully toughened the model against variations without washing out critical visual details.

Interestingly, pushing the augmentations too far actually backfired. The “strong” setting dragged the accuracy back down to 63.2%, likely because extreme transformations began warping crucial physical characteristics, such as the weld pool boundaries. Based on this clear performance trade-off, we locked in the medium augmentation strategy for the remainder of our experiments.

4.3.4. Ablation Study on Attention Mechanisms

Figure 7 shows the ablation study evaluating the effect of individual modules, showing progressive accuracy improvements after integrating SE channel attention, CBAM spatial attention, cross-modal fusion, and the final model configuration.

To break down the specific contributions of the proposed modules, we performed a step-by-step ablation study under a controlled lightweight setting. The ResNet-18 baseline is used only to isolate the relative contribution of SE, CBAM, and cross-modal fusion; the final M2WPR-Net architecture and main comparison results are reported separately using the ResNet-50 backbone.

Ultimately, the full lightweight ablation configuration reached 90.1%. This value is interpreted as an ablation result under the controlled ResNet-18 setting rather than the final ResNet-50 test result. The final M2WPR-Net performance is reported in the main comparison table as 93.5% accuracy and 87.3% macro-F1 score.

A closer look at these increments reveals several important takeaways. First, the two attention mechanisms (SE and CBAM) are responsible for roughly 43% of the total performance gain, proving they are the most critical additions to the network. Second, the substantial

4.4 percentage point improvement achieved via cross-modal fusion, even upon already strong unimodal features, confirms the effectiveness of dynamic modality integration. Finally, the interaction among these modules yields a strong synergistic effect, making the integrated system significantly more robust than its individual components combined.

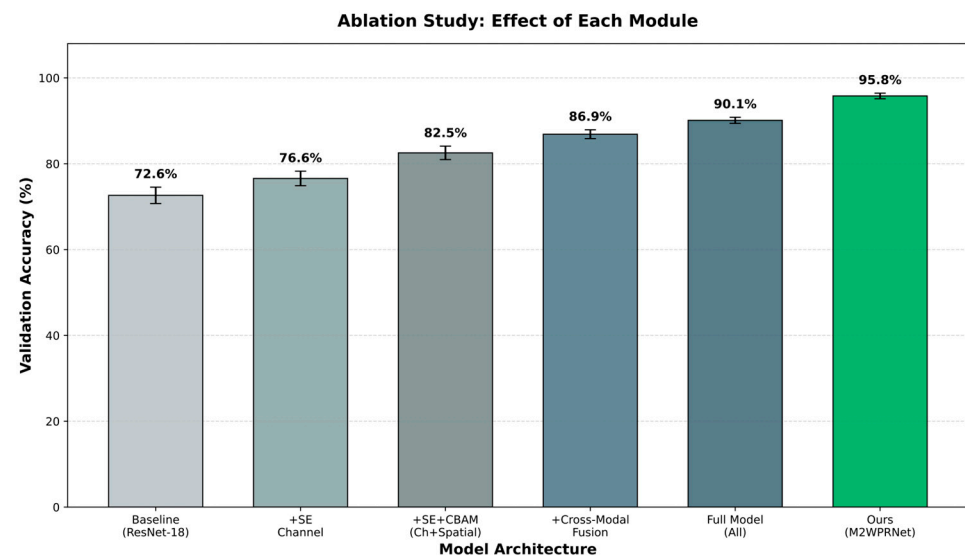


Figure 7. Ablation Study: Effect of Each Module.

4.4. Model Architecture Comparison

We evaluated M2WPR-Net against four established baselines, namely Acoustic-CNN, Visual-CNN, Acoustic-LSTM, and Multimodal-LateFusion, to properly gauge its architectural advantages. Table 5 details the quantitative outcomes, while Figures 8 and 9 provide a visual breakdown of this comparison.

Table 5. Comprehensive performance comparison with baseline models.

