Journal of Manufacturing Processes 156 (2025) 252–267

Contents lists available at ScienceDirect

Journal of Manufacturing Processes

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

An effective penetration depth and width prediction method in pulsed GTA welding based on multimodal transformer-serial fusion network

Yuqing Xu, Qiang Liu, Jingyuan Xu, Shanben Chen *

Intelligentized Robotic Welding Technology Laboratory, School of Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200240, PR China

ARTICLE INFO

Keywords:
Prediction of backside weld width
Prediction of penetration depth
Multimodal fusion

ABSTRACT

Sensor technology application is the key for intelligent welding quality monitoring(WQM). Multimodal sensor information fusion has shown their significant advantages over single sensor which can only provide limited information. However, given the fusion process of different sensor information, the feature mining mechanism of the deep learning model is still under-explained, and the direct complementarity of the different information remains unclear. In this paper, a multimodal sensor feature fusion methodology was presented to automatically evaluate backside weld width and penetration depth in real time for Al alloy in pulsed gas tungsten arc welding by means of online molten pool images, arc sound signals and infrared thermal images. Based on the developed feature extraction algorithms in low and high frequencies, multimodal features were successively extracted from raw signals. After the synchronization of multimodal heterogeneous information, the multimodal feature fusion was conducted by establishing a network called AM-TSFNet. The test results indicate that multimodal sensor-based network achieves a higher R^2 value of 0.97 and a lower MSE of 0.16 than single sensor-based one in term of prediction of backside weld width and penetration depth. This paper serves as a progressive extension of our previous work on multispectral channel attention mechanisms for welding state prediction. While the former focused on classifying typical weld states using multimodal features, the present work advances toward precise regression of quantitative weld quality metrics, such as backside weld width and penetration depth.

1. Introduction

In the era of Industry 5.0, artificial intelligence (AI) has emerged as a primary solution to address the complexity and the unpredictability present in contemporary manufacturing systems. As a vital step in industrial production, welding is experiencing considerable evolution. Manual welding is gradually being substituted by robotic systems because of its inefficiency and lack of uniformity [1]. Nonetheless, the majority of current welding robots still rely on the teach-and-repeat model, which restricts their flexibility when dealing with intricate welding scenarios and fluctuating process conditions. Consequently, there is a growing demand for intelligent welding solutions in modern industry.

Welding quality monitoring(WQM) serves as a fundamental enabling technology for intelligent welding manufacturing, ensuring both high efficiency and superior quality [2,3]. Among various quality indicators, the backside weld width and penetration depth are key determinants of WQM. Accurate monitoring of penetration depth is essential in welding research to ensure full fusion, prevent defects such as lack of penetration or porosity, and optimize process parameters for improved joint reliability [4]. To achieve reliable WQM, it is essential to develop an efficient and robust detection method capable of accurately predicting these metrics from sensor data. Such a system would facilitate real-time adjustment of process parameters—such as current and welding speed—thereby maintaining the weld within an optimal operational state.

To enable real-time WQM, a variety of sensing technologies have been adopted. For instance, microphones have been utilized to acquire acoustic signals [5]. Visual sensors such as charge-coupled devices (CCDs) [6], complementary metal-oxide-semiconductor (CMOS) sensors [7], high-dynamic range (HDR) cameras [8], and high-speed cameras equipped with special filters [9] have been employed to capture images of the molten pool. Spectrometers [10,11] are used to collect optical signals, including visible light (VIS), infrared (IR), and ultraviolet (UV) spectra. Thermal information can be obtained through IR cameras [12,13] and coaxial pyrometers [14]. However, considering that

* Corresponding author.
E-mail address: sbchen@sjtu.edu.cn (S. Chen).

https://doi.org/10.1016/j.jmapro.2025.10.116
Received 26 May 2025; Received in revised form 20 September 2025; Accepted 30 October 2025
Available online 6 November 2025

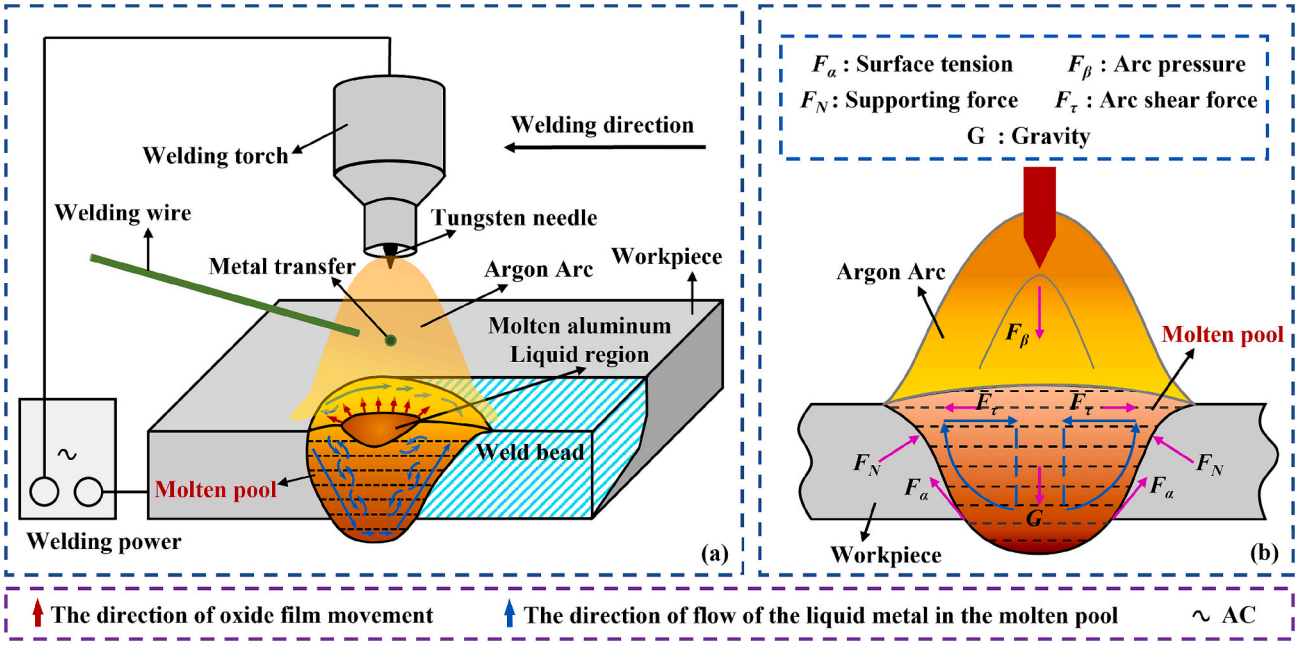

Fig. 1. Illustration of the P-GTAW process (inspired by the design in [49]).

welding is a highly nonlinear and time-varying process, single sensor deployed in these studies can only capture partial aspects of the welding state. Moreover, such sensors are highly susceptible to various disturbances, such as spatter, electromagnetic interference, acoustic noise, and intense arc light. In contrast, multimodal sensor fusion offers significant advantages over single sensor approaches by leveraging complementary information, thereby greatly enhancing the robustness and stability of the monitoring system. This technique has been widely applied in additive manufacturing [15,16], fault detection [17,18], and has recently shown promising potential in WQM.

Although multimodal sensors can provide richer and more comprehensive information, extracting and fusing heterogeneous data from different sensor types remains a major challenge. Existing multimodal WQM methods can be generally categorized into three types: (1) signal and image processing-based approaches, (2) machine learning (ML)-based approaches, and (3) end-to-end deep learning (DL)-based approaches.

Signal and image processing-based methods extract features such as pool length, width, perimeter, aspect ratio, and area [19,20] by applying filtering, edge detection, thresholding, and morphological operations. Temporal signals such as arc sound, current, voltage, and spectroscopic data are analyzed in time and frequency domains to derive features like energy, root mean square (RMS), variance, kurtosis, peak factor, and the variance ratio of H/Ar [21,22]. However, these methods heavily rely on well-designed data acquisition systems to ensure high signal-to-noise ratios—uniform illumination for imaging and clean, stable waveforms for signal analysis. Consequently, they exhibit limited adaptability to poor imaging conditions and highly dynamic welding environments, making them vulnerable to real-world disturbances.

Traditional ML methods have also been widely applied in WQM. These approaches typically involve the extraction and selection of relevant features during preprocessing, followed by the use of trained ML models to make predictions. For instance, Gao et al. [23] constructed a TAN Bayesian network to predict weld width and relative penetration, thereby enabling weld state classification. Zhang et al. [24] employed a support vector machine (SVM) with 10-fold cross-validation to establish a feature-level fusion classifier for weld defect identification. In another study, a deep belief network (DBN) was developed to monitor welding states, and a genetic algorithm was adopted to optimize the model parameters. Despite their success, ML-based methods largely rely on expert knowledge and the careful design of manually extracted features to achieve optimal performance. The omission of critical features may severely impair the model's fitting and generalization capabilities.

In contrast, DL techniques offer accurate, robust, and end-to-end solutions by automatically learning complex feature representations and patterns from raw inputs. Recently, numerous studies have explored DL-based methods for multimodal weld quality assessment. For example, Gao et al. [25] adopted a DenseNet combined with an atrous spatial pyramid pooling (ASPP) module to predict the backside melting width in gas metal arc welding (GMAW). In another work, a Multimodal Continuous Signals Characteristic Reinforcement Network (MCRNet) was proposed to monitor full-penetration groove cold metal transfer (CMT) welding in real time [26]. Zhang et al. [27–29] analyzed and identified the challenges in adaptive robotic welding and proposes a series of modern approaches to address these challenges. An innovative 3-D vision sensing system is used to measure the characteristic parameters such as width, length and convexity of the weld pool in real-time in gas tungsten arc welding. A predictive control algorithm was developed to control the characteristic parameters in a closed form without the need for online optimization. The measured characteristic parameters were used to estimate the backside bead width by an adaptive neuro-fuzzy inference system. However, the presence of redundant or weakly correlated information across modalities can increase the computational burden of the network. Moreover, the harsh welding environment, characterized by extensive noise and disturbances, may impair the model's ability to extract discriminative features, limiting its effectiveness in identifying informative patterns.

This paper focuses on robotic pulsed gas tungsten arc welding (P-GTAW), an essential technique for aluminum alloy joining. Owing to its advantages—precise heat input control, narrow heat-affected zone, stable arc behavior, and low oxidation risk—P-GTAW has been widely adopted in industries such as aerospace, automotive manufacturing, and shipbuilding [1,2]. The weld quality of joints is critical to the performance and reliability of these products. However, due to the unique thermal conductivity and pool dynamics of aluminum alloys, the molten metal flow within the molten pool exhibits highly complex behavior. The final weld formation is highly sensitive to heat input distribution, as illustrated in Fig. 1 [49]. Consequently, intelligent technologies and automated monitoring systems are urgently required for real-time quality assessment in P-GTAW processes. Although DL techniques

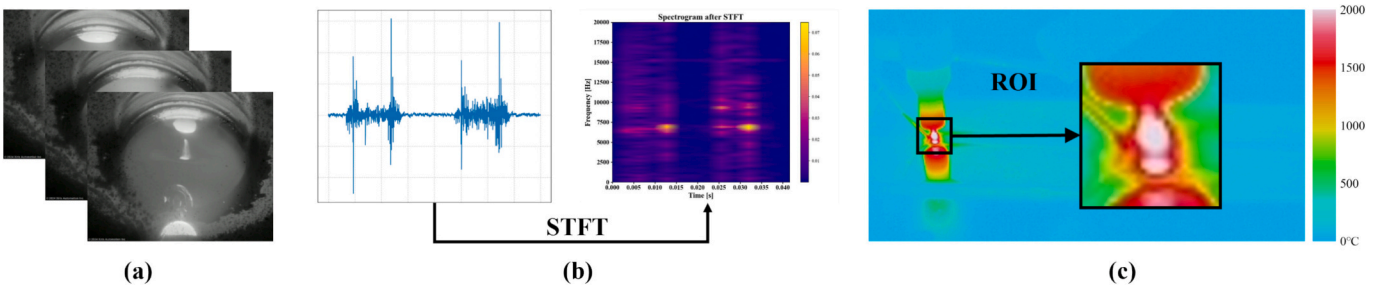

Fig. 2. Representative input samples: (a) molten pool image, (b) arc sound spectrogram, and (c) infrared thermal ROI image.

have demonstrated significant potential in various WQM scenarios, their applicability to welding processes characterized by highly dynamic and complex molten pool behavior—such as P-GTAW—requires further investigation. To address this challenge, this paper proposes an attention-based multimodal Transformer-serial fusion network (AM-TSFNet), designed to predict backside weld width and penetration depth during the P-GTAW process. By effectively integrating global contextual dependencies and local spatial structural features from heterogeneous sensor modalities, AM-TSFNet enables accurate regression of weld formation metrics. The proposed method has been experimentally validated through extensive welding trials, demonstrating its feasibility and effectiveness. It holds great promise for broader applications in WQM in manufacturing, supporting the real-time adjustment or shutdown of process parameters and laying the groundwork for closed-loop control systems.

