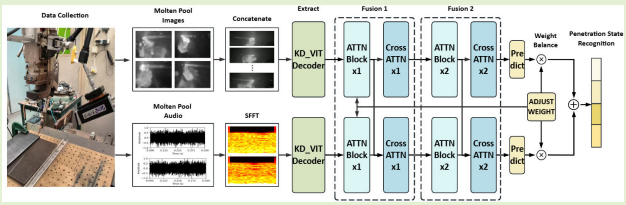
DMAW: A Dynamic Multimodal Measurement Fusion Network With Attention for Reliable Welding Process Monitoring Under Harsh Industrial Environments

Jiawei Fan, Ting Yuan, Haonan Zhang, Songlin Li, Uwe D. Hanebeck, *Fellow, IEEE*, Zhuguo Li, Edmond Q. Wu, *Senior Member, IEEE*, and Jiuchao Qian

***Abstract*—In laser welding, harsh environmental conditions such as arc light interference, spatter, and strong background noise pose significant challenges to accurate process monitoring and measurement. Traditional single-modality sensing methods are often inadequate for a comprehensive characterization of welding states and remain highly vulnerable to noise contamination. To overcome the limitations, a dynamic multimodal attention-based weighting (DMAW) network is proposed, which integrates visual and acoustic information to achieve more comprehensive welding state characterization. First, modality-specific feature extractors are trained using unsupervised knowledge distillation (KD) to learn welding-relevant semantic representations. Next, the extracted features are passed through a cross-attention module to enable intermodal interaction and suppress noise. Finally, a dynamic reliability-weighted fusion is proposed that adaptively adjusts modality contributions, thereby reducing measurement uncertainty and enhancing robustness under varying conditions. Experimental validation on a dedicated laser welding platform demonstrates that the proposed framework achieves superior accuracy and resilience in welding state monitoring, highlighting its potential as a reliable solution for intelligent industrial welding systems.**



***Index Terms*—Cross-attention, dynamic balance, laser welding, multimodal measurement, unsupervised feature learning.**

I. Introduction

TRADITIONAL welding state detection methods, based on machine learning techniques, often suffer from limited adaptability and accuracy. With the advancement of deep learning technologies, numerous studies have explored methods such as convolutional neural networks (CNNs) [7], [8], [9], [10], unsupervised methods [11], [12], [13], [14], and attention-based models [4], [13], [15] to improve the performance of welding state detection. Compared with traditional techniques, deep learning methods show more robustness in different scenarios by virtue of their strong adaptive ability. CNN-based methods have been widely employed for visual feature extraction in welding state detection. Ren et al. [7] introduced a time–frequency spectrogram-based CNN for classifying welding states in arc welding, while Feng et al. [8] proposed an ensemble CNN detection model to improve classification accuracy. Similarly, Ma et al. [16] developed a YOLO-based model for weld seam feature point extraction. The above methods leverage CNNs' ability to automatically learn local spatial features. However, CNNs in welding detection may struggle to capture complex global patterns and dynamic temporal features, while also being prone to overfitting due to the scarcity of samples and the diversity of welding conditions, affecting the model's generalization ability.

Besides CNNs, attention models have become a key component in the development of deep learning models [15], with existing research demonstrating the potential in the field of welding detection. Zhao et al. [4] proposed an auditory

attention model that integrates attention mechanisms with LSTM networks for penetration state recognition in gas tungsten arc welding, significantly improving detection accuracy. In addition, Caron et al. [13] explored the application of attention architectures in visual detection algorithms, revealing that attention-based models outperform traditional CNNs in capturing global features during welding processes. Therefore, attention mechanisms offer substantial potential to enhance the accuracy and robustness of welding state detection in complex industrial settings.

Meanwhile, self-supervised strategies, which do not rely on large amounts of labeled data, enhance semantic extraction capabilities by learning from the internal structure of the samples [17]. The characteristic is particularly valuable in welding state detection and has attracted significant attention from researchers in the field. Wan et al. [12] proposed an unsupervised feature mapping model for anomaly detection, which proves especially effective in detecting welding defects when labeled data is scarce. Similarly, Cui et al. [14] employed a Transformer-based encoder trained on defect-free images to improve representation learning. The advances highlight the potential of self-supervised methods to overcome data limitations, enabling detection algorithms to have greater adaptability in varying environments.

However, the dependence on unimodal inputs makes the above methods vulnerable to domain shifts when applied to different industrial conditions [18]. Single-modal deep learning methods can achieve high accuracy in environments with regular data quality distribution. Due to environmental effects such as noise interference or smoke occlusion, effective features are easy to be completely masked in some time periods, and the performance of single-modal methods will deteriorate. The limited robustness of single-modal methods highlights the need for multimodal learning [19], where integrating complementary information enhances detection reliability and generalization capabilities in complex scenarios [5].

Multimodal fusion strategies are commonly categorized into early and late fusion [6], [20], [21], [22], [23], [24], [25], [26]. Early fusion integrates raw data from different modalities before feature extraction [20], [21], while late fusion processes each modality independently and combines the results later [6], [22], [23]. In welding state detection, due to the high-intensity industrial noise and laser contamination, late fusion, where features are extracted first before fusion, is typically more suitable. However, existing generic multimodal fusion methods primarily assume that both modalities provide high-quality information during fusion, aiming to enhance overall accuracy. In contrast, the quality of modalities in welding processes is dynamic; at certain moments, one modