Journal of Manufacturing Processes 115 (2024) 431–440

Contents lists available at ScienceDirect

Journal of Manufacturing Processes

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

Online penetration prediction based on multimodal continuous signals fusion of CMT for full penetration

Peng Gao, Xiaocong Su, Zijian Wu, Jun Lu *, Jing Han, Lianfa Bai, Zhuang Zhao *

Jiangsu Key Laboratory of Spectral Imaging and Intelligent Sense, Nanjing University of Science and Technology, Nanjing 210094, China

ARTICLE INFO

Keywords:
Audio-visual signal
Penetration state
Deep learning

ABSTRACT

Online penetration monitoring for complex butt welding is challenging due to steel plate's groove instability and welding heat deformation. In this paper, automatic cold metal transfer (CMT) welding is used to join two complex bevelled austenitic stainless steel with SS304 as the base metal. This work reports a hybrid approach combining deep learning, computer vision, and sound signal processing to monitor groove welding penetration under full penetration in real time. Sequence signals such as video and sound can complimentarily characterize the melt pool state. In this paper, the proposed Multimodal continuous signals Characteristic Reinforcement Network (MCRNet) utilizes 3D convolution and multiscale convolution with channel attention to considerably improve the performance of lightweight networks. At the same time, a new fusion method with similarity loss is proposed to cope with the input of visual and acoustic signals. That improves the effect by at least 18 % compared with the single-modal signal input. The experimental results show that the Mean Square Error (MSE) of MCRNet improved the performance by 44 % compared with the mainstream deep learning framework. Meanwhile, the inference speed under multimodal input reaches 57 frames per second (FPS). MCRNet finally achieves online penetration accurate prediction of the melt pool.

1. Introduction

Welding technology is the most common way of joining metals. Traditionally, welding quality inspection is usually carried out after welding. Especially for butt welds, the penetration quality dramatically affects the joined metal's rigidity at both ends [1]. In industrial production, the penetration quality of butt welds is precarious due to unstable manual welding and the quality of steel plate processing [2]. The penetration quality is invisible from the front side of the steel plate during or even after welding. The plate needs to be analyzed from the backside or cut open to estimate the penetration quality [3]. With the development of intelligent welding, manual welding has gradually been replaced by robotic welding. While robotic welding offers more stable and consistent results, it does come with particular challenges. Namely, welding must be performed using predetermined parameters and a fixed path and cannot be adjusted in real time based on human experience and intuition [4]. Additionally, any defects in the welding process can lead to costly and sometimes irreparable damage, further exacerbating the challenges associated with robotic welding [5].

Several methods have been proposed to detect errors that may occur during the welding process automatically. These sensors include acoustic [6,7], current [8,9], temperature [10,11], and spectral sensors [12], which have been used to monitor basic welding conditions. In one of these examples, neural networks analyzed signals from acoustic and current sensors to classify the penetration state [9]. These efforts indeed cope with simple tasks in the welding process, but most only extract shallow features. The deeper features of the data are not exploited, and there is no confidence in handling complex tasks.

Vision-based sensing technologies can provide richer information. This provides the potential to identify different welding states in complex processes or workpiece conditions, providing timely correction solutions. Contemporary work exists using both traditional machine vision [13–15] and deep learning [16–18] approaches to characterize the welding states. Deep learning approaches have better robustness and generalization in classification tasks of penetration state than traditional machine learning approaches, especially in complex environments. In particular, Convolutional Neural Network (CNN) such as ResNet and DenseNet offer opportunities for melt pool online monitoring tasks, as convolution is highly efficient for image feature extraction, especially sensitive to image pixel position [19,20]. Thus, they are trained to

* Corresponding authors.
E-mail addresses: lujunchenhao@njust.edu.cn (J. Lu), zhaozhuang@njust.edu.cn (Z. Zhao).

https://doi.org/10.1016/j.jmapro.2024.02.033
Received 28 September 2023; Received in revised form 21 January 2024; Accepted 16 February 2024
Available online 20 February 2024

classify or segment melt pool images [21,22]. Currently, most vision tasks for butt welding focus on the variation of steel plate shape [23,24]. Little research has been done on varying groove angles in butt welding, but the change of penetration state caused by the angle in groove welding is more significant than that caused by steel plate shape.

Meanwhile, the above works extract features from one single melt pool image. It should be noted that welding is a continuous physical process, and it takes time for heat diffusion and solidification of the melt pool [25]. In an actual welding experiment, it will be found that the end of the weld still shows the red colour caused by high temperature when the welding is completed. The shape of the surface is still unformed, not to mention the hotter interior. The instantaneous image is insufficient to reflect the current penetration state [26,27]. Sequence signals such as video and sound can complimentarily characterize the melt pool state. Therefore, it is necessary to determine the penetration state from successive frames of melt pool images. How to handle multi-frame information extraction is the first challenge.

With complex grooves, the forces on the melt pool and droplets change as the welding environment changes, greatly affecting the penetration depth. Furthermore, there is access limitation by the groove, and data cannot be obtained directly from the backside, which increases the difficulty of sensing the welding process [28]. Raw information containing semantic information is needed to characterize and predict penetration indirectly. Additional sound information is introduced as another modal information supplement. Information from different modalities can represent the same melt pool state from different directions and can be used as complementary information. Based on the limited current work on audio-visual based tasks, a notable representative work is a method proposed by Wu et al., which uses hybrid deep learning to integrate image and sound for penetration classification [29]. This work is pioneering, but not applicable to subsequent online control due to issues related to data acquisition speed and shortcomings in the classification task. Many studies have already begun to explore online control with the use of visual modalities [30–32]. Online control represents a future trend, and a fundamental aspect of online control is obtaining accurate weld seam regression values, as opposed to mere classification. An advanced neural network is needed to fuse the video and sound information of the melt pool. For the butt welding used in this paper, the backside melt width after cooling was used as a quantitative representation of the penetration state. Simultaneously, to prevent the occurrence of under and over penetration defects, our research focuses on the precise regression of the backside melt width under the condition of full penetration. By doing so can unfavorable trends be identified before defects occur.

This work reports a fast and scalable CNN that fuses sequence data from video and sound modalities for regression of penetration of the melt pool. The method has been applied to variable groove welding penetration detection with outstanding results. The paper is organized as follows: Section 2 analyses the penetration state while demonstrating why video and audio information is needed; Section 3 gives an overview of the experimental setup and data generation; Section 4 details the MCRNet framework. Section 5 describes the results and discussion of the validation and ablation experiments. Finally, conclusions are drawn in Section 6.

2. Penetration analysis

CMT welding is a highly effective process where the melting and transfer of the welding wire take place in a precise, step-by-step manner. That significantly reduces spatter and oxidation compared to traditional Gas Metal Arc Welding (GMAW) welding while minimizing part distortion and annealing for superior welding speed and quality. One of the unique benefits of CMT is its special process that allows for a base and peak moment during welding, which enables greater control and precision throughout the entire process. During the base moment, the arc is deliberately stopped, allowing for a crystal-clear image of the melt pool to be captured using passive vision.