This paper serves as a progressive extension of our previous research [3], which proposed MFCA-Net to classify typical welding states. While MFCA-Net demonstrated high classification accuracy for welding states, it focused primarily on discrete state identification. This paper advances this research by introducing AM-TSFNet aimed at accurately predicting quantitative metrics of WQM, namely backside weld width and penetration depth. In this paper, WQM specifically refers to the real-time prediction and estimation of backside weld width and penetration depth based on multimodal sensing data.

In summary, the main contributions of this paper are as follows:

1. A novel hybrid architecture, AM-TSFNet, is proposed for the prediction of weld penetration depth and backside weld width. The architecture integrates Transformer and partial convolution modules in series, enabling the extraction of both global contextual features and local structural information from multimodal welding data. This enhances the model's representational capacity in complex, dynamic welding processes.
2. A Soft Thresholding Shrinkage Layer (STSL) is developed to adaptively suppress noise-related features while emphasizing informative ones, thereby improving model robustness and interpretability under high-noise welding conditions. In addition, an attention-based cross-modal fusion module is designed to selectively integrate complementary information derived from molten pool images, spectrograms, and infrared thermal images, facilitating more effective intermodal information exchange.
3. The robustness of AM-TSFNet is evaluated under various welding conditions. The model demonstrates strong generalization capabilities on datasets collected under different scenarios and performs well on unseen data. Moreover, its effectiveness is thoroughly validated through ablation studies and comparisons with mainstream baseline models.

The remainder of this paper is organized as follows. Section 2 presents the details of the proposed AM-TSFNet architecture and algorithm design. Section 3 describes the experimental platform and the

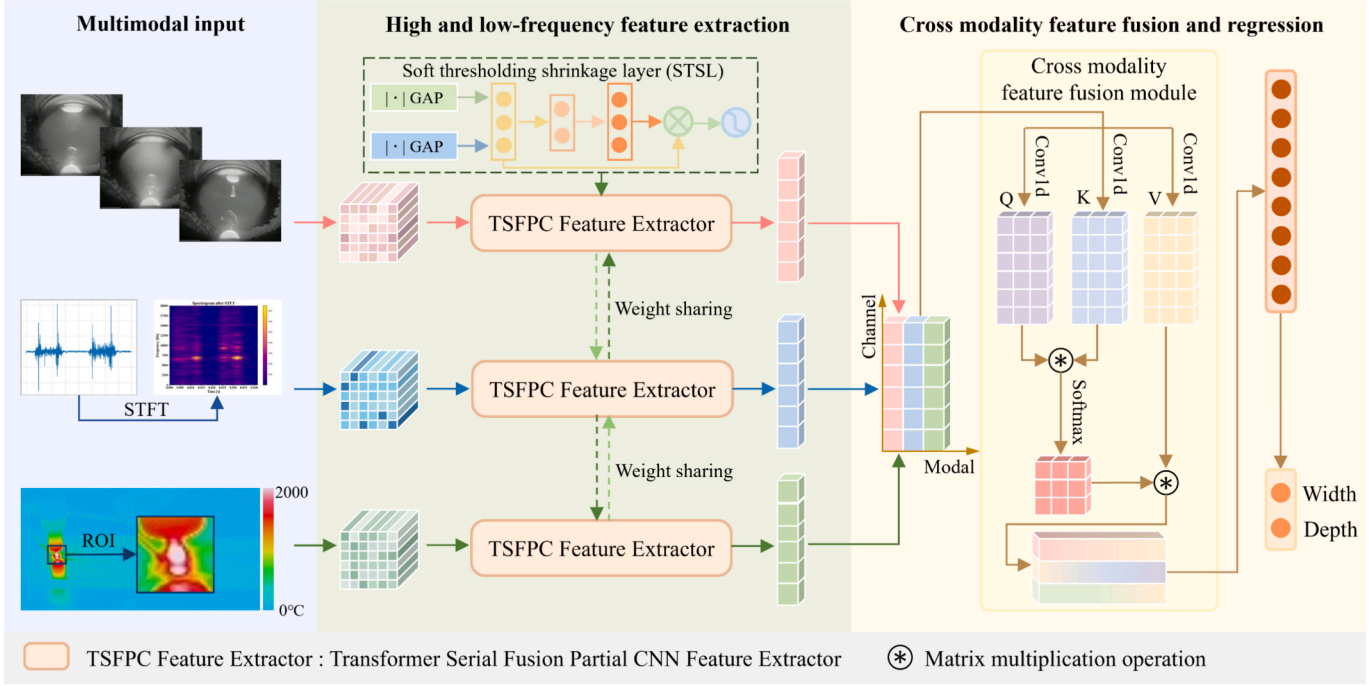

Fig. 3. Architecture of proposed AM-TSFNet.

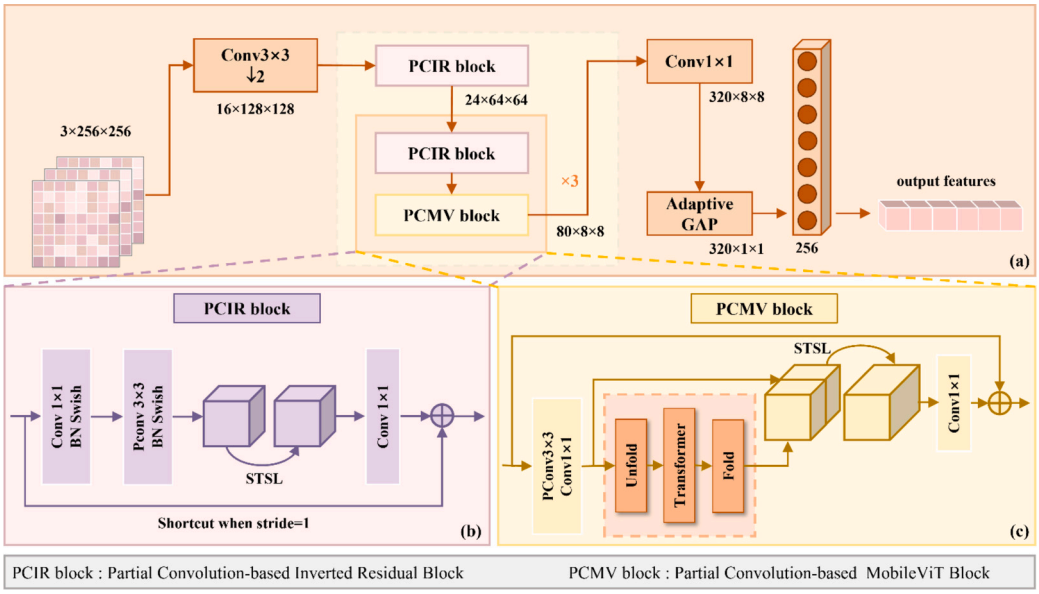

Fig. 4. Architecture of proposed TSFPC feature extractor (a) Overall framework (b) PCIR block framework (c) PCMV block framework.

acquisition of multimodal sensor data. Section 4 provides experimental validation and performance analysis of the model on aluminum alloy P-GTAW data, along with comparative evaluations and ablation studies. Finally, Section 5 concludes the paper.

2. Architecture of AM-TSFNet

2.1. Framework overview

As illustrated in Fig. 3, the proposed attention-based multimodal Transformer-serial fusion network (AM-TSFNet) framework consists of three main components: multimodal input processing, high and low-frequency feature extraction, and cross-modality feature fusion and regression. Specifically, three types of sensory data—molten pool images, acoustic signals, and infrared thermal images—are used as inputs. The acoustic signals are first converted into time-frequency representations via Short-Time Fourier Transform (STFT), while Regions of Interest (ROI) are extracted from the IR images.

Representative samples from the three input modalities are shown in Fig. 2. Molten pool images exhibit distinct pool boundaries and arc structures, while arc sound spectrograms reveal temporal-frequency patterns associated with arc dynamics. Specifically, the arc sound signals were sampled at 40 kHz and transformed using STFT with a Hamming window of 256 samples (6.4 ms) and 50 % overlap, yielding a time resolution of 3.2 ms and a frequency resolution of 156.25 Hz. As shown in Fig. 2(b), the spectrograms depict the temporal evolution of frequency components, where the horizontal axis represents time (ms) and the vertical axis represents frequency (Hz). To facilitate interpretation, an amplitude scale bar indicating the relative spectral power (in dB, relative to the reference sound pressure of 20 μPa) is added, showing the range between the minimum and maximum amplitude values, which is consistent with common practices reported in the literature [30,31]. Infrared thermal images capture heat distribution and are preprocessed via ROI segmentation. These inputs are resized to 256 × 256 and synchronized based on frame index. All three modalities are used as input channels during model training and testing.

Each modality is then processed through a shared feature extraction module, which integrates a soft-thresholding shrinkage layer (STSL) to adaptively suppress redundant low-frequency components and enhance informative high-frequency features. To ensure consistency and reduce model complexity, weight sharing is applied across all modality-specific feature extractors.

Subsequently, the extracted features are fused using the proposed Cross-Modality Feature Fusion Module, which employs a self-attention mechanism based on query (Q), key (K), and value (V) projections to model long-range dependencies and cross-modal interactions. Finally, the fused representation is fed into a regression head to simultaneously predict the backside weld width and penetration depth, providing accurate and interpretable WQM.

To prevent overfitting and improve the model's generalization capability, several regularization strategies were employed. Dropout layers with a rate of 0.5 were inserted in the fully connected layers to reduce co-adaptation of neurons. Additionally, L2 regularization (weight decay of 1e-4) was applied to all trainable parameters. Early stopping with a patience of 10 epochs was also used to terminate training once the validation loss stopped improving.

Note that in this study, the term ‘feature fusion’ is used in the context of machine learning to denote the integration of multiple data modalities or features. It should not be confused with ‘fusion welding,’ which refers to the metallurgical joining of materials through melting.

2.2. Architecture of the proposed feature extractor

2.2.1. Transformer serial fusion partial CNN (TSFPC) feature extractor

To comprehensively capture features from multimodal inputs and improve long-term prediction accuracy, a novel feature extractor—-Transformer Serial Fusion Partial CNN (TSFPC)—is proposed in this paper Specifically, Transformer modules [32] are first employed to extract localized features across spatial and time-frequency domains, followed by convolutional neural networks (CNNs) for modeling long-range dependencies. Transformer-CNN hybrid frameworks have been widely applied in various domains such as facial expression recognition and autonomous driving [33,34], demonstrating strong feature extraction capabilities and high prediction accuracy across multiple tasks. To further enhance computational efficiency, reduce frequent memory access, and eliminate redundant operations, partial convolution (PConv) [35] is adopted in place of standard convolution. PConv performs convolution only on a subset of input channels, while leaving the remaining channels unchanged, thus achieving a favorable balance between efficiency and representational capacity. In the proposed TSFPC feature extractor, local features derived from the Transformer and global features extracted by PConv are serially fused, as illustrated in Fig. 4, to

Table 1
Structure of the proposed TSFPC feature extractor.