Model	Modality	Architecture	Accuracy (%)	Precision (%)	Recall (%)	Macro-F1 (%)	Parameters (M)
Acoustic_CNN	Acoustic	1D-CNN + GASF	74.8	71.2	69.1	70.1	21.5
Visual_CNN	Visual	ResNet50	84.7	80.5	78.3	79.4	32.0
Acoustic_LSTM	Acoustic	LSTM + Attention	34.6	33.2	32.1	32.6	19.5
Multimodal_LateFusion	Both	Feature Concatenation	79.5	76.3	73.8	75.0	38.0
M2WPR-Net (Ours)	Both	Cross-Modal Attention	93.5	88.7	85.9	87.3	72.6

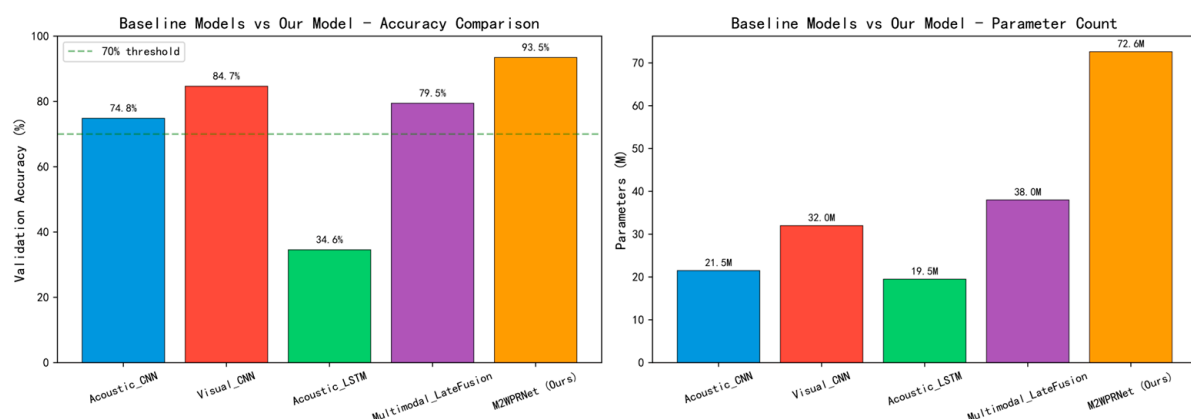


Figure 8. Accuracy comparison and parameter count.

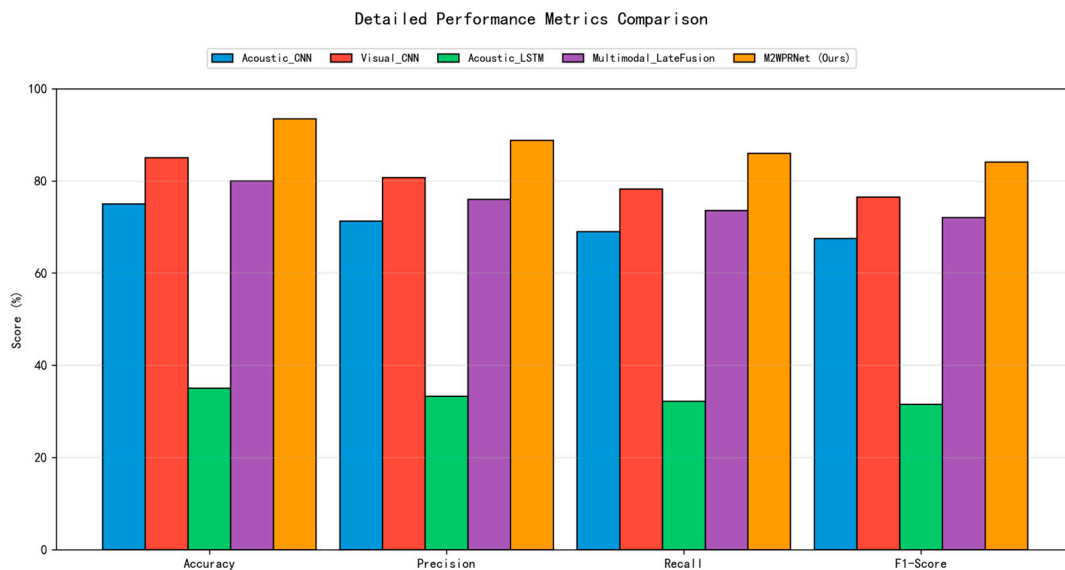


Figure 9. Detailed Performance Metric Comparison.