Heat input can somewhat determine the degree of penetration. However, the same heat input can lead to different penetration states for different groove angles, as shown in Fig. 1. Weld penetration is a crucial indicator of the mechanical robustness of a weld, making it an essential parameter for evaluating the quality of this method. Depending on the extent of this process, the weld can be classified into three categories: under-penetration, full-penetration, and over-penetration. Observing the backside penetration state is difficult in actual production, but skilled workers can directly determine the penetration state with their expert eyes and expertise. In the CMT classification experiments, raw visual and audio information could easily distinguish the difference between different penetration states.

Through the comparison of the three states, some intriguing findings have emerged by analyzing images of the melt pool formation.

(1) For a groove angle of 60°, the limited heat input is insufficient to penetrate the groove fully, and the melt pool will overflow onto the surface of the groove. Due to the surface tension, the melt pool develops into a protruding convex shape. Furthermore, the width of the pool is broader while the length appears comparatively shorter, resulting in under-penetration.

(2) In contrast, for a groove angle of 120°, the surface of the melt pool tends to be lower than the surface of the groove and forms a soothing concave shape. Accordingly, the width of the melt pool is narrower while the length appears longer, leading to over-penetration.

(3) Based on the observations and classification characteristics of the melt pool images, there is an indirect relationship between the morphology of the melt pool under different groove angles and its corresponding penetration state. Fig. 1 shows that the backside melt width is 0 mm for under-penetration and the backside melt width limit at 6 mm for over-penetration. When the backside melt width is 0–6 mm, molten droplets seep from the back side, forming a stable weld internally [33].

Meanwhile, the acoustic features during welding under complex groove conditions are one of the important melt pool characterizations. In practical experiments, the human ear discerns different welding states distinctly. When there is under-penetration, the arc sound is harsh with crackling noises. When there is full-penetration, the sound is relatively soft and stable. When there is over-penetration, the sound is lower with splattering noises. Based on the Nyquist sampling theorem, to accurately recover the original signal from a sampled signal without distortion, the sampling frequency should be greater than twice the highest frequency of the signal. The audible frequency range of the human ear is typically considered to be 20–20,000 Hz. Therefore, this study sets the sampling frequency of the audio signal to 51,200 Hz. The complete weld is transformed into a spectrogram through operations such as Fourier Transform. As the groove size increases, the welding state transitions from incomplete penetration to overpenetration. This transition is reflected in the spectrogram by a gradual decrease in the intensity of low-frequency sound signals, which is apparent from Fig. 2.

In the analysis above, preliminary differentiation of different penetration states could be achieved through human visual and auditory monitoring. However, this analysis remains limited to qualitative

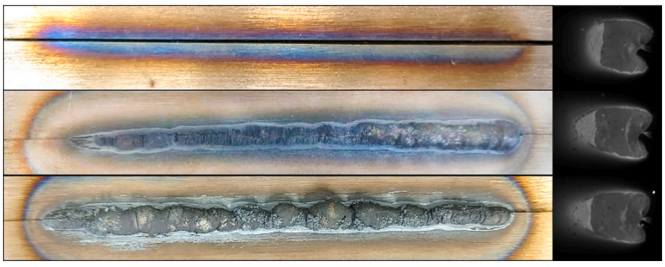

Fig. 1. Backside view of weld with melt pool image classification experiments: under-penetration, full-penetration, and over-penetration.

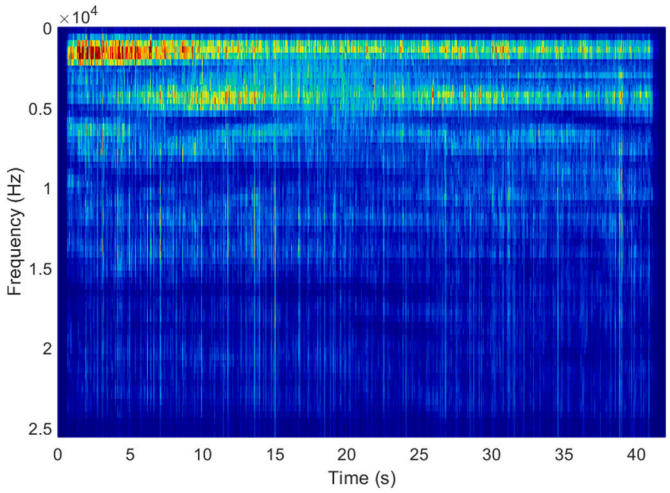

Fig. 2. Acoustic spectrum of welding with gradual groove angle.

analysis, and quantitative analysis still requires the assistance of algorithms. Despite the regular classification tasks expertly identifying the boundaries of the three states, it falls short in addressing the intricacies of intelligent welding as defects have already taken place. It's necessary to anticipate potential defects and make the necessary adjustments to the parameters. With a broad range of backside melt widths at full-penetration, this allows the other two states to be predicted in advance.

3. Experimental setup and data generation

This system is designed with a vision module and sound module for a complementary approach, which is shown in Fig. 3. The vision module consists of an industrial camera and a filter. The sound module consists of a microphone and a data acquisition card. The two modules are positioned symmetrically relative to the torch to guarantee non-interference. The camera and microphone are mounted on the robot arm and move synchronously with the welding torch. Specifically, the camera is Basler 1920-155um with an 850 high-pass filter, the microphone is MPA201, and the data acquisition card is ADLink USB2405. To reduce the bending of the steel plate in the experiments due to its small size compared to the actual industrial production, custom-made clamps were made to hold the steel plates in place stably.

A frame rate of 70 Hz was chosen for capturing the melt pool image with a size of 500 × 700, which is approximately the same period as the CMT. To mitigate the jittering issue caused by the follow-up equipment during welding, a random crop of 0–50 pixels was applied horizontally and vertically to the original image. Additionally, pixels were added in the opposite direction to maintain the integrity of the melt pool area. A two-step process was implemented to address the challenges posed by variable light intensities in complex welding environments. First, a sharpening operation to the image was applied, followed by a random brightness enhancement ranging from −10 % to 10 %. The hue saturation of the image was also randomly adjusted using the same method. Fig. 4 illustrates the preprocessing steps applied to the melt pool image.

The acquisition frequency is set to 51,200 Hz for acquiring sound signals. The frequency is set so high to retain as much information as possible in both high and low dimensions while ensuring that the subsequent Short Time Fourier Transform (STFT) requires twice the amount of data. Frequency analysis of sound is commonly used in various tasks. Frequency domain signals contain rich feature information, while sound spectrograms can express richer time-varying characteristics of frequency domain signals. Fourier transform can convert complex time and space signals to the frequency domain. However, its frequency domain characteristics are time-varying for some unstable signals, such as welding sound signals. At this point, the STFT is more suitable for obtaining the main frequency characteristics of the sound on the partial time period [34]. The STFT treats the non-smooth process as a superposition of a series of short-time smooth signals, and the shortness can be achieved by adding windows to the time. The equation for the STFT is as follows.

$$STFT(f,k) = \sum_{n=0}^{N-1} s(n)\left[W(n-k)e^{-j2\pi fn/N}\right] \quad (1)$$

For the discrete signal s, the window function W is used to do the Fourier transform by sliding along the direction of $s(n)$. For the sound signal acquired by our system, the window length is set to 128, the step size is set to 32 and STFT is used to generate several 64*64 sound spectrograms according to the welding time. The preprocessing process of the sound signal is shown in Fig. 5.