Layer	Output size	Kernel	Stride	Padding	Output Channels
RGB Image	256*256	–	–	–	3
Conv1	128*128	3*3	2	1	16
PCIR Block	64*64	–	–	–	24
PCIR+PCMV Block1	32*32	–	–	–	48
PCIR+PCMV Block2	16*16	–	–	–	64
PCIR+PCMV Block3	8*8	–	–	–	80
Conv2	8*8	1*1	1	0	320
Adaptive GAP layer	1*1	–	–	–	320
Fully connected layer	–	–	–	–	256

enhance the overall expressive power of the network. The architecture design of TSFPC takes into account spatial inductive biases and convergence behavior of deep models [36]. Furthermore, to mitigate the impact of noise and enhance both the interpretability and generalization ability of the TSFPC module, a Soft Thresholding Shrinkage Layer (STSL) is incorporated. This layer is specifically designed to strengthen the feature learning capability from high-noise sensor inputs.

As shown in Fig. 4(a), the proposed TSFPC feature extractor consists of a 3 × 3 convolutional layer, four PCIR blocks, three PCMV blocks, a 1 × 1 convolutional layer, an adaptive global average pooling layer, and a fully connected layer. Initially, the module receives a 256 × 256 RGB image with three channels as input. This input is processed by a convolutional layer utilizing a 3 × 3 kernel and a stride of 2 to achieve down-sampling [37,38]:

$$y^{l(i,j)} = K_i^l * x^{l(r^j)} = \sum_{j}^{\omega-1} K_i^{l(j^i)} x^{l(j+j')} \quad (1)$$

In the formula, $K_i^{l(j^i)}$ represents the j^i th element of the ith convolution filter in the l th layer. The term $x^{l(r^j)}$ denotes the j th segment in the l th layer involved in the convolution. The symbol ω indicates the width of the convolution filter, as illustrated in Fig. 4(b) and (c).

By stacking PCIR and PCMV blocks, the framework achieves sequential integration of global and local feature representations. The data flow of the proposed TSFPC feature extractor is summarized in Table 1.

As shown in Fig. 4(b), a 1 × 1 convolutional layer is first applied to increase the dimensionality of the input feature maps. For local feature modeling, the PCIR block utilizes partial convolution, which introduces spatial inductive bias and accelerates network convergence. Finally, a 1 × 1 convolution is employed to reduce the dimensionality of the output features.

In the PCMV block, a Mobile Transformer operator is used to capture global features. As illustrated in Fig. 4(c), the process begins with a partial convolution layer (3*3 kernel), followed by a standard 1*1 convolution. The input image is subsequently segmented into patches via the Unfold operation, and these segments are augmented with positional information. The resulting data is processed by a Transformer encoder and reassembled through the Fold mechanism. This approach effectively limits redundant data and lowers computational complexity. Subsequently, the convolutional and transformer features are merged along the channel to realize feature fusion. Finally, the feature dimensionality is adjusted by a convolution operation with kernel size of 1 × 1.

To avoid gradient vanishing and solve the degeneracy problem, we introduced identity shortcuts in both blocks, allowing the gradients to directly flow back to the initial layers. At the end of the TSFPC feature extractor, a 1 × 1 convolutional layer is first employed to adjust the number of channels, followed by an adaptive global average pooling layer to reduce the spatial dimensions. Finally, a fully connected layer is used to produce the compact output feature representation.

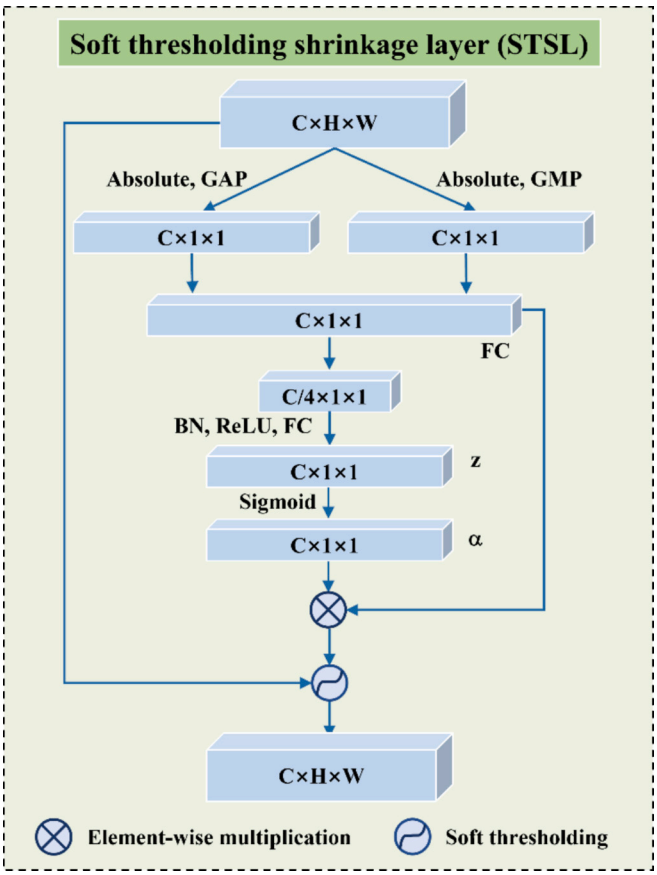

Fig. 5. Architecture of the Soft Thresholding Shrinkage Layer (STSL).

Unlike early fusion strategies that directly concatenate multimodal features at the input or intermediate levels, the proposed serial fusion architecture in TSFPC first extracts and refines modality-specific features independently and then integrates them hierarchically through Transformer and partial CNN blocks. This staged fusion helps reduce feature interference across modalities, preserve modality-specific information, and enhance the model's ability to focus on discriminative features. By sequentially combining global representations from Transformers with local structural cues from CNNs, the network benefits from both long-range contextual awareness and spatial inductive biases. This serial fusion design has shown superior performance in multimodal tasks, especially under noisy and complex data conditions of welding monitoring.

2.2.2. *The soft thresholding shrinkage layer (STSL)*

The multimodal information collected from welding sites often contain a significant amount of noise. When processing high-noise information, the feature learning capability of TSFPC feature extractor tends to degrade. In TSFPC feature extractor, partial convolution and the Transformer operator are used as feature extractors, but due to the interference from noise, welding-state-related features may not be detected effectively. In such cases, the high-level features learned at the output layer are often insufficient for accurate classification of welding states. To address this issue, we design the Soft Thresholding Shrinkage Layer (STSL) to enhance the feature learning ability from high-noise information, aiming to improve the prediction performance of backside weld width and penetration depth. The architecture of the proposed STSL is illustrated in Fig. 5.

The STSL module is applied to feature maps from all input

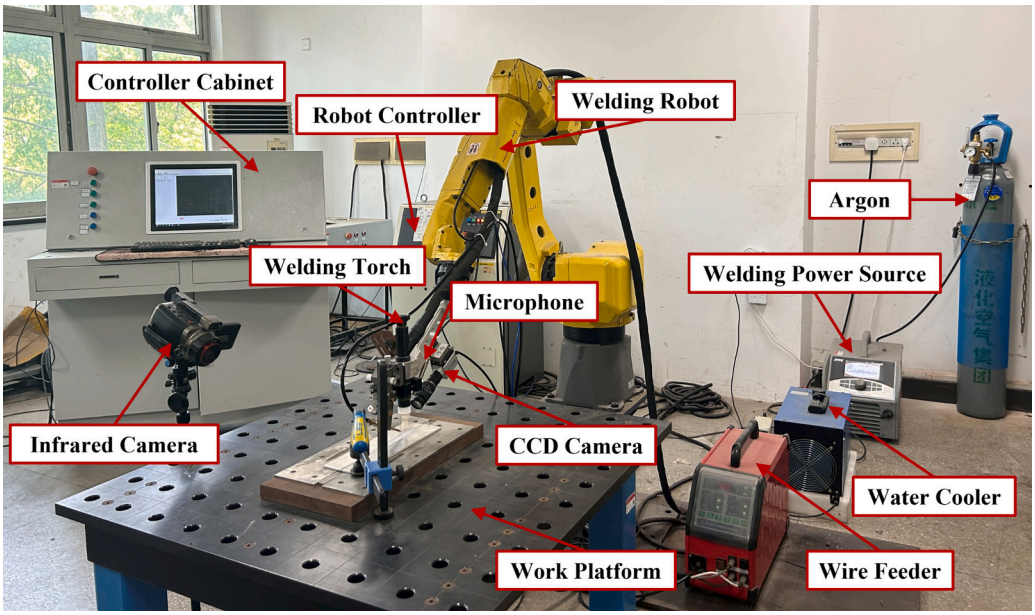

Fig. 6. Schematic diagram of P-GTAW experiment system.

modalities, including visual, acoustic, and infrared signals. Near-zero activations typically stem from background noise, non-weld regions, or low-salience frequency components. These features generally lack discriminative value, and soft thresholding effectively suppresses them while retaining physically meaningful negative responses, improving signal-to-noise separation in feature space.

Specifically, we employ soft thresholding as a nonlinear transformation layer, which is inserted into PCIR blocks and PCMV blocks to suppress irrelevant or redundant features, thereby enabling the model to focus more effectively on information that is closely related to weld prediction. The function of soft thresholding can be expressed by [39].

$$y = \begin{cases} x - \tau & x > \tau \\ 0 & -\tau \le x \le \tau \\ x + \tau & x < -\tau \end{cases} \tag{2}$$

where x is the input feature, y is the output feature, and τ is the threshold, i.e., a positive parameter. Instead of setting the negative features to zero in the ReLU activation function, soft thresholding sets the near-zero features to zeros, so that useful negative features can be preserved. Soft thresholding does not set negative features to zero, as in the case of the ReLU activation function, but instead sets features close to zero to zero, allowing useful negative features to be preserved.

Moreover, considering that it is generally challenging to set proper values for the thresholds, the developed STSL integrate a few specialized neural networks as trainable modules to automatically determine the thresholds, so that professional expertise on signal processing is not required. Additionally, each input can have its own set of thresholds.

In this paper, spatial location information and time–frequency domain features play a critical role in both analysis and model training, as derived from molten pool images, arc sound spectrograms, and infrared thermal images. A combined architecture that integrates partial convolutional networks with Transformer components enriches feature representation by utilizing both global context and local details. The integration of STSL further improves the model's robustness to noise and its ability to focus on features relevant to weld penetration.

Moreover, the proposed cross-modal fusion module effectively integrates complementary information across different modalities, allowing the model to fully exploit the rich correlations among heterogeneous sensor inputs. As a result, the proposed AM-TSFNet architecture not only improves the robustness and interpretability of feature learning under noisy conditions but also contributes to achieving more stable and accurate regression performance in predicting backside weld width and penetration depth.

3. Experimental setup

3.1. Experiment platform

As illustrated in Fig. 6, the experimental system consisted of the following components: a P-GTAW system (including a power source for welding, an automatic wire feeder, a cooling unit, and a welding torch), a FANUC six-axis robot arm with a control unit, a work platform equipped with fixtures, 99.99 % pure argon gas, and a data acquisition system comprising a CCD camera, microphone, and IR camera. A control cabinet housing a data acquisition card, programmable logic controller (PLC), and industrial personal computer (IPC) was also integrated.

To perform butt welding, two parallel aluminum alloy plates were fixed side-by-side in a groove on the work platform. The P-GTAW torch was mounted vertically at the end of the robotic arm, positioned directly above the workpieces. The CCD camera and microphone were mounted to move synchronously with the welding torch, capturing the molten pool and arc sound at different spatial positions along the weld seam. Positioned at a distance of 350 mm from the welding torch, the CCD camera and microphone were angled horizontally at 45° and 75° with respect to the workpiece, respectively, to ensure optimal signal acquisition.