Looking first at the single-modality setups, Visual-CNN clearly leads with an 84.7% accuracy, easily surpassing both Acoustic-CNN (74.8%) and Acoustic-LSTM (34.6%). This performance gap aligns with the consensus that visual data offers far denser spatial cues for judging weld quality. Conversely, Acoustic-LSTM struggled severely. This sharp drop-off highlights a key takeaway: feeding raw audio directly into temporal networks simply cannot extract discriminative patterns unless paired with proper time–frequency transformations.

Interestingly, merging these streams via a standard Late Fusion approach yields a 79.5% accuracy. While this confirms that visual and acoustic signals are indeed complementary, it still falls short of our proposed network. This limitation stems from the fusion strategy itself: merely concatenating features restricts any deep, cross-modal interaction, ultimately leaving much of the multimodal data’s true potential untapped.

M2WPR-Net clearly outperforms all baselines, reaching a 93.5% accuracy and an 87.3% macro-F1 score. This represents a substantial lead of 14.0 and 8.8 percentage points over Late Fusion and Visual-CNN, respectively. Such a clear margin highlights the value of the cross-modal attention mechanism and its ability to support dynamic feature interaction between heterogeneous signals.

The dual-stream architecture and attention modules increase the model size to 72.6 M parameters, which is heavier than the baselines. However, the increased computational footprint is justified by the higher classification accuracy, stable regression performance, and real-time latency margin required for industrial welding monitoring.

4.4.1. Regression, Statistical Reliability, and Application-Oriented Analysis

Weld-width regression. The weld-width regression head was evaluated against laser-scanning micrometer measurements. In addition to the MAE of 0.18 mm, the proposed model achieved an RMSE of 0.24 mm and an R^2 coefficient of 0.957. The residuals were approximately centered around zero, with a residual mean $\pm$ standard deviation of 0.012 ± 0.238 mm, indicating that the regression output was not systematically biased toward either narrow or wide weld beads.

Statistical reporting. Because five-fold stratified cross-validation was used, the main results are reported as mean $\pm$ standard deviation across folds. M2WPR-Net achieved an accuracy of $93.5\% \pm 0.7\%$ and a macro-F1 score of $87.3\% \pm 1.0\%$. Compared with the strongest unimodal baseline (Visual-CNN) and the late-fusion multimodal baseline,

paired statistical testing showed significant improvements, with $p = 0.004$ and $p = 0.002$, respectively. These results suggest that the observed performance gains are unlikely to be attributable to a single favorable fold split.

False-alarm analysis. For practical online monitoring, the sample-level false-positive rate was converted into an expected alarm burden over a one-hour process. Using the deployment decision frequency of 0.2 Hz after temporal smoothing and a false-positive rate of 0.006 for normal welds incorrectly classified as defective, the expected number of false alarms is $0.2 \times 3600 \times 0.006 = 4.32$ alarms per hour. Although the model can process approximately 39 samples per second computationally, alarm decisions are updated at 0.2 Hz after temporal smoothing to reduce transient false alarms. This value provides an application-oriented estimate of operator alarm burden rather than only reporting sample-level classification accuracy.

Inference efficiency. Real-time feasibility was evaluated with batch size 1 under the same hardware environment as the main experiments. The average GASF preprocessing latency was 3.8 ms per sample, the M2WPR-Net forward inference latency was 21.6 ms per sample, and the total computational latency was 25.4 ms per sample, excluding physical sensor exposure/acquisition time. This latency corresponds to approximately 39 frames per second and satisfies the online monitoring requirement of the proposed welding-quality assessment scenario.

4.4.2. Complete Training Dynamics

Figure 10 plots M2WPR-Net's training progression across 30 epochs. As expected, the training loss drops steeply during the first 15 epochs before leveling off. Interestingly, the validation loss shows a brief spike near epoch 9. This points to a temporary overfitting phase, which our regularization strategies (dropout and weight decay) successfully brought under control as training continued.