The Wiiboox Reeyee 5 M 3D scanner can accurately capture the backside melt width. Once the backside 3D point cloud data is obtained, the weld edges are carefully marked by hand. To determine the melt width, the distance between the identified boundaries is calculated based on the direction of the weld. This data is then used to fit a distance-melt width curve.

4. Network framework

This section describes the essential components of MCRNet. The backbone is the soul of a network, and its design determines the model's performance, number of participants, and inference speed. Due to the abrupt increase in the amount of data, the design of neural networks should achieve a balance between depth and efficiency. Based on our preliminary research on the mainstream backbone networks, the current networks with better results, such as ResNet, DenseNet, Transformer, ConvNeXt, etc., all have a certain depth in the structure of the network [19,20,35,36]. That is because each layer of the neural network corresponds to extracting different levels of a feature. The deeper the network

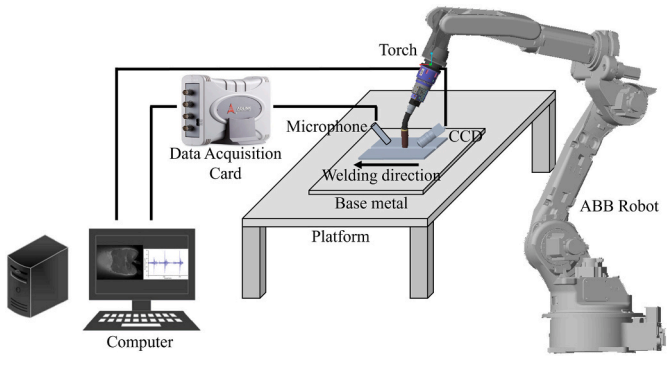

Fig. 3. Overview of the welding process and equipment.

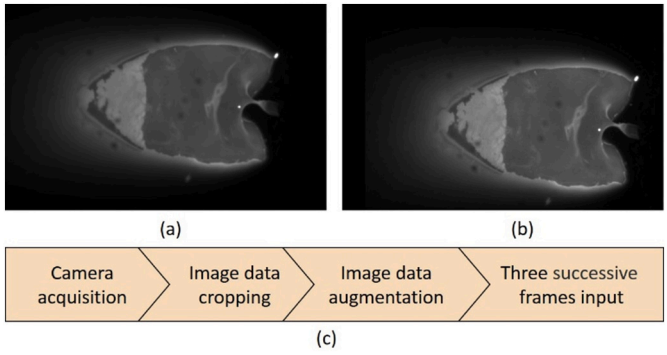

Fig. 4. Preprocessing of the melt pool image. (a) Original melt pool image. (b) Augmented image. (c) Steps of preprocessing.

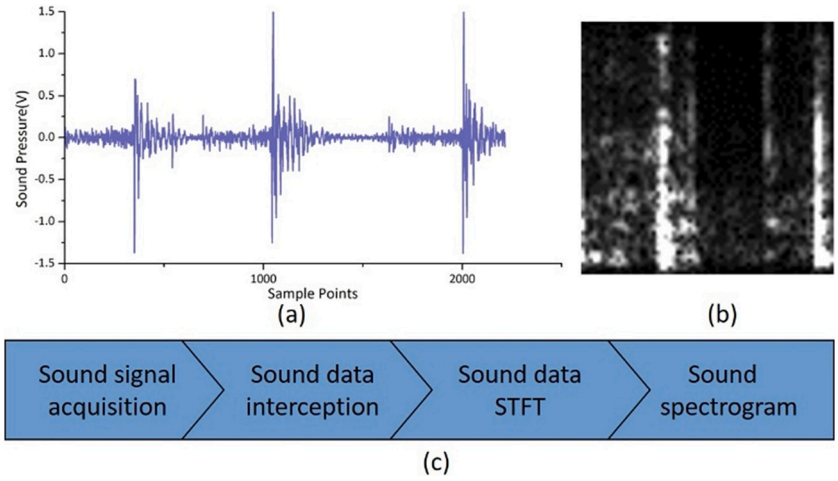

Fig. 5. Preprocessing of the sound signal. (a) Original sound signal. (b) Sound spectrogram. (c) Steps of preprocessing.

is, the more information can be extracted from different levels. However, a deeper network results in a longer inference time, which is unfriendly for online monitoring. Therefore, there is a need to balance time and accuracy with limited resources, giving a higher priority to accuracy.

Regarding this, MCRNet ensures efficient melt pool feature extraction within a limited depth. It combines 3D convolution, multiscale 2D convolution, and channel attention to reinforce the extraction and screening of continuous signal features. The fusion module combines features from different modalities. The idea of our network is to enhance the extraction of features with continuous correlated signals, such as video sequences and sound sequences. The structure of MCRNet is shown in Fig. 6. In the network framework, the input variables include video sequences and spectrograms, which are used to characterize the melt pool state. The video sequence input size is $3 \times 256 \times 256$, and the spectrogram input size is $1 \times 128 \times 128$. The output variable is the backside weld width, representing the penetration state.

4.1. 3Dcov block

For multi-frame images, three consecutive frames are integrated into separate channels, resulting in a network input of dimensions $3 \times 256 \times 256$. The adoption of 3D convolution is motivated by their capacity to capture spatiotemporal information from consecutive video frames. The preference for three-dimensional convolutions over their two-dimensional counterparts is rooted in their ability to seamlessly combine spatial and temporal data, thereby enhancing the model's understanding of temporal sequences.

An additional benefit of employing three-dimensional convolutions lies in the maintenance of feature consistency. In multi-frame images, it is common for adjacent frames to share similar features. The use of three-dimensional convolutions leverages this feature consistency, reducing model complexity, improving computational efficiency, and mitigating overfitting risks.

Consequently, this convolutional approach contributes to enhancing the model's performance, enabling it to better grasp the temporal relationships and dynamic variations within the data. To facilitate the direct extraction of multi-frame image information, 3D convolution is introduced at the initial stage of video data processing.

4.2. MFS module

The Multi-Feature Screening (MFS) Module consists of a Multi-Feature Extraction (MFE) block and a Squeeze-and-Excitation (SE) block. The MFE block occupies a central and indispensable role within our network's framework. It encompasses five distinct branches, each dedicated to the extraction of diverse features. These branch outputs are subsequently harmonized to culminate in the ultimate network output. To enrich our model's capacity for capturing intricate spatial details along both the horizontal and vertical axes, asymmetric convolutions of dimensions 1×3 and 3×1 are introduced. The inclusion of these convolutions amplifies the diversity of extracted features. Moreover, 1

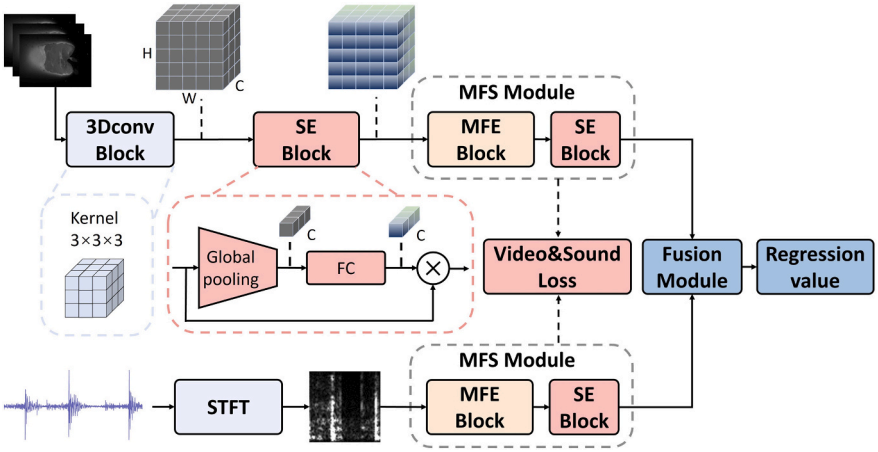

Fig. 6. MCRNet framework.

× 1 convolutions have been thoughtfully integrated to bolster the network's non-linear processing capabilities.

To fortify the network against overfitting, residual connections and a proven regularization strategy are needed. In the absence of considering the activation function and batch normalization layer, the relationship between the network's input and output can be succinctly expressed as follows:

$$y_i = F_{3\times3}(x_i, \{W_i\}) + F_{1\times3}(x_i, \{W_i\}) + F_{3\times1}(x_i, \{W_i\}) + F_{1\times1}(x_i, \{W_i\}) + x_i \tag{2}$$

$F(x)$ encapsulates the network's learned transformations, where $F(x_i, \{W_i\})$ denotes the different convolutions on x_i with the corresponding weights W_i. This architectural configuration ensures a more robust and generalizable network performance while preserving the integrity of the input information. According to reparameterization [37], these multiscale convolution parameters are superimposed in the inference stage after transforming them into 3 × 3.

SE block consists mainly of a global pooling layer and a Fully Connected (FC) layer. It adds channel attention mechanism on the basis of the above feature extraction and adds a weight to each feature channel to realize the screening of core features.

4.3. Fusion module

Our approach is centred on transforming multimodal data into a feature map of matching dimensions by utilizing linear and Batch Normalization (BN) layers. This transformation leverages the inherent characteristics of the image data as it traverses the backbone network. Rather than pursuing a straightforward concatenation of the data for fusion, our approach embraces a more intricate and thoughtful strategy. In our method, the 1D vectors corresponding to the two distinct features are converted into 2D representations through the application of vector multiplication. This 2D transformation facilitates a more intimate and cohesive interplay between the features. Intriguingly, our experimentation reveals that employing a shallow network to process these amalgamated features proves more productive than direct fusion. Furthermore, other research indicates that employing two successive linear layers yields results on par with a self-attention module [38], emphasizing the proficiency of our chosen strategy.

Recognizing the common objective shared by video and sound features, both characterizing the identical penetration state, a loss function is designed to encourage similarity between these two modalities. This constraint incentivizes the alignment of expressions between the two sets of features, fostering a more harmonious representation of the underlying data. The structural layout of our fusion module is visually depicted in Fig. 7, illustrating the architecture's design and functionality.

4.4. Similarity loss

In our quest to foster a meaningful association among the video, sound, and penetration data modalities, a similarity loss function is introduced. Despite the inherent disparities between video and sound data, they possess the potential to complement each other, particularly given their shared capacity to express the same penetration state. Thus, it becomes paramount that the features derived from the video and sound within the same group exhibit the utmost similarity. By optimizing the resemblance between these two modalities, our overarching aim is to enhance the accuracy of target value predictions and, concomitantly, elevate the overall system's performance. To achieve this, our model is augmented with the loss term denoted as $L_{V\&S}$, which evaluates the cosine similarity between the two feature sets.

The cosine similarity is computed as follows:

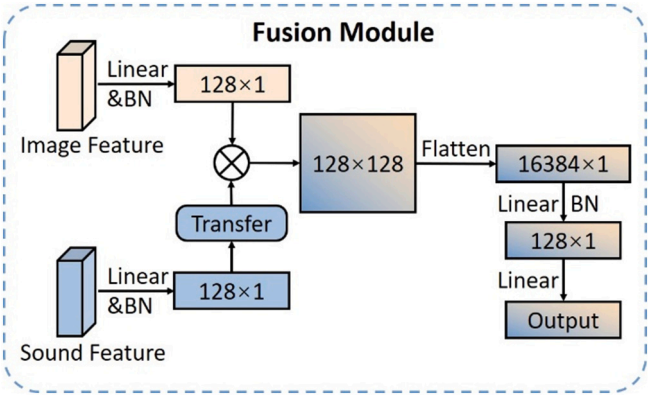

Fig. 7. Fusion module structure.

$$similarity = cos(\theta) = \frac{V \cdot S}{\|V\|\|S\|} = \frac{\sum_{i=1}^{n} V_i \times S_i}{\sqrt{\sum_{i=1}^{n} (V_i)^2} \times \sqrt{\sum_{i=1}^{n} (S_i)^2}} \tag{3}$$

Here, V_i is the feature map obtained from the melt pool video, and S_i is the feature map obtained from the sound spectrograms. The loss term $L_{V\&S}$ is defined as:

$$L_{V\&S} = 1 - \frac{cos^{-1}(similarity)}{\pi} \tag{4}$$

The direction of the feature vectors conveys the decision tendencies of their respective classes, with closely aligned vector angles indicating harmonious decisions. Euclidean distance is not used in this part because the information of different modalities has different degrees of influence on the decision. The difference between multimodal is reflected in the length of the vectors. In order not to miss this part of the information, a relative difference like cosine similarity is therefore used as the metric.

For the regression task, MSE is employed to enforce constraint, expressed as:

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n} [y_i - \widehat{y}_i]^2 \tag{5}$$

where y_i is the network prediction result, and $\widehat{y}_i$ is the true value. The total loss function can be expressed as follows:

$$L_{Relative} = \lambda_1 L_{V\&S} + \lambda_2 L_{MSE} \tag{6}$$

Here, λ_1 and λ_2 are hyper-parameters that balance the two parts of the loss. $L_{Relative}$ is the final loss function. The process of hyperparameter tuning yielded the values for λ_1 and λ_2, which were established as 10 and 1, respectively. These parameter settings are finely tuned to optimize the model's performance in the context of the MCRNet architecture.

5. Results and discussion

This paper focused on using 304 stainless steel as the base material, with austenitic stainless steel (Cr19Ni9) used for the welding wire. The welding power source was the Fronius CMT advanced 4000R. As a Directed Energy Deposition (DED) process, wire arc additive manufacturing (WAAM) has become an increasingly popular and viable alternative method for producing large metal parts, showcasing its impressive capabilities and versatility.

Data were collected from 20 welds in a full-penetration state in the first stage. A total of 27,342 sets of video, sound, and backside melt width data were gathered to train the model. The melt pool images, sound spectrograms, and backside melt width are manually aligned based on the arc start and rest times concerning their respective positions. Our observations indicate that each set of melt pool images and sound spectrograms theoretically correspond to the penetration position

at the exact time they were recorded. The total welding distance for a single weld is 12 cm. The first and last 1 cm of each weld's data is discarded to eliminate instability during the arc start and rest stages.

A rigorous validation strategy was employed to ensure the robustness and generalizability of the model. Data from the collected welds were intentionally perturbed to simulate real-world variabilities and complexities. Subsequently, a stratified sampling approach was adopted, allocating 90 % of the data to the training set and reserving 10 % for the test set. This distribution was meticulously designed to prevent any overlap between the training and test sets, thereby ensuring the validity and reliability of the model's performance evaluation.

Subsequently, the trained MCRNet was applied to monitor online data from three additional welds and served as a validation set. These welds were also included in ablation studies and comparative experiments to assess the model's effectiveness further. The details of the training and experimental dataset are outlined in Table 1.

The groove angle variation range is 60°-120° (Fig. 8). Note that our data acquisition is continuous during online monitoring, so it is possible to go back and retrieve the first few frames of images. The data relationship is shown in Fig. 9, which takes the weld cross-section at the torch's position during welding to illustrate the data. Each video input consists of the target image at the given time, along with the two preceding frames. The sound input includes a 64 × 64 spectrogram from 2144 sound sequences recorded within the previous period of approximately 2–3 CMT pulse periods. The integration prediction process is all based on the current torch position, and there is an overlap of data used in the backtracking process. That reflects the fact that penetration is a cumulative process.

The MCRNet is implemented using the PyTorch framework and trained with a batch size of 32. To optimize the network's performance, the AdamW optimizer is employed, configured with a momentum parameter of 0.99 and a weight decay value of 0.01. A cosine annealing strategy is used, which is a significant improvement compared to the common gradient learning rate strategy. All models are trained on an NVIDIA TITAN RTX GPU and trained with 100 epochs.

The predicted and actual melt widths of the three experimental welds were compared using Mean Absolute Error (MAE) and MSE as comparators. Fig. 10 shows the error comparison between the actual and predicted values for the three verification welds in the full penetration process, which is divided into upper and lower parts. The upper section presents a direct comparison between predicted and actual backside weld width values, while the lower section displays the absolute error for each data set. The errors were calculated for the three beads: the MAE reached 0.2538 mm, and the MSE reached 0.1190 mm. An MSE of 0.12 mm is roughly 2 % (0.12 mm divided by 6 mm) of the typical backside melt width.

5.1. Ablation experiments

In this research, ablation experiments were conducted to assess the effectiveness of certain improvements made to the network architecture. These improvements include the implementation of the 3Dcov block, the MFE (Multi-Feature Extraction) block, and the SE (Squeeze-and-Excitation) block. The purpose of these experiments was to determine the individual contributions of these components to the overall performance of the network.

The process involved modifying or removing these specific blocks from the network and observing the changes in performance. This method allows for an empirical evaluation of each component's utility.

Table 1
Dataset.

Type	Training set	Test set	Validation set
Images	24,606	2736	3768
Sound	24,606	2736	3768

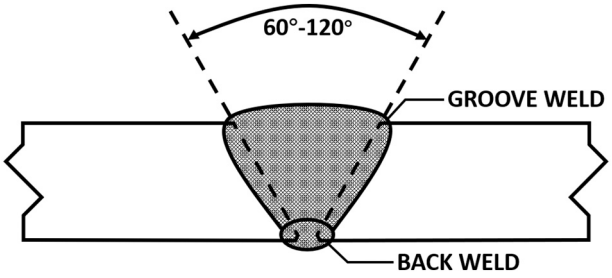

Fig. 8. Schematic diagram of complex groove welding penetration.

The 3Dcov block is intended for improved spatiotemporal feature extraction, the MFE block for enhanced feature aggregation, and the SE block for efficient feature recalibration. The results of these experiments are presented in Table 2.

Table 2 provides a critical insight into the functional significance of the well-designed modules within the MCRNet architecture. The results show that MCRNet can predict the backside melt width using either video or audio data independently. However, a significant enhancement in predictive accuracy was observed when the network utilized multimodal information, integrating both video and audio data. This increase in accuracy with multimodal inputs aligns with the overarching design philosophy of MCRNet, which emphasizes the synergistic potential of multi-feature analysis. The integration of diverse data types allows for a more comprehensive understanding of the welding process, leading to more accurate predictions.

A notable observation from the data is the pronounced impact resulting from the omission of the multi-feature screening module, which comprises the MFE and SE blocks. This module is identified as the cornerstone of the MCRNet, playing a pivotal role in its performance. In its absence, the network's operation is reduced to the capabilities of a standard 3 × 3 convolution, leading to a marked decrease in performance. Specifically, the MSE for multimodal input in such a scenario was observed to be inferior to that of a ResNet-34 model in subsequent experiments. Additionally, these experiments afforded the opportunity to evaluate the influence of different loss functions on the results, providing a comprehensive understanding of their impact on model performance.

Furthermore, the MSE metric, compared to the MAE, is more indicative of the model's robustness. Consequently, MSE was primarily utilized for analysis in the subsequent experiments.

The research also addressed the critical requirement of online monitoring, which imposes specific timing constraints on the model. In this context, the reparameterization mechanism within the MFE block assumes a significant role. It was found to expedite the inference time of MCRNet with multimodal inputs by 40 %, as evidenced in Table 3. This enhancement not only contributes to the efficiency of the model but also underscores the thoughtful and effective design of MCRNet. Such optimizations are instrumental in establishing the network's superiority in terms of both accuracy and real-time performance.

5.2. Comparison experiments

Experiments validate the rationale and superiority of the neural network proposed by comparing it with mainstream methods. As the novel approach suggested in this paper incorporates the fusion of video and sound on the melt pool time series, no existing work is available for direct comparison. In our comparison experiments, the performance of single images, video frames, and sound spectrograms was evaluated using state-of-the-art models, including ResNet, DenseNet, Swin Transformer, and ConvNeXt. Additionally, the fusion results of both modalities were also compared. The results of these comparative experiments are documented in Table 4. For clarity, the input dimensions for the video, image, and sound data were standardized to 3 × 256 × 256, 1 ×

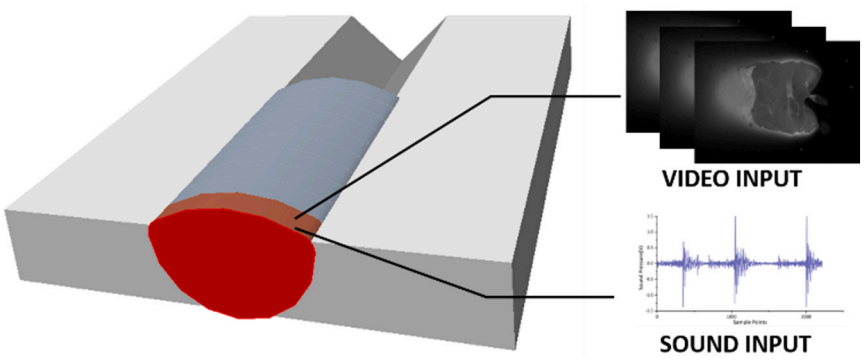

Fig. 9. Raw data of groove welding penetration.

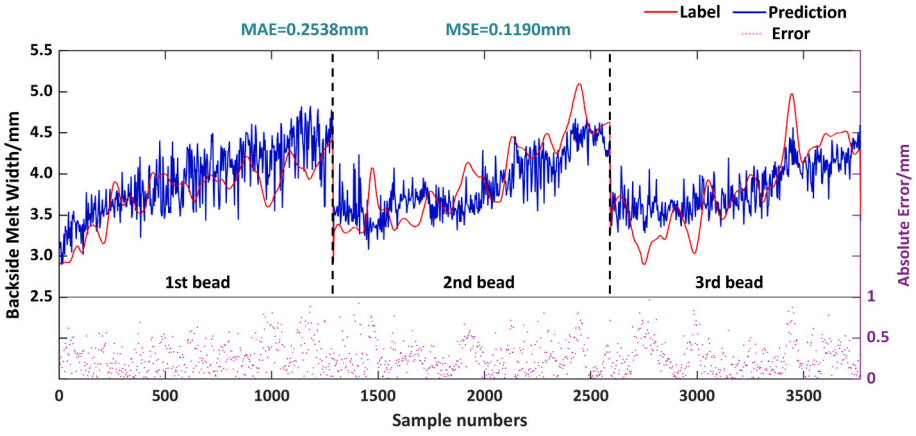

Fig. 10. Results of validation experiments.

Table 2
Ablation experiment results.[a]

Network	MAE (mm)	MSE (mm)
MCRNet	**0.2538**	**0.1190**
MCRNet-Video only	0.2833	0.1555
MCRNet-Sound only	0.2893	0.1796
Without $L_{V\&S}$	0.2754	0.1314
Without 3Dcov block	0.2876	0.1514
Without MFE block	0.3565	0.2041
Without SE block	0.2984	0.1833

[a] In the table, bolded results indicate data obtained using the network model proposed in this paper.

256 × 256, and 1 × 128 × 128 respectively. All networks employed in these experiments were configured to use the Mean Squared Error (L_{MSE}) loss function. The metrics compared include MSE, inference time, Params, and FLOPs.

5.2.1. Neural network model performance

Our experiments compared the MSE and inference time of several mainstream networks on the validation set in Fig. 11. This investigation revealed notable disparities in MSE, particularly when processing solely video or image inputs. ResNet-34, Swin-T, and ConvNeXt-T demonstrated diminished proficiency in minimizing MSE. In contrast, DenseNet-121 exhibited relatively enhanced performance, albeit with a 26 % shortfall in efficiency compared to our proposed MCRNet.

The MCRNet architecture, distinctively incorporating a 3D extraction module, excels in handling multi-frame images, especially within

Table 3
Ablation experiment results (time).[a]

Network	Time (ms)
MCRNet	**17.4**
MCRNet- without reparameterization	24.3

[a] In the table, bolded results indicate data obtained using the network model proposed in this paper.

Table 4
Comparison experiment results.[a]

Network	Input source	MSE (mm)	Time (ms)	Params (M)	FLOPs(M)
MCRNet	Video + Sound	**0.1190**	**17.4**	99.27	6242.71
MCRNet	Video	**0.1555**	**9.9**	13.19	4439.43
CNN-LSTM	Video	0.2088	50.2	22.50	14,416.24
ResNet-34	Video	0.2278	11.4	21.80	4804.73
DenseNet-121	Video	0.1970	20.2	43.82	3819.37
Swin-T	Video	0.2642	25.2	5.45	1261.19
ConvNeXt-T	Video	0.3626	9.5	27.80	5818.47
ResNet-34	Image	0.2556	11.4	21.80	4701.97
DenseNet-121	Image	0.1810	20.2	43.81	3716.61
Swin-T	Image	0.2695	25.2	5.45	1257.00
ConvNeXt-T	Image	0.3857	9.5	27.80	5805.89
MCRNet	Sound	**0.1796**	**9.6**	13.19	1053.48
ResNet-34	Sound	0.1942	11.1	23.96	4530.78
DenseNet-121	Sound	0.1893	20.0	11.05	924.03
Swin-T	Sound	0.2025	23.8	5.45	314.18
ConvNeXt-T	Sound	0.2286	9.2	27.80	1451.47

[a] In the table, bolded results indicate data obtained using the network model proposed in this paper.

sequence signal inputs. Notably, CNN-LSTM [26] is a baseline method for multi-frame video processing. Resnet-34 was used as the CNN part of CNN-LSTM, and it found that it did improve the results, but only to a limited extent. Instead, due to the Long Short-Term Memory (LSTM) algorithm looping multiple times based on the number of frames, the time consumption reached 50 ms, which is not the ideal time consumption.

When considering sound input alone, the sound spectrogram, similar to a single image, yielded more stable results across the networks. However, MCRNet still demonstrated a minimum 5 % improvement due to its network design that considers sound continuity.

Regarding inference time, MCRNet, ResNet-34, and ConvNeXt-T exhibited approximately half the processing time compared to DenseNet-121 and Swin-T. Notably, the inference time was independent of the input image channel. The inference time for multi-channel inputs is the same as that for single-channel inputs. This is also why CNN is more suitable than LSTM for online monitoring.

5.2.2. *Structural and computational complexity*

In the realm of computational efficiency and network architecture complexity, our research offers an in-depth comparison of Parameters (Params) and Floating Point Operations (FLOPs) across various neural network models. The Params metric quantitatively represents the total number of trainable parameters within a model, serving as an indicator of both the model's size and its inherent complexity. FLOPs, on the other hand, provide a measure for evaluating the computational complexity and efficiency of models. Typically, models with lower FLOPs are associated with reduced computational demands and heightened operational efficiency.

Our empirical investigations revealed that the Params are predominantly influenced by the complexity inherent to the linear layers of the network. A noteworthy observation was that the addition of an extra linear layer could potentially result in a doubling of the Params, while the corresponding impact on FLOPs remained marginal. This delineates a crucial aspect of network design, where layer complexity significantly alters the model's parameter count without substantially influencing its computational load. Conversely, the FLOPs metric is substantially affected by the size of the input images. It is imperative to note that, although FLOPs are a critical factor in assessing computational complexity, they do not exhibit a direct correlation with the inference time. However, they are instrumental in facilitating comparative assessments of performance across diverse network models.

Referring to Table 4 in our study, when utilizing a single modality input, the MCRNet maintains a low Params profile, with FLOPs slightly inferior to those of the ResNet-34 model. This efficiency is attributed to our innovative multimodal feature extraction module, which horizontally processes information across different feature layers. This approach effectively reduces the overall depth of the network while simultaneously enhancing data processing capabilities.

In scenarios involving dual modality inputs, the integration of our fusion module results in a significant escalation in Params. However, this increase is meticulously counterbalanced by a well-regulated growth in FLOPs. The substantial rise in Params primarily stems from the computational demands of the numerous linear layers, which, despite their complexity, have a relatively modest impact on the overall computational load. In contrast, the front-end convolutional computations, which form the more time-intensive segment of the network, are designed for optimal efficiency.