The IR camera was fixed laterally beside the work platform due to its hardware design and weight. To cover the full weld length, a large field of view was set, and for each frame, a ROI corresponding to the molten pool was extracted from the infrared thermal images. This ROI changes along the weld seam, and all ROIs were timestamp-aligned with the corresponding molten pool images and acoustic signals, ensuring that each multimodal sample accurately represents the state of the molten pool at a specific location.

Both the CCD camera and the IR camera capture the molten pool region, but the information they record is fundamentally different. The CCD camera acquires visible-light images, reflecting the optical appearance of the molten pool, whereas the IR camera captures infrared thermal images, representing the temperature distribution of the molten pool. Therefore, although the two cameras overlap spatially, they provide complementary information—one optical and the other

Table 2
Basic welding experiment parameters.

Parameters	Value	Parameter	Value
Material type	Aluminum alloy LF6	Welding wire type	ER5183
Joint type	Butt joint	Welding position	Flat
Current polarity	AC Pulse	Plate thickness(mm)	4
AC duty ratio (%)	35	Welding speed (mm/s)	3
AC frequency (Hz)	50	Argon flow (L/min)	15
Pulse duty ratio (%)	50	Welding wire diameter (mm)	1.2
Pulse frequency (Hz)	100	Electrode diameter (mm)	3.2

thermal—enabling the model to leverage both visual and thermal features of the molten pool.

During the welding process, optical images of the molten pool, arc sound signals, and infrared thermal images were captured by the CCD camera, microphone, and IR camera, respectively, and transmitted to the IPC for subsequent processing. The CCD camera operated at a frame rate of 59 fps. Detailed P-GTAW experimental parameters are listed in Table 2.

The welding parameters listed in Table 2 were selected based on a combination of standard practice in P-GTAW for aluminum alloy joints, prior literature guidance, and preliminary experimental trials. In particular, the AC pulse mode was adopted to achieve enhanced arc stability and reduced heat input. Core parameters such as AC duty ratio, pulse frequency, and welding speed were tuned to ensure the generation of all three penetration states (LP, NP, and OP), which is critical for robust regression modeling. For example, the chosen AC pulse frequency of 100 Hz and duty ratio of 50 % are commonly adopted for arc stabilization in aluminum welding. Welding speed (3 mm/s) and gas flow (15 L/min) were set to balance penetration depth and arc shielding. Other parameters such as filler wire type (ER5183), plate thickness (4 mm and 6 mm), and electrode size were chosen in accordance with recommended configurations for aluminum alloy LF6 [40–43].

3.2. Dataset creation

In order to generate data for training, validation, and evaluation of the AM-TSFNet model, multiple P-GTAW experiments were carried out using two aluminum alloy plates joined by butt welding. The detailed welding parameters are listed in Table 4. To ensure accurate synchronization of multimodal data, a unified data acquisition system was used, in which all sensors—including the optical CCD camera, microphone, and IR camera—were triggered by a common synchronization signal. Each multimodal sample was aligned based on timestamp matching to ensure temporal consistency. A post-processing verification step was also performed to discard frames with mismatched timestamps beyond a predefined threshold. Moreover, robustness analysis revealed that the proposed model maintained stable performance even when slight temporal offsets (e.g., ±1 frame) were introduced during testing, indicating that the AM-TSFNet architecture is resilient to minor synchronization errors. Due to the presence of a DC component introduced by the sound regulator, the collected acoustic signals exhibited zero-drift, which must be removed prior to feature extraction. The drift correction is performed using the following formula [44]:

$$x_i^* = x_i - \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3}$$

where x_i represents the original amplitude of the acoustic signal at time index i, and x_i^* denotes the corrected signal after removing the DC offset. The term $\frac{1}{n}\sum_{i=1}^{n} x_i$ calculates the mean value of the signal over the entire sequence of length n, which is subtracted from each sample to eliminate the zero-drift caused by the DC component.

Table 3
Definition of different penetration states.

	LP	NP	OP
c/mm	–	≤15	>15
h_1/mm	–	0.3–2	<0.3
h_2/mm	<0.5	0.5–2.5	>2.5
h_3/mm	–	≤0.9	–

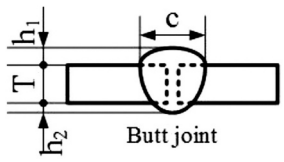

Fig. 7. Classification standards of welding states.

In addition, a high-pass filter was applied to eliminate low-frequency drift in the acoustic signals. To extract the time–frequency characteristics of the arc sound, the recorded raw audio signals were processed using the Short-Time Fourier Transform (STFT). The STFT is defined as follows [45]:

$$\mathrm{STFT}\{x(t)\}(m,\omega) = \int_{-\infty}^{\infty} x(t)\cdot w(t-m)\cdot e^{-j\omega t}dt \tag{4}$$

where $x(t)$ denotes the input time-domain signal, $w(t-m)$ is a window function centered at time m, and ω is the angular frequency.

To accelerate computation, threshold-based segmentation was employed to extract the region of interest containing the molten pool from the infrared thermal images. For each sample, the optical image, arc sound spectrogram, and the segmented infrared thermal ROI from the same frame were combined to form a multimodal input, each resized to 256 × 256 pixels.

To address data imbalance and overfitting, we adopted a targeted augmentation strategy aimed specifically at enriching less LP and OP categories during training. For molten pool images, we employed random horizontal flipping and random cropping within ±10 % of image size. For acoustic spectrograms, slight time-shifting and amplitude scaling (±5 %) were applied. For IR thermal images, Gaussian noise and brightness jitter were introduced to simulate sensor variability. These augmentation strategies enhanced the model's robustness and helped improve generalization performance.

Each input was labeled with two regression targets: backside weld width and penetration depth, both of which were obtained through post-weld metallographic analysis. To generate accurate ground truth labels for training and evaluation, metallographic cross-sections were extracted at 5 mm intervals along the weld seam. Each sample underwent a standardized preparation procedure including mounting, sequential grinding with abrasive papers ranging from 320 to 7000 grit, polishing with magnesium oxide, and etching with Keller's reagent. The prepared sections were imaged using a 4800-B digital metallographic microscope at 24× magnification. Based on this known magnification, pixel-level measurements of weld penetration depth and backside weld width were converted to actual physical dimensions. Given the progressive nature of weld formation in pulsed GTAW, both penetration depth and backside width exhibit smooth variation along the weld seam. Therefore, interpolation-based curve fitting was used to estimate label values between metallographic samples, ensuring that each multimodal sample had a precise and consistent regression target. All welds were categorized based on three penetration states: lack of penetration(LP), normal penetration(NP), and over penetration(OP). This paper refers to ISO 10042:2018 for the quantified standard definition of different penetration states, as shown in Table 3 and Fig. 7. It provides a comprehensive and standardized framework for the identification and classification of penetration states.

Table 4
Detailed welding parameters.

Experiment		I_b (A)	I_p (A)	v_{wf} (cm/min)	Penetration states
#1	No.1	65–80-100	130–160-200	100	LP、NP、OP
	No.2	65–80-100	130–160-200	100	LP、OP
	No.3	80	140	100	LP
	No.4	80	160	100	LP、NP
	No.5	140	200	100	OP
#2	No.1	110	170	100	LP、NP
	No.2	120	180	100	LP、OP

Table 5
Dataset statistics of P-GTAW.

Datasets		Samples
Data 1	Train	8000
	Valid	2000
	Total	10,000
Data 2	Generalizability test	1233

The CCD camera recorded at 59 fps. Each weld seam was approximately 100–120 mm long, and the welding torch traveled at 3 mm/s, giving a welding duration of 33–40 s per seam. This corresponds to approximately 1950–2350 molten pool images per weld seam, depending on the actual welding duration. The arc sound spectrograms and infrared thermal images were synchronized with the molten pool images, producing the same number of multimodal samples per seam.

To assess the performance of the AM-TSFNet model, five welding experiments featuring varied predefined penetration levels were carried out according to the settings described in Experiment #1 in Table 4. After excluding frames with poor synchronization or occlusion, a total of 10,000 high-quality multimodal samples were retained from the five weld seams in Experiment #1 to construct a dataset referred to as *Data 1*, with 80 % used for training and 20 % for validation.

To evaluate the generalization capability of the AM-TSFNet trained on *Data 1*, two additional welding trials with varying penetration states were conducted under the experimental conditions of Experiment #2 in Table 4. Using the same data acquisition and processing pipeline, 1233 high-quality multimodal samples were obtained, forming an independent test dataset referred to as *Data 2*. *Data 1* and *Data 2* were collected under different experimental settings, particularly with distinct values of base current I_b, peak current I_p, and wire feeding speed v_{wf}, ensuring an unbiased assessment of model generalization.

Thus, a total of 11,233 high-quality multimodal samples were obtained. The statistical summary of both datasets is provided in Table 5. Each of the 11,233 multimodal samples consists of one molten pool image from the CCD camera, one infrared thermal ROI image from the IR camera, and one arc sound segment converted into a spectrogram. All three modalities are synchronized by timestamp, so each sample represents the same temporal instance of the welding process across the three sensors.

3.3. Evaluation metrics

To quantitatively evaluate the regression performance of the proposed model, two commonly used metrics were employed: the Mean Squared Error (MSE) and the R-Squared score (R^2). These metrics assess the consistency between the predicted and ground truth values from different perspectives.

Mean Squared Error (MSE) measures the average squared difference between the predicted values and the actual values. A lower MSE indicates that the predicted results are closer to the ground truth. It is defined as [46]:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5}$$

Table 6
The prediction results of the proposed AM-TSFNet.

	Metrics	Test results
Average	MSE	0.16
	R^2	0.97
Width	MSE	0.25
	R^2	0.98
Depth	MSE	0.07
	R^2	0.96

where n is the total number of samples, y_i denotes the ground truth value of the $i-$th sample, and $\hat{y}_i$ represents the corresponding predicted value.

R-Squared score (R^2) evaluates the proportion of variance in the dependent variable that is predictable from the independent variables. An R^2 value closer to 1 indicates better prediction performance. It is calculated as [47]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6}$$

where $\bar{y}$ is the mean of all ground truth values. The numerator quantifies the residual sum of squares, while the denominator measures the total sum of squares.

By jointly considering both MSE and R^2, a comprehensive evaluation of model accuracy and generalization can be achieved. These metrics are adopted in this paper to assess the predictive performance of the proposed model on welding penetration estimation tasks.

3.4. Test environment

The proposed AM-TSFNet was implemented and executed on an industrial IPC running Windows 10, equipped with an Intel Core i7-11700K CPU (3.60 GHz), an NVIDIA GeForce RTX 4060 Ti GPU (16 GB VRAM), and 32 GB of RAM. The training was conducted for 100 epochs. The original images were resized to 256 × 256 pixels. Initially, a classification model was trained to distinguish among three penetration states: LP, NP, and OP. The pre-trained parameters obtained from this classification model were then used to initialize the proposed regression model. During evaluation, the model weights achieving the best performance on the validation set were selected. To ensure reproducibility, a uniform random seed was applied across all experiments.

To evaluate the real-time feasibility of AM-TSFNet, we measured the inference speed. The model achieved an average inference speed of approximately 65 FPS, which is sufficient to meet the 59 FPS acquisition rate of the CCD camera used in the system. Multimodal data streams (optical images, arc sound spectrograms, and infrared thermal images) were synchronized per frame, forming temporally aligned inputs for the model. This indicates that AM-TSFNet can reliably support real-time operation for online quality monitoring in P-GTAW applications. For practical deployment, the model can be further optimized via quantization, pruning, or deployment on embedded systems such as Jetson AGX Orin to reduce latency and memory footprint.