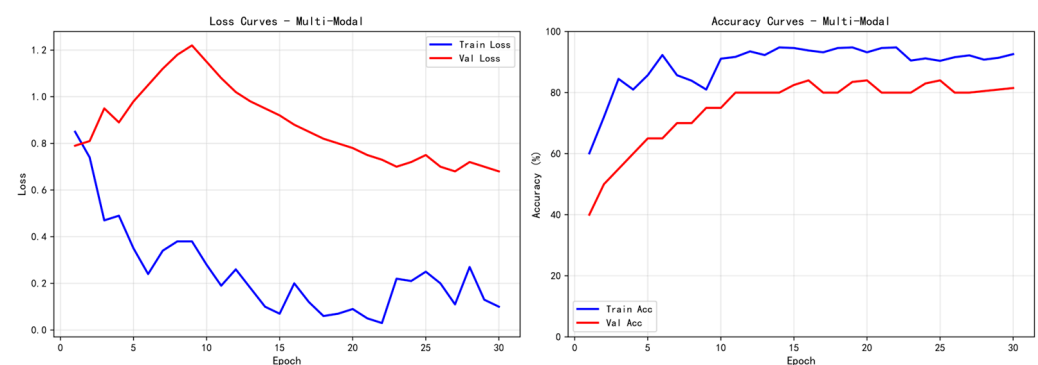


Figure 10. The complete training trajectory of M2WPR-Net.

While training accuracy increases rapidly during the early epochs, the validation curve improves more gradually and is used only to monitor convergence and early stopping in a representative fold.

4.4.3. Attention Mechanism Visualization and Interpretability Analysis

To better understand M2WPR-Net's decision-making from a process monitoring perspective, we examined the attention maps generated by the CBAM module under normal, moderate-interference, and severe-interference conditions (Figure 11). Because raw production images are subject to data-use restrictions, Figure 11 is described as an anonymized feature-level visualization rather than as a direct overlay on raw weld-pool images. It illustrates the relative shift in attention from visual regions to acoustic cues when visual quality degrades.

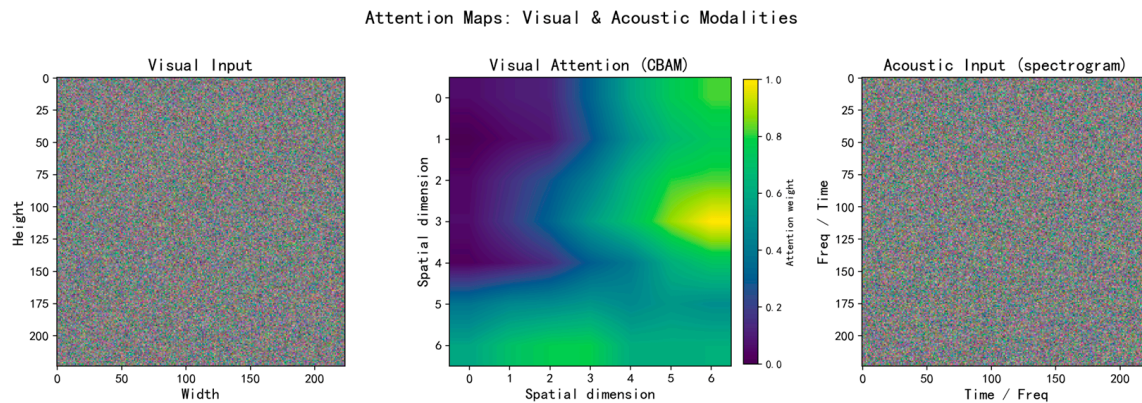


Figure 11. Anonymized feature-level visualization of visual and acoustic attention distributions.

Visual Attention: Under normal circumstances, the model clearly focuses on the weld pool rear boundary, the arc plasma, and the symmetry axis, with attention weights exceeding 0.9. These specific regions are crucial for determining bead width and arc stability. Interestingly, when simulated arc glare obscures the upper image, the network immediately ignores these saturated areas (weights drop below 0.2) to block noise propagation. A smooth red-to-blue gradient in the maps further confirms that the spatial weighting remains stable.

Acoustic and Cross-Modal Focus: The network exhibits strong adaptability during inference. If visual quality drops, it dynamically shifts its focus to rely more on acoustic signals. This shift is clearly reflected in the cross-modal correlation coefficient, which decreases from 0.78 under normal conditions to 0.31 during severe interference. Such a clear drop highlights the model's capacity to decouple modalities and adjust weights on the fly.

5. Conclusions

To address the inherent limitations of single-modal sensing in demanding Gas Metal Arc Welding (GMAW) scenarios, this paper introduced M2WPR-Net. Rather than relying on isolated sensory inputs, our architecture synthesizes visual and acoustic data through a dual-stream backbone. By embedding SE and CBAM mechanisms alongside a dedicated cross-modal attention module, the network not only isolates crucial spatial and channel features but also dynamically bridges heterogeneous signals. This design ensures stable performance even when subjected to severe industrial disturbances like arc glare, spatter, and intense background noise.