In terms of overall complexity, our model exhibits a parallelism with mainstream deep learning models in size. The strategic incorporation of the reparameterization concept within our network architecture substantially diminishes computational time, thereby enhancing the network's suitability for real-time monitoring applications. This design philosophy ensures that while the network remains robust and capable of handling complex data inputs, it does so with an efficiency that aligns with the operational demands of online monitoring systems.

In summary, MCRNet is specifically designed to address the online detection of melt pools. Therefore, the network does not prioritize depth but focuses on extracting rich features within a limited depth. This approach allows for high prediction accuracy and low time consumption. With dual inputs, MCRNet achieves an inference time of 17.4 ms (equivalent to 57 FPS), which is sufficient for online monitoring. Additionally, it demonstrates at least an 18 % improvement in MSE compared to a single input and at least a 44 % improvement when compared to other deep learning networks.

6. Conclusion

This paper introduces a novel network called MCRNet, which aims to enhance the characteristics of multimodal continuous signals for effective online penetration prediction in robotic welding. The key contributions of this research can be summarized as follows.

(1) A deep CNN network based on multimodal continuous signals is presented, and the model effectively combines video frames and sound signals to improve the accuracy of regression results. A fusion module is introduced, which reconstructs and integrates the information from both modalities, ensuring a seamless fusion. Additionally, the similarity loss function constrains the learned representations to enhance the multimodal features further.

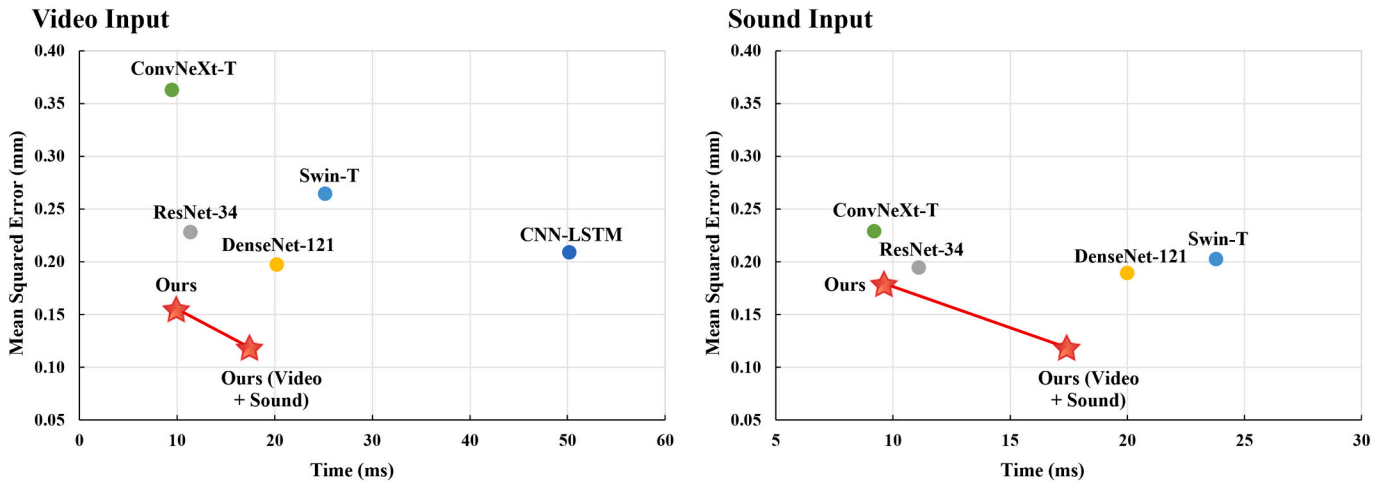

Fig. 11. MCRNet vs. other network of comparison experiments.

(2) This paper delves into a new feature extraction backbone designed to extract richer features within a limited depth. To accomplish this, a multi-feature screening module effectively captures informative features. By leveraging this module, MCRNet achieves high prediction accuracy while maintaining low time consumption. Additionally, a 3Dconv module enhances the feature extraction process for video sequences, further improving the network's performance.
(3) The experimental results demonstrate that the proposed network exhibits exceptional efficiency and is well-suited for intelligent industrial assembly line production deployment. Our model achieves an MSE of 0.1190 mm on the validation set, showcasing its accuracy in real-time penetration prediction. Furthermore, the network operates at an inference speed of 57 frames per second (FPS), ensuring timely and accurate predictions in a production environment.

This work demonstrates an example of multimodal information fusion using the tool of deep learning rationally and advances the level of penetration monitoring. Our approach is to use deep learning to guide automated applications to reduce errors that can be applied to various manufacturing processes. With enough data and expert knowledge, it can enable automatic corrections that outperform human operators.

This can be extended to monitor various indicators of melt pool condition, while monitoring information can be extended to temperature, stress, spectrum, 3D point cloud, etc. However, the inclusion of these multimodal information needs to be carefully considered. One aspect is the difficulty and stability of obtaining them, while another aspect is whether they can truly reflect the penetration state. In today's rapidly evolving hardware and software advancements, the inclusion of more relevant welding information and faster computational speeds can make intelligent welding a possibility.

Declaration of competing interest

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

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62101265, U23A20283, 62271263), the China Postdoctoral Science Foundation (2021M691592), and the Fundamental Research Funds for the Central Universities (No. 30922010705).