4. Experimental validation

4.1. Effectiveness of the model

The prediction results of the proposed AM-TSFNet on *Data 2* are presented in Table 6. For backside weld width, the model achieved a mean squared error (MSE) of 0.25 mm and a coefficient of determination (R^2) of 0.98. For penetration depth, the MSE reached 0.07 mm, and the R^2 was 0.96. The average MSE across both regression tasks was 0.16 mm, with an overall R^2 of 0.97, demonstrating the strong prediction

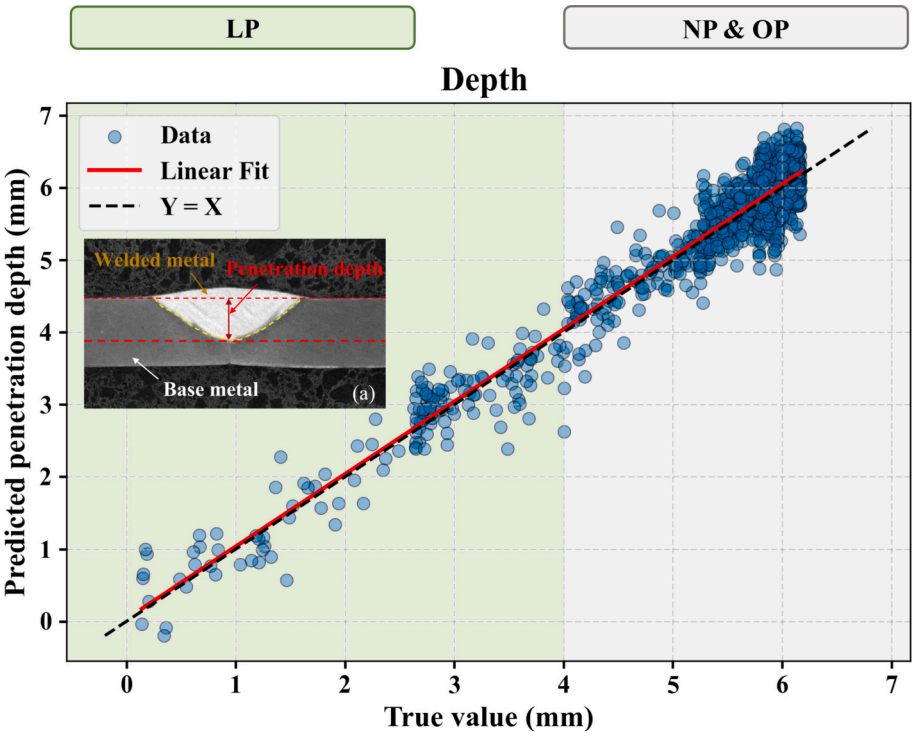

Fig. 8. Prediction results of penetration depth.

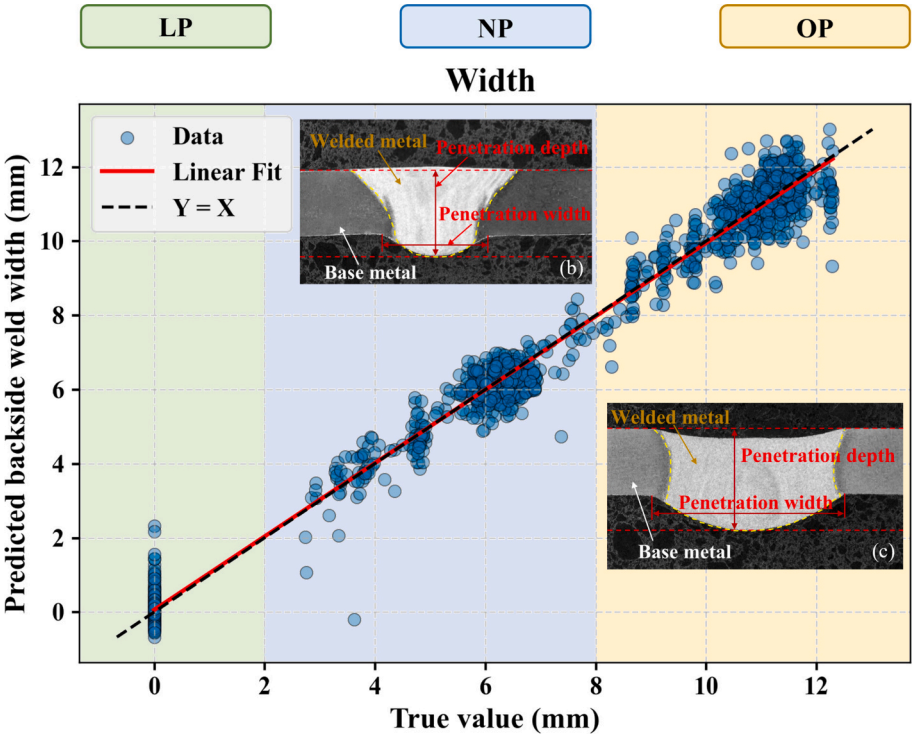

Fig. 9. Prediction results of backside weld width.

accuracy and robustness of the proposed AM-TSFNet in weld quality monitoring tasks.

Figs. 8 and 9 present the regression performance of AM-TSFNet on the tasks of backside weld width and penetration depth prediction. The data distribution is segmented into three regions corresponding to different penetration states: LP, NP, and OP. As illustrated in Fig. 8(a), LP refers to welds where the molten metal fails to fully penetrate the entire thickness of the workpiece in P-GTAW. In contrast, NP denotes welds where the molten metal extends through the entire thickness of the base material, as shown in Fig. 9(b). OP refers to excessive fusion that extends beyond the backside of the workpiece, as depicted in Fig. 9 (c).

Following the butt welding process, a clear boundary is observed between the molten metal and the base material. For LP cases, as shown

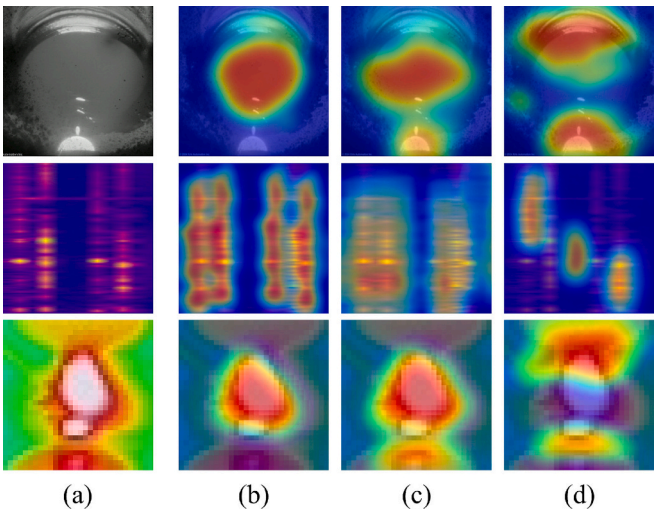

Fig. 10. Grad-CAM visualization across modalities and activation functions: (a) original input, (b) STSL, (c) ReLU, and (d) baseline (no activation).

in Fig. 8(a), the penetration depth is defined as the vertical distance from the top surface of the base material to the deepest point of complete fusion, excluding the partially melted heat-affected zone (HAZ). And the back weld width is zero. For NP and OP conditions, as shown in Fig. 9(b) and (c), a backside weld bead is formed beyond the full thickness of the base metal. In these cases, penetration depth refers to the length from the top surface of the base material to the tip of the backside bead. The backside weld width is defined as the lateral dimension of the fused metal at the back surface of the workpiece. Overall, the term "penetration depth" refers to the vertical distance from the top surface of the base metal to the deepest point of complete fusion, as measured on the cross-sectional metallographic image. The "backside weld width" is defined as the lateral dimension of the fully fused region on the backside surface of the weld, corresponding to the width of the penetration bead. These definitions are consistent with ISO 10042:2018 standards and are illustrated in the metallographic insets within Figs. 8 and 9.

As shown in Fig. 8, most data points lie close to the diagonal line (Y = X), indicating a high level of agreement between the predicted and true values. For the weld width prediction, the model achieved a MSE of 0.25 mm and a R^2 of 0.98, demonstrating excellent fitting capability. Fig. 9 presents the results for penetration depth prediction. Although slight dispersion is observed in the low-value region, the overall trend aligns well with the Y = X reference line. The model achieved an MSE of only 0.07 mm and an R^2 of 0.96, confirming its strong ability to capture depth-related features effectively. Overall, the AM-TSFNet achieved an average MSE of 0.16 mm and an average R^2 of 0.97 across the two regression tasks, which demonstrates the robustness and generalization capability of the proposed AM-TSFNet. These results suggest that AM-TSFNet is well-suited for practical deployment in industrial WQM applications.

As shown in Figs. 8 and 9, the predicted penetration depth and width generally align well with the truth, though some dispersion is observed. This deviation mainly stems from sensor limitations and environmental disturbances. Under LP and NP conditions, weak thermal contrast and partial occlusion (e.g., by the arc torch or filler wire) can degrade the quality of infrared features. In OP scenarios, arc instability introduces fluctuations in both sound and molten pool signals, impairing time-frequency consistency. Additionally, in shallow penetration (< 2 mm), the visual and acoustic cues become less distinctive, lowering the signal-to-noise ratio and making accurate regression more challenging.

To further evaluate the interpretability and effectiveness of the proposed STSL, we conducted visual analyses using both Grad-CAM [48] attention maps and intermediate feature maps across modalities.

We compared three settings: (1) a baseline model without activation, (2) with ReLU, and (3) with STSL. Grad-CAM was employed to generate attention heatmaps shown in Fig. 10 based on the influence of each modality—molten pool images, arc sound spectrograms, and infrared thermal images—on the model's predictions. The results indicate that STSL produces more focused and semantically meaningful attention distributions. For molten pool images, the highlighted areas accurately

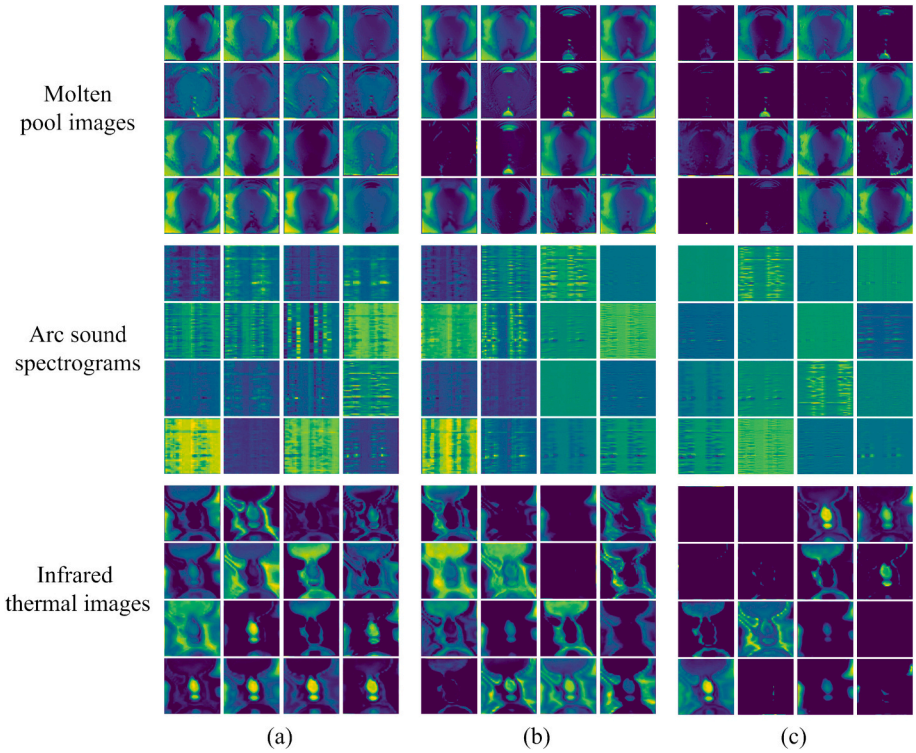

Fig. 11. Comparison of intermediate feature maps with (a) STSL, (b) ReLU, and (c) baseline.