Performance evaluations confirmed the reliability of the system, achieving a classification accuracy of 93.5% and a macro-F1 score of 87.3% for quality identification. Subsequent ablation studies systematically elucidated the specific contributions of our core design choices, which include attention modules, fusion strategies, and physical consistency constraints. These experiments demonstrated the individual value of each component across both classification and regression tasks. Crucially, interpretability analyses verified that M2WPR-Net grounds its predictions on physically meaningful regions, primarily the weld pool boundaries and the arc plasma. This finding confirms that the decision-making process of the model aligns with expert domain knowledge instead of relying on spurious statistical artifacts.

Ultimately, M2WPR-Net offers a dependable and transparent framework for on-site welding inspection within the tested GMAW experimental environment. Nevertheless, the current validation is limited to the available experimental setup and has not yet covered multiple independent factories or production lines. Future work will focus on broader multi-site industrial validation, architecture streamlining for lower latency,

modeling temporal dynamics, and introducing thermal or electrical current data into the multimodal framework.

Author Contributions: Methodology and writing—original draft preparation, A.H.; formal analysis, T.Z.; resources, Y.P.; data duration, H.C.; software, J.Z.; investigation, H.Y.; writing—review and editing and supervision, P.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Key Project of the National Undergraduate Innovation and Entrepreneurship Training Program of Nantong University under Grant 2026103040043, and the Project of State Grid Jiangsu Electric Power Co., Ltd. Ultra High Voltage Branch under Grant J2026090.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon reasonable request, subject to institutional and industrial confidentiality restrictions. Representative anonymized samples, preprocessing details, and the labeling protocol are provided in Github repository where permitted.

Acknowledgments: All individuals included in this section have consented to the acknowledgement.

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

References

- Li, H.; Ma, Y.; Duan, M.; Wang, X.; Che, T. Defects detection of GMAW process based on convolutional neural network algorithm. *Sci. Rep.* **2023**, *13*, 21219. [CrossRef] [PubMed]
- Lopez, A.; Bacelar, R.; Pires, I.; Santos, T.G.; Sousa, J.P.; Quintino, L. Non-destructive testing application of radiography and ultrasound for wire and arc additive manufacturing. *Addit. Manuf.* **2018**, *21*, 298–306. [CrossRef]
- Santoro, L.; Sesana, R.; Molica Nardo, R.; Curá, F. Infrared in-line monitoring of flaws in steel welded joints: A preliminary approach with SMAW and GMAW processes. *Int. J. Adv. Manuf. Technol.* **2023**, *128*, 2655–2670. [CrossRef]
- Li, T.; Cao, Y.; Zhang, Y. Analysis of weld pool region constituents in GMAW for dynamic reconstruction through characteristic enhancement and LSTM U-Net networks. *J. Manuf. Process.* **2024**, *127*, 573–588. [CrossRef]
- Sassi, P.; Tripicchio, P.; Avizzano, C.A. A smart monitoring system for automatic welding defect detection. *IEEE Trans. Ind. Electron.* **2019**, *66*, 9641–9650. [CrossRef]
- Xia, L.; Zhou, J.; Xue, R.; Li, X.; Liu, H. Real-time seam tracking during narrow gap GMAW process based on the wide dynamic vision sensing method. *J. Manuf. Process.* **2023**, *101*, 820–834. [CrossRef]
- Xu, H.; Yan, Z.H.; Ji, B.W.; Huang, P.F.; Cheng, J.P.; Wu, X.D. Defect detection in welding radiographic images based on semantic segmentation methods. *Measurement* **2022**, *188*, 110569. [CrossRef]
- Cullen, M.; Ji, J.C. Online defect detection and penetration estimation system for gas metal arc welding. *Int. J. Adv. Manuf. Technol.* **2025**, *136*, 2143–2164. [CrossRef]
- He, K.; Li, J.; Li, Z.; Tao, W.; Huang, J.; Zhou, R. Detection method of austenitic stainless steel laser welding quality based on the AE signal. *Nondestruct. Test. Eval.* **2025**, 1–25. [CrossRef]
- Jabeen, S.; Li, X.; Amin, M.S.; Bourahla, O.; Li, S.; Jabbar, A. A review on methods and applications in multimodal deep learning. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 1–41. [CrossRef]
- Zhao, F.; Zhang, C.; Geng, B. Deep multimodal data fusion. *ACM Comput. Surv.* **2024**, *56*, 1–36. [CrossRef]
- Huang, Y.-W.; Gao, X.-D.; Gao, P.P.; Ma, B.; Zhang, Y.-X. Laser welding monitoring techniques based on optical diagnosis and artificial intelligence: A review. *Adv. Manuf.* **2025**, *13*, 337–361. [CrossRef]
- Madhvacharyula, A.S.; Pavan, A.V.S.; Gorthi, S.; Chitral, S.; Venkaiah, N.; Kiran, D.V. In situ detection of welding defects: A review. *Weld. World* **2022**, *66*, 611–628. [CrossRef]
- Erfani, S.M.; Rajasegarar, S.; Karunasekera, S.; Leckie, C. High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattern Recognit.* **2016**, *58*, 121–134. [CrossRef]
- Kshirsagar, R.; Jones, S.; Lawrence, J.; Tabor, J. Prediction of bead geometry using a two-stage SVM-ANN algorithm for automated tungsten inert gas (TIG) welds. *J. Manuf. Mater. Process.* **2019**, *3*, 39. [CrossRef]
- Wu, D.; Hu, M.; Huang, Y.; Zhang, P.; Yu, Z. In situ monitoring and penetration prediction of plasma arc welding based on welder intelligence-enhanced deep random forest fusion. *J. Manuf. Process.* **2021**, *66*, 153–165. [CrossRef]
- Yang, G.; Guan, K.; Zou, L.; Sun, Y.; Yang, X. Weld defect detection of a CMT arc-welded aluminum alloy sheet based on arc sound signal processing. *Appl. Sci.* **2023**, *13*, 5152. [CrossRef]