References

[1] Guo S, Peng Y, Cui C, Gao Q, Zhou Q, Zhu J. Microstructure and mechanical characterization of re-melted Ti-6Al-4V and Al-Mg-Si alloys butt weld. Vacuum 2018:154. https://doi.org/10.1016/j.vacuum.2018.04.048.
[2] Chang Y, Yue J, Guo R, Liu W, Li L. Penetration quality prediction of asymmetrical fillet root welding based on optimized BP neural network. J Manuf Process 2020:50. https://doi.org/10.1016/j.jmapro.2019.12.022.
[3] Hensel J, Köhler M, Uhlenberg L, Castro JD e., Dilger K, Faß M, et al. Laser welding of 16MnCr5 butt welds with gap: resulting weld quality and fatigue strength assessment. Welding in the World 2022;66. doi:https://doi.org/10.1007/s40194-022-01306-4.
[4] Shen W, Hu T, Zhang C, Ye Y, Li Z. A welding task data model for intelligent process planning of robotic welding. Robot Comput Integr Manuf 2020:64. https://doi.org/10.1016/j.rcim.2020.101934.
[5] Mortazavian E, Wang Z, Teng H. Repair of light rail track through restoration of the worn part of the railhead using submerged arc welding process. Int J Adv Manuf Technol 2020:107. https://doi.org/10.1007/s00170-020-05208-x.
[6] Wu D, Huang Y, Chen H, He Y, Chen S. VPPAW penetration monitoring based on fusion of visual and acoustic signals using t-SNE and DBN model. Mater Des 2017:123. https://doi.org/10.1016/j.matdes.2017.03.033.
[7] Gao Y, Wang Q, Xiao J, Zhang H. Penetration state identification of lap joints in gas tungsten arc welding process based on two channel arc sounds. J Mater Process Technol 2020:285. https://doi.org/10.1016/j.jmatprotec.2020.116762.
[8] 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. https://doi.org/10.1016/j.jmapro.2021.11.055.
[9] 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. https://doi.org/10.1016/j.jmapro.2020.08.028.
[10] Yu P, Xu G, Gu X, Zhou G, Tian Y. A low-cost infrared sensing system for monitoring the MIG welding process. Int J Adv Manuf Technol 2017:92. https://doi.org/10.1007/s00170-017-0515-7.
[11] Reza Tabrizi T, Sabzi M, Mousavi Anijdan SH, Eivani AR, Park N, Jafarian HR. Comparing the effect of continuous and pulsed current in the GTAW process of AISI 316L stainless steel welded joint: microstructural evolution, phase equilibrium, mechanical properties and fracture mode. J Mater Res Technol 2021:15. https://doi.org/10.1016/j.jmrt.2021.07.154.
[12] Zhang Z, Ren W, Yang Z, Wen G. Real-time seam defect identification for Al alloys in robotic arc welding using optical spectroscopy and integrating learning. Measurement (Lond) 2020:156. https://doi.org/10.1016/j.measurement.2020.107546.
[13] Jaypuria S, Bondada V, Kumar Gupta S, Kumar Pratihar D, Chakrabarti D, Jha MN. Prediction of electron beam weld quality from weld bead surface using clustering and support vector regression. Expert Syst Appl 2023:211. https://doi.org/10.1016/j.eswa.2022.118677.
[14] Lei Z, Shen J, Wang Q, Chen Y. Real-time weld geometry prediction based on multi-information using neural network optimized by PCA and GA during thin-plate laser welding. J Manuf Process 2019:43. https://doi.org/10.1016/j.jmapro.2019.05.013.
[15] Xiong J, Zou S. Active vision sensing and feedback control of back penetration for thin sheet aluminum alloy in pulsed MIG suspension welding. J Process Control 2019:77. https://doi.org/10.1016/j.jprocont.2019.03.013.
[16] Lu J, He H, Shi Y, Bai L, Zhao Z, Han J. Quantitative prediction for weld reinforcement in arc welding additive manufacturing based on melt pool image and deep residual network. Addit Manuf 2021:41. https://doi.org/10.1016/j.addma.2021.101980.
[17] Nomura K, Fukushima K, Matsumura T, Asai S. Burn-through prediction and weld depth estimation by deep learning model monitoring the melt pool in gas metal arc welding with gap fluctuation. J Manuf Process 2021:61. https://doi.org/10.1016/j.jmapro.2020.10.019.
[18] Yang L, Song S, Fan J, Huo B, Li E, Liu Y. An automatic deep segmentation network for pixel-level welding defect detection. IEEE Trans Instrum Meas 2022:71. https://doi.org/10.1109/TIM.2021.3127645.
[19] K. He, X. Zhang SR and JS. "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 770–778. doi: https://doi.org/10.1109/CVPR.2016.90. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016.
[20] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, vol. 2017- January, 2017. doi:https://doi.org/10.1109/CVPR.2017.243.
[21] Wang Y, Han J, Lu J, Bai L, Zhao Z. TIG stainless steel melt pool contour detection and weld width prediction based on res-Seg. Metals (Basel) 2020:10. https://doi.org/10.3390/met10111495.
[22] Baek D, Moon HS, Park SH. In-process prediction of weld penetration depth using machine learning-based melt pool extraction technique in tungsten arc welding. J Intell Manuf 2022. https://doi.org/10.1007/s10845-022-02013-z.
[23] Yu R, Kershaw J, Wang P, Zhang YM. Real-time recognition of arc weld pool using image segmentation network. J Manuf Process 2021:72. https://doi.org/10.1016/j.jmapro.2021.10.019.
[24] Lv N, Zhong J, Chen H, Lin T, Chen S. Real-time control of welding penetration during robotic GTAW dynamical process by audio sensing of arc length. Int J Adv Manuf Technol 2014:74. https://doi.org/10.1007/s00170-014-5875-7.
[25] Meng Q, Zhou X, Li J, Cui Z, Wang Y, Zhang H, et al. High-throughput laser fabrication of Ti-6Al-4V alloy: part I. Numerical investigation of dynamic behavior in melt pool. J Manuf Process 2020:59. https://doi.org/10.1016/j.jmapro.2020.10.008.
[26] Yu R, Kershaw J, Wang P, Zhang YM. How to accurately monitor the weld penetration from dynamic weld Pool serial images using CNN-LSTM deep learning model? IEEE Robot Autom Lett 2022:7. https://doi.org/10.1109/LRA.2022.3173659.
[27] Liu T, Wang J, Huang X, Lu Y, Bao J. 3DSMDA-net: an improved 3DCNN with separable structure and multi-dimensional attention for welding status recognition. J Manuf Syst 2022:62. https://doi.org/10.1016/j.jmsy.2021.01.017.
[28] Chen Z, Wang Z, Wang F, Liang X, Liu W, Wu S, et al. Feasibility study on sensing and prediction of backside weld geometry in cold metal transfer welding of X65 pipeline in the vertical-up position. J Manuf Process 2023:85. https://doi.org/10.1016/j.jmapro.2022.12.031.
[29] Wu D, Huang Y, Zhang P, Yu Z, Chen H, Chen S. Visual-acoustic penetration recognition in variable polarity plasma arc welding process using hybrid deep learning approach. IEEE. Access 2020:8. https://doi.org/10.1109/ACCESS.2020.3005822.
[30] Liu Y, Zhang W, Zhang Y. Dynamic neuro-fuzzy-based human intelligence modeling and control in GTAW. IEEE Trans Autom Sci Eng 2015:12. https://doi.org/10.1109/TASE.2013.2279157.
[31] Gao P, Wu Z, Wang Y, Lu J, Zhao Z. Method for monitoring and controlling penetration of complex groove welding based on online multimodal data. J Intell Manuf 2023. https://doi.org/10.1007/s10845-023-02107-2.

[32] Zhang Y, Wang Q, Liu Y. Adaptive intelligent welding manufacturing. Weld J 2021: 100. https://doi.org/10.29391/2021.100.006.
[33] Zheng B, Li Y, Ao S, Zhang X, Zhang D, Manladan SM, et al. Narrow gap welding of X80 steel using laser-CMT hybrid welding with misaligned laser and arc. Crystals (Basel) 2022:12. https://doi.org/10.3390/cryst12060832.
[34] Sturmel N, Daudet L. Signal reconstruction from STFT magnitude: A state of the art. DAFx: Proceedings of the International Conference on Digital Audio Effects; 2011.
[35] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. Proceedings of the IEEE International Conference on Computer Vision 2021. https://doi.org/10.1109/ICCV48922.2021.00986.
[36] Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2022- June, 2022. doi:https://doi.org/10.1109/CVPR52688.2022.01167.
[37] Ding X, Zhang X, Han J, Ding G. Diverse branch block: building a convolution as an inception-like unit. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2021. https://doi.org/10.1109/CVPR46437.2021.01074.
[38] Guo MH, Liu ZN, Mu TJ, Hu SM. Beyond self-attention: external attention using two linear layers for visual tasks. IEEE Trans Pattern Anal Mach Intell 2023:45. https://doi.org/10.1109/TPAMI.2022.3211006.