Table 7
Results of ablation study.

Exp. No.	Input			Feature Fusion	Feature Extraction		Width		Depth		Average	
	img	audio	ir		-STSL	+STSL	MSE	R^2	MSE	R^2	MSE	R^2
1	√			–	√		0.98	0.92	0.22	0.82	0.60	0.87
						√	0.76	0.94	0.18	0.86	0.47	0.90
2		√		–	√		1.10	0.91	0.32	0.77	0.71	0.84
						√	1.05	0.92	0.23	0.82	0.64	0.87
3			√	–	√		1.61	0.87	0.31	0.77	0.96	0.82
						√	1.31	0.91	0.29	0.79	0.80	0.85
4	√	√		√	√		0.63	0.94	0.23	0.86	0.43	0.90
						√	0.61	0.95	0.15	0.89	0.38	0.92
5	√		√	√	√		0.71	0.94	0.21	0.84	0.46	0.89
						√	0.46	0.96	0.16	0.88	0.31	0.92
6		√	√	√	√		1.37	0.89	0.23	0.83	0.80	0.86
						√	0.66	0.94	0.22	0.84	0.44	0.89
7	√	√	√	–	√		0.59	0.96	0.17	0.88	0.38	0.92
						√	0.38	0.97	0.12	0.91	0.25	0.94
8	√	√	√	√	√		0.30	0.97	0.10	0.93	0.20	0.95
						√	0.25	0.98	0.07	0.96	0.16	0.97

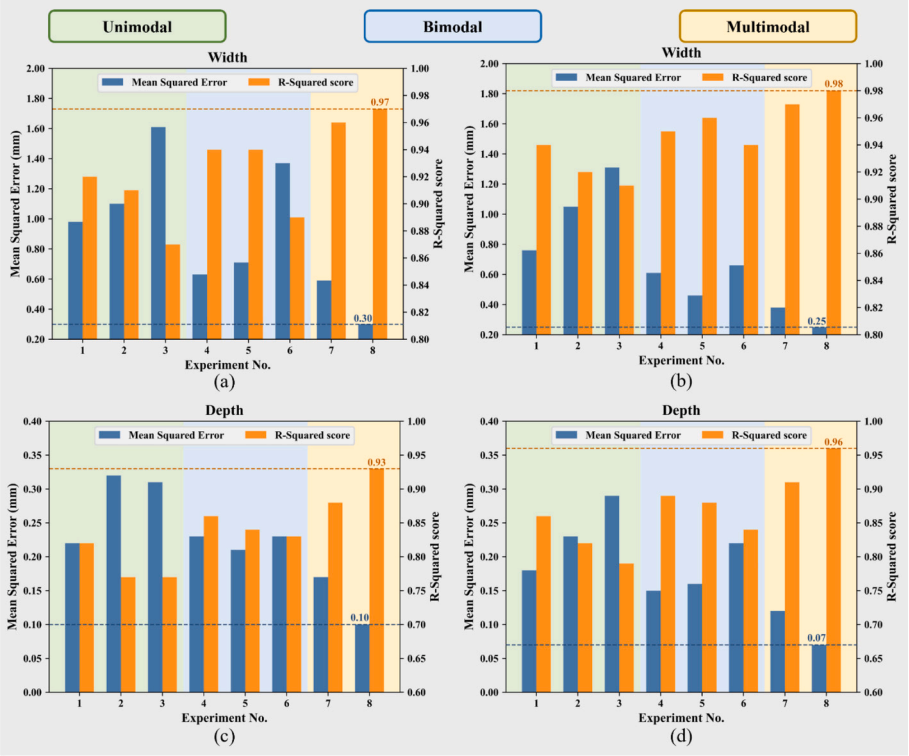

Fig. 12. Ablation study results histogram (a) Prediction results for backside weld width without STSL; (b) prediction results for backside weld width with STSL; (c) prediction results for penetration depth without STSL; (d) prediction results for penetration depth with STSL.

correspond to the molten region, which directly reflects fusion behavior. In the case of arc sound spectrograms, the attention is concentrated in high-energy time-frequency regions indicative of arc stability and penetration status. For infrared thermal images, the most salient zones are located near the arc boundary and molten pool region, where thermal gradients are strongest and most informative for quality estimation. These results visually support the claim that STSL facilitates more task-relevant attention compared to conventional activations.

In parallel, the corresponding feature maps shown in Fig. 11 demonstrate that STSL enhances the contrast and structural clarity of intermediate representations. Compared to ReLU, STSL suppresses background noise and emphasizes critical features across modalities. These include well-defined molten pool contours, distinct transitions in acoustic patterns, and temperature gradients in thermal images. These results visually confirm that the learned thresholds adaptively filter out modality-specific noise and guide the model to focus on task-relevant information. Together, these visualizations provide strong support for the effectiveness and interpretability of STSL in multimodal welding quality prediction.

4.2. Ablation study

To evaluate the effectiveness of each component within the proposed AM-TSFNet architecture, a comprehensive ablation study was conducted. The experiments were designed to examine the contributions of input modality combinations, the proposed feature fusion module, and the STSL to the overall regression performance. MSE and R^2 were used as evaluation metrics. Note that the experiments discussed in this section

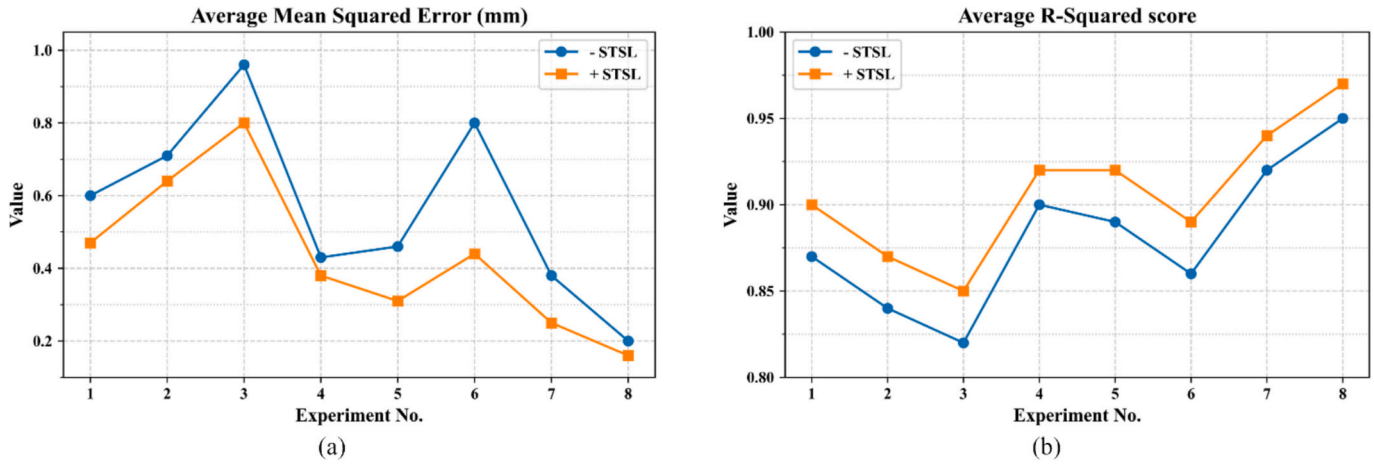

Fig. 13. Line chart of ablation study results (a) Average MSE of experiments without/with STSL module; (b) Average R^2 scores of experiments without/with STSL module.

refer to different model variants used in the ablation study, rather than physical welding experiments.

In this study, three distinct modalities were utilized: (1) molten pool images captured by a CCD camera, which provide critical visual cues such as pool shape and boundary definition; (2) arc sound signals, processed into time–frequency spectrograms using STFT, which capture dynamic acoustic patterns associated with arc stability and droplet detachment; and (3) infrared thermal images, which contain spatial heat distribution information related to thermal cycles and penetration behavior. Each modality contributes complementary information that enhances model perception and predictive accuracy.

As summarized in Table 7, the ablation experiments include three configurations with unimodal input (Experiments 1–3), three with bimodal input (Experiments 4–6), and two with multimodal input (Experiments 7–8). To further illustrate the prediction outcomes, Fig. 12 presents scatter plots comparing predicted and true values for weld width and penetration depth. Fig. 12(a) and (c) show results without the STSL module, while Fig. 12(b) and (d) include STSL for comparison. For clarity, the term "unimodal input" used in Experiments 1–3 refers to models trained using only one modality at a time (image, audio, or infrared), although all data were collected synchronously using a multimodal sensor system.

The results reveal a clear performance improvement as the input modality becomes more diverse. Transitioning from unimodal to bimodal and finally to multimodal input leads to consistent reductions in MSE and increases in R^2 across both regression tasks. This performance gain is attributed to the complementary nature of the heterogeneous input signals: arc sound and infrared data provide dynamic and thermal context, while molten pool images supply crucial spatial and structural features.

Additionally, across all experiments, configurations that include image data (e.g., Experiments 1, 4, 5, and 7–8) outperform those without, confirming the dominant role of visual features in weld formation analysis. A notable performance boost is also observed when comparing Experiments 7 and 8, demonstrating that the proposed attention-based feature fusion module significantly enhances the integration of multimodal information, enabling the network to leverage intermodal correlations and improving generalization.

Overall, the ablation results validate the importance of incorporating diverse sensing modalities—specifically, CCD molten pool images, arc sound spectrograms, and infrared thermal images—as well as the proposed feature extraction and fusion mechanism, in achieving robust and accurate predictions. These results are particularly significant under complex welding scenarios characterized not only by varying penetration states and fluctuations in welding parameters (such as current, pulse frequency, and wire feed speed), but also by real-world disturbances including background acoustic noise, image occlusion or reflection artifacts, and thermal signal inconsistencies caused by fluctuating emissivity or ambient temperature. The consistent improvements observed across all configurations emphasize the model's ability to generalize and remain resilient in noisy and dynamically changing production environments.

Table 8
Test metric scores of different methods on test data.

Model	Width		Depth		Average	
	MSE	R^2	MSE	R^2	MSE	R^2
AF-FTTSnet [49]	1.18	0.91	0.24	0.83	0.71	0.87
ViT [32] + Cross-attention [50]	1.12	0.92	0.32	0.76	0.72	0.84
ResNet18 [51] + Cross-attention [50]	0.83	0.93	0.23	0.83	0.53	0.88
MobileVit [36] + Cross-attention [50]	0.42	0.97	0.15	0.89	0.29	0.93
ViT [32] + FFM	0.75	0.94	0.19	0.86	0.47	0.90
ResNet18 [51] + FFM	0.59	0.96	0.17	0.88	0.38	0.92
MobileVit [36] + FFM	0.41	0.97	0.11	0.93	0.26	0.95
Ours	0.25	0.98	0.07	0.96	0.16	0.97

The impact of the STSL module on regression performance is clearly illustrated in Fig. 13. Across all experimental configurations, models equipped with STSL consistently achieve lower average MSE and higher R^2 scores compared to their counterparts without STSL. This improvement highlights the effectiveness of STSL in enhancing feature representation. As a soft-thresholding-based attention mechanism, STSL adaptively emphasizes informative spatial-temporal features while suppressing irrelevant or noisy signals. This selective enhancement facilitates more accurate prediction by improving the model's sensitivity to subtle structural variations in the input data. The consistent performance gains across different experiments confirm that STSL contributes significantly to both learning efficiency and model generalization in multimodal regression tasks.