18. Chen, C.; Xiao, R.; Chen, H.; Lv, N.; Chen, S. Prediction of welding quality characteristics during pulsed GTAW process of aluminum alloy by multisensory fusion and hybrid network model. *J. Manuf. Process.* **2021**, *68*, 209–224. [\[CrossRef\]](#)
19. Nam, K.; Ki, H. One camera-based laser keyhole welding monitoring system using deep learning. *J. Manuf. Process.* **2023**, *104*, 17–27. [\[CrossRef\]](#)
20. Gao, A.; Fan, Z.; Li, A.; Le, Q.; Wu, D.; Du, F. YOLO-weld: A modified YOLOv5-based weld feature detection network for extreme weld noise. *Sensors* **2023**, *23*, 5640. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Cong, S.; Zhou, Y. A review of convolutional neural network architectures and their optimizations. *Artif. Intell. Rev.* **2023**, *56*, 1905–1969.
22. Wei, W.; Liu, Y.; Wu, J.; Wei, Z.; Zhou, Z.; Long, Y. In-situ monitoring method of femtosecond laser welding between glass and copper with acoustic emission. *Measurement* **2025**, *240*, 115568. [\[CrossRef\]](#)
23. Liu, L.; Chen, H.; Chen, S. Quality analysis of CMT lap welding based on welding electronic parameters and welding sound. *J. Manuf. Process.* **2022**, *74*, 1–13. [\[CrossRef\]](#)
24. Mobaraki, M.; Ahani, S.; Gonzalez, R.; van Heusden, K.; Dumont, G.A. Multi-modal data improves real-time defect classification deep learning models for fillet joints in gas metal arc welding. *Int. J. Adv. Manuf. Technol.* **2025**, *138*, 5309–5317. [\[CrossRef\]](#)
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
26. Tsai, Y.-H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.-P.; Salakhutdinov, R. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Florence, Italy, 2019; pp. 6558–6569.
27. Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; Sun, C. Attention bottlenecks for multimodal fusion. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 14200–14213.
28. Wang, Z.; Oates, T. Imaging time-series to improve classification and imputation. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2015; pp. 3939–3945.
29. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE Computer Society: Los Alamitos, CA, USA, 2018; pp. 7132–7141.
30. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2018; pp. 3–19.
31. Girshick, R. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2015; pp. 1440–1448.
32. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.

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.