4.3. Comparison with mainstream models

To comprehensively evaluate the effectiveness of the proposed model, we conducted comparisons with seven representative baseline methods: AF-FTTSnet, ViT + Cross-attention, ResNet18 + Cross-attention, MobileViT v3 + Cross-attention, ViT + FFM, ResNet18 + FFM, and MobileViT v3 + FFM. These models span a variety of backbone architectures and multimodal fusion strategies, enabling a fair and comprehensive performance comparison. The MSE and R^2 were adopted as

Table 9
Detailed welding parameters of 6 mm-thick LF6 aluminum plates.

Experiment	I_b (A)	I_p (A)	v_{wf} (cm/min)	Penetration states
No.1	170	250	100	LP
No.2	200	260	100	NP

evaluation metrics, and the results are summarized in Table 8.

The ResNet18 model is built on standard 2D convolution operations and lacks the ability to capture global dependencies. While ViT architectures are capable of modeling long-range interactions, their performance is limited by the high structural similarity across welding datasets, which makes it difficult to achieve high prediction accuracy. The MobileViT model integrates lightweight Transformer modules to enhance global awareness while preserving local features; however, it struggles to maintain regression stability under complex welding conditions. Although Cross-attention-based models enable interaction among multiple modalities, they lack an explicit mechanism to suppress redundant features, which may introduce irrelevant or noisy information and undermine fusion effectiveness. As a representative model integrating modality-aware fusion and attention mechanisms, AF-FTTSnet improves performance to a certain extent, but still suffers from limitations in inter-feature interaction and information selection.

In contrast, the proposed AM-TSFNet incorporates a soft-thresholding-based attention mechanism, which effectively preserves informative features while suppressing noise and redundancy. This mechanism facilitates more precise feature selection and representation, allowing the model to better capture spatial structural dynamics, time–frequency patterns, and infrared thermal variations. Furthermore, the multimodal fusion strategy designed in this work demonstrates strong coordination and alignment capabilities when dealing with heterogeneous data. As a result, the proposed AM-TSFNet achieves the highest accuracy and robustness in both width and depth regression tasks, further validating the effectiveness and generalizability of the method.

4.4. Generalization evaluation on varying material thickness

To further evaluate the generalization capability of the proposed AM-TSFNet, additional experiments were conducted using 6 mm-thick LF6 aluminum plates, extending beyond the original 4 mm dataset. The corresponding welding parameters are listed in Table 9, covering two representative penetration states: LP and NP. The weld samples and cross-section metallographic images are shown in Fig. 14. The data were collected using the same multimodal sensing setup, labeling procedure, and preprocessing methods described in Section 3.2, and were entirely excluded from the training and validation phases. The dataset comprises 1200 samples per modality.

The trained model was directly applied to this external test set to assess predictive performance. As summarized in Table 10, the model achieved an average MSE of 0.17 and R^2 of 0.97, demonstrating excellent generalization. For individual targets, the MSE for predicted width and depth were 0.23 and 0.11, respectively, with corresponding R^2 values of 0.99 and 0.95.

Regression results in Fig. 15 illustrate a strong alignment between predicted and ground-truth values across LP and NP cases. Although the material thickness increased from 4 mm to 6 mm, leading to changes in thermal behavior and fusion morphology, the predictions for both width and depth remained highly accurate and consistent. These findings confirm the robustness of AM-TSFNet in adapting to base metal thickness variations without requiring retraining or fine-tuning.

5. Results

The main results of this paper are as follows:

(1) The proposed AM-TSFNet can accurately predict backside weld width and penetration depth in P-GTAW. It achieved an average MSE of 0.16 mm and R^2 of 0.97 on unseen test data, demonstrating strong generalization ability and robustness.

Table 10
The prediction results of the proposed AM-TSFNet on 6 mm-thick LF6 aluminum plates.

	Metrics	Test results
Average	MSE	0.11
	R^2	0.97
Width	MSE	0.09
	R^2	0.99
Depth	MSE	0.12
	R^2	0.95

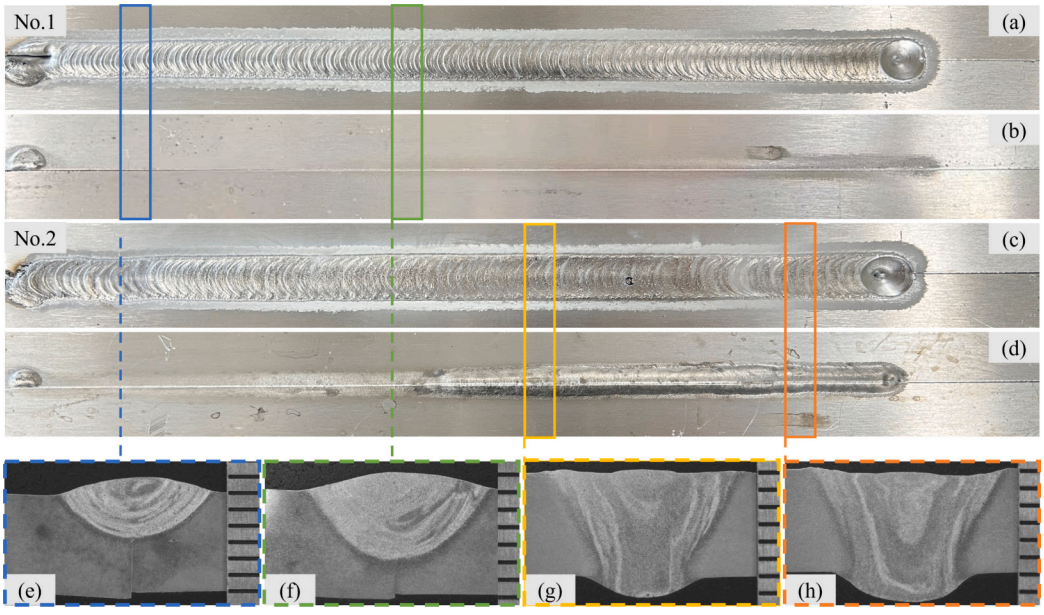

Fig. 14. Schematic diagram of the weld seams of 6 mm thick aluminum alloy plates.

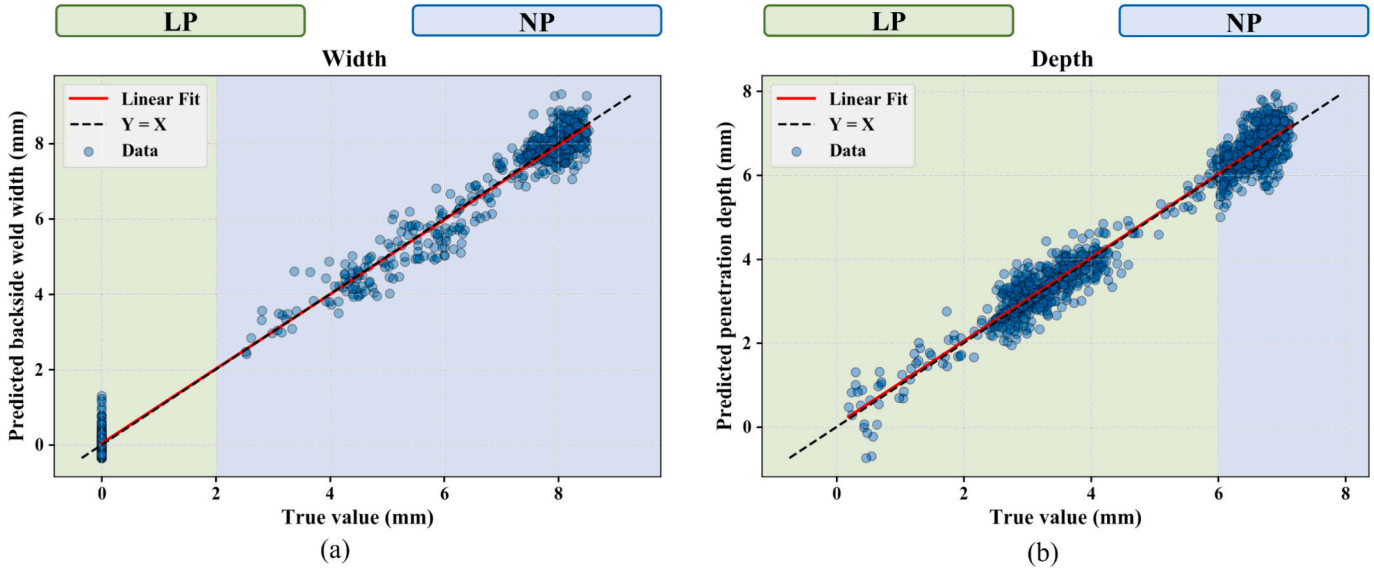

Fig. 15. Prediction results of 6 mm-thick LF6 aluminum plates. (a) Predicted backside weld width; (b) Predicted penetration depth.

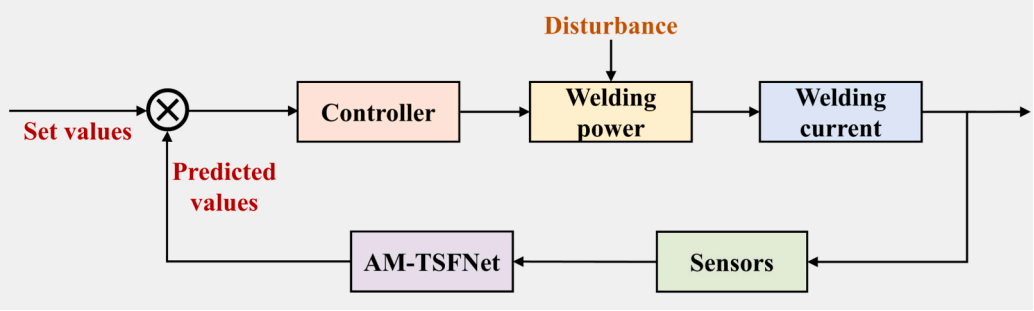

Fig. 16. Conceptual schematic of a closed-loop welding control system incorporating AM-TSFNet predictions.

(2) Ablation study confirmed that the combination of multimodal inputs, the Transformer-serial fusion architecture, and the cross-modal fusion strategy jointly contributed to steady improvements in regression performance.
(3) Compared with other mainstream models, AM-TSFNet significantly reduced prediction error (MSE) and improved fitting accuracy (R^2), highlighting the superiority of its attention-guided and noise-suppressing design in complex welding environments.
(4) The proposed weld quality monitoring framework demonstrated high accuracy and efficiency for both weld width and penetration prediction tasks, making it suitable for weld quality assessment and providing a basis for the real-time adjustment of P-GTAW process parameters.

Compared to our earlier work [3], which concentrated on multimodal classification of weld states, the proposed AM-TSFNet represents a significant advancement by enabling quantitative regression of weld geometry. This extension broadens the applicability of multimodal fusion techniques for intelligent welding quality monitoring.

In current practical settings, the predicted values of backside weld width and penetration depth can be displayed in real time to assist engineers in monitoring weld quality. When deviations from target values occur, engineers can manually adjust process parameters such as welding current or speed to maintain product consistency.

In addition, this work lays the foundation for integrating real-time weld quality prediction into closed-loop control systems. By embedding the proposed AM-TSFNet as a real-time quality estimator, predicted weld width and penetration depth can serve as direct feedback signals. These can be compared with preset quality targets to drive automatic parameter adjustments (e.g., arc current or travel speed), as illustrated in the conceptual control framework shown in Fig. 16. Such predictive-feedback integration is expected to enhance process stability and reduce quality deviations under dynamic welding conditions. We consider this a promising direction for future research and plan to explore it further by developing a fully adaptive, closed-loop welding system based on the AM-TSFNet framework.

6. Conclusions

We propose an online weld quality monitoring method based on multimodal sensing, including visible molten pool images, arc sound signals, and infrared thermal images, applied to robotic P-GTAW. AM-TSFNet effectively predicts backside weld width and penetration depth by leveraging a Transformer-serial fusion architecture, the STSL for noise suppression, and an attention-guided cross-modal fusion module.

The approach exhibits excellent accuracy, robustness, and generalization capability, making it a promising solution for real-time weld quality monitoring and intelligent process control. In future work, we aim to extend this framework to other welding techniques such as Gas Metal Arc Welding and Plasma Arc Welding, as well as arc-based additive manufacturing scenarios, and to explore the development of a fully adaptive, closed-loop welding system based on the AM-TSFNet framework.

CRediT authorship contribution statement

Yuqing Xu: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Qiang Liu:** Investigation. **Jingyuan Xu:** Investigation. **Shanben Chen:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition.

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 is partly supported by the National Natural Science Foundation of China under the Grant No. 51575349 and 61873164.

References

[1] Chen SB, Lv N. Research evolution on intelligentized technologies for arc welding process. J Manuf Processes 2014;16:109–22. https://doi.org/10.1016/j.jmapro.2013.07.002.
[2] Xu Y, Liu Q, Xu J, Xiao R, Chen S. Review on multi-information acquisition, defect prediction and quality control of aluminum alloy GTAW process. J Manuf Processes 2023;108:624–38. https://doi.org/10.1016/j.jmapro.2023.11.025.
[3] Xu Y, Liu Q, Xu J, Xiao R, Chen S. A multi-spectral channel attention mechanism for prediction of welding state during pulsed GTAW. J Manuf Processes 2025;134:1021–33. https://doi.org/10.1016/j.jmapro.2025.01.023.
[4] Bahedh AS, Mishra A, Al-Sabur R, Jassim AK. Machine learning algorithms for prediction of penetration depth and geometrical analysis of weld in friction stir spot welding process. Metall Res Technol 2022;119:305. https://doi.org/10.1051/metal/2022032.
[5] Zhao Z, Lv N, Xiao R, Chen S. A novel penetration state recognition method based on LSTM with auditory attention during pulsed GTAW. IEEE Trans Ind Inform 2022:1–11. https://doi.org/10.1109/TII.2022.3229837.
[6] Zhao Z, Lv N, Xiao R, Liu Q, Chen S. Recognition of penetration states based on arc sound of interest using VGG-SE network during pulsed GTAW process. J Manuf Process 2023;87:81–96. https://doi.org/10.1016/j.jmapro.2022.12.034.
[7] Chandrasekhar N, Muthukumaran V, Das CR. Real-time determination of weld penetration status during A-TIG welding of stainless steel employing deep learning approach. Weld World 2025. https://doi.org/10.1007/s40194-025-02021-6.
[8] Zhang B, Shi Y, Cui Y, Wang Z, Hong X. Prediction of keyhole TIG weld penetration based on high-dynamic range imaging. J Manuf Process 2021;63:179–90. https://doi.org/10.1016/j.jmapro.2020.03.053.
[9] Xue B, Chang B, Du D. Monitoring of high-speed laser welding process based on vapor plume. Opt Laser Technol 2022;147:107649. https://doi.org/10.1016/j.optlastec.2021.107649.
[10] Xu J, Liu Q, Xiao R, Xu Y, Zhou W, Chen S. A real-time prediction method for weld porosity of 5A06 aluminum alloy based on arc spectra using elemental attention, multi-scale convolution and LSTM during pulsed GTAW process. Opt Laser Technol 2025;186:112708. https://doi.org/10.1016/j.optlastec.2025.112708.
[11] Li S, Jiang P, Gao Y, Song M, Shu L. A penetration depth monitoring method for Al-cu laser lap welding based on spectral signals. J Mater Process Technol 2023;317:117972. https://doi.org/10.1016/j.jmatprotec.2023.117972.
[12] Wang Y, Lee W, Jang S, Truong VD, Jeong Y, Won C, et al. Prediction of internal welding penetration based on IR thermal image supported by machine vision and ANN-model during automatic robot welding process. J Adv Join Process 2024;9:100199. https://doi.org/10.1016/j.jajp.2024.100199.
[13] Jiang R, Xiao R, Chen S. Prediction of penetration based on infrared thermal and visual images during pulsed GTAW process. J Manuf Process 2021;69:261–72. https://doi.org/10.1016/j.jmapro.2021.07.046.
[14] Xiao X, Liu X, Cheng M, Song L. Towards monitoring laser welding process via a coaxial pyrometer. J Mater Process Technol 2020;277:116409. https://doi.org/10.1016/j.jmatprotec.2019.116409.
[15] Zhang Y, Yan W. Applications of machine learning in metal powder-bed fusion in-process monitoring and control: status and challenges. J Intell Manuf 2023;34:2557–80. https://doi.org/10.1007/s10845-022-01972-7.
[16] Chen L, Bi G, Yao X, Tan C, Su J, Ng NPH, et al. Multisensor fusion-based digital twin for localized quality prediction in robotic laser-directed energy deposition. Robot Comput Integr Manuf 2023;84:102581. https://doi.org/10.1016/j.rcim.2023.102581.
[17] Wang H, Zhu H, Li H. A rotating machinery fault diagnosis method based on multi-sensor fusion and ECA-CNN. IEEE Access 2023;11:106443–55. https://doi.org/10.1109/ACCESS.2023.3320065.
[18] Mao G, Li H, Xue L, Li Y, Cai Z, Noman K. FedPM-SGN: a federated graph network for aviation equipment fault diagnosis by multi-sensor fusion in decentralized and heterogeneous setting. Inf Fusion 2025;117:102876. https://doi.org/10.1016/j.inffus.2024.102876.
[19] 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 Processes 2021;68:209–24. https://doi.org/10.1016/j.jmapro.2020.08.028.
[20] Hong Y, Jiang Y, Yang M, Chang B, Du D. Intelligent seam tracking in foils joining based on spatial–temporal deep learning from molten pool serial images. Rob Comput Integr Manuf 2025;91:102840. https://doi.org/10.1016/j.rcim.2024.102840.
[21] Zhang Z, Chen H, Xu Y, Zhong J, Lv N, Chen S. Multisensor-based real-time quality monitoring by means of feature extraction, selection and modeling for Al alloy in arc welding. Mech Syst Sig Process 2015;60–61:151–65. https://doi.org/10.1016/j.ymssp.2014.12.021.
[22] Zhu K, Wang Q, Chen W, Li X, Xiao R, Chen H. Robotic MAG welding defects and quality assessment with a defect threshold decision model-driven method. Mech Syst Sig Process 2025;224:112056. https://doi.org/10.1016/j.ymssp.2024.112056.
[23] Gao X, Li Z, Wang L, Zhou X, You D, Gao PP. Detection of weld imperfection in high-power disk laser welding based on association analysis of multi-sensing features. Opt Laser Technol 2019;115:306–15. https://doi.org/10.1016/j.optlastec.2019.01.053.
[24] Zhang Z, Chen S. Real-time seam penetration identification in arc welding based on fusion of sound, voltage and spectrum signals. J Intell Manuf 2017;28:207–18. https://doi.org/10.1007/s10845-014-0971-y.
[25] Gao P, Wu Z, Wang Y, Lu J, Zhao Z. Method for monitoring and controlling penetration of complex groove welding based on online multi-modal data. J Intell Manuf 2024;35:1247–65. https://doi.org/10.1007/s10845-023-02107-2.
[26] Gao P, Su X, Wu Z, Lu J, Han J, Bai L, et al. Online penetration prediction based on multimodal continuous signals fusion of CMT for full penetration. J Manuf Process 2024;115:431–40. https://doi.org/10.1016/j.jmapro.2024.02.033.
[27] Zhang Y, Wang Q, Liu Y. Adaptive intelligent welding manufacturing. Weld J 2021;100:63–83. https://doi.org/10.29391/2021.100.006.
[28] Liu Y, Zhang Y. Control of 3D weld pool surface. Control Eng Pract 2013;21:1469–80. https://doi.org/10.1016/j.conengprac.2013.06.019.
[29] Liu YK, Zhang YM. Model-based predictive control of weld penetration in gas tungsten arc welding. IEEE Trans Control Syst Technol 2014;22:955–66. https://doi.org/10.1109/TCST.2013.2266662.
[30] Cayo EH, Alfaro SCA. A non-intrusive GMA welding process quality monitoring system using acoustic sensing. Sens 2009;9:7150–66. https://doi.org/10.3390/s90907150.
[31] Alvarez Bestard G, Absi Alfaro SC. Measurement and estimation of the weld bead geometry in arc welding processes: the last 50 years of development. J Braz Soc Mech Sci Eng 2018;40:444. https://doi.org/10.1007/s40430-018-1359-2.
[32] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2021. https://doi.org/10.48550/arXiv.2010.11929.
[33] Lai-Dang Q-V. A Survey of Vision Transformers in Autonomous Driving: Current Trends and Future Directions. 2024. https://doi.org/10.48550/arXiv.2403.07542.
[34] Wang L, Kang X, Ding F, Nakagawa S, Ren F. MSSTNet: A Multi-Scale Spatio-Temporal CNN-Transformer Network for Dynamic Facial Expression Recognition. ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2024. p. 3015–9. https://doi.org/10.1109/ICASSP48485.2024.10446699.
[35] Park S, Yeo Y-J, Shin Y-G. PConv: simple yet effective convolutional layer for generative adversarial network. Neural Comput Applic 2022;34:7113–24. https://doi.org/10.1007/s00521-021-06846-2.
[36] Wadekar SN, Chaurasia A. MobileViTv3: Mobile-Friendly Vision Transformer with Simple and Effective Fusion of Local, Global and Input Features. 2022. https://doi.org/10.48550/arXiv.2209.15159.
[37] Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998;86:2278–324. https://doi.org/10.1109/5.726791.
[38] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;60:84–90. https://doi.org/10.1145/3065386.
[39] Zhao M, Zhong S, Fu X, Tang B, Pecht M. Deep residual shrinkage networks for fault diagnosis. IEEE Trans Industr Inform 2020;16:4681–90. https://doi.org/10.1109/TII.2019.2943898.
[40] Zhao Z, Lv N, Xiao R, Chen S. A novel penetration state recognition method based on LSTM with auditory attention during pulsed GTAW. IEEE Trans Ind Inform 2023;19:9565–75. https://doi.org/10.1109/TII.2022.3229837.
[41] Zhang Z, Wen G, Chen S. Weld image deep learning-based on-line defects detection using convolutional neural networks for Al alloy in robotic arc welding. J Mater Process Technol 2012;212:1654–62. https://doi.org/10.1016/j.jmapro.2019.06.023.
[42] Xu Y, Yu H, Zhong J, Lin T, Chen S. Real-time seam tracking control technology during welding robot GTAW process based on passive vision sensor. J Mater Process Technol 2012;212:1654–62. https://doi.org/10.1016/j.jmatprotec.2012.03.007.
[43] Chen C, Lv N, Chen S. Welding penetration monitoring for pulsed GTAW using visual sensor based on AAM and random forests. J Manuf Processes 2021;63:152–62. https://doi.org/10.1016/j.jmapro.2020.04.005.
[44] Oppenheim A, Schafer R. Discrete-Time Signal Processing. Upper Saddle River: Pearson; 2010.
[45] Allen J. Short term spectral analysis, synthesis, and modification by discrete Fourier transform. IEEE Trans Acoust Speech Signal Process 1977;25:235–8. https://doi.org/10.1109/TASSP.1977.1162950.

[46] Pattern Recognition and Machine Learning. n.d.
[47] Applied Regression Analysis. n.d.
[48] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. 2019. https://doi.org/10.48550/arXiv.1610.02391.
[49] Hong Y, He X, Xu J, Yuan R, Lin K, Chang B, et al. AF-FTTSnet: an end-to-end two-stream convolutional neural network for online quality monitoring of robotic welding. J Manuf Syst 2024;74:422–34. https://doi.org/10.1016/j.jmsy.2024.04.006.
[50] Chen C-F, Fan Q, Panda R. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. 2021. https://doi.org/10.48550/arXiv.2103.14899.
[51] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2015. https://doi.org/10.48550/arXiv.1512.03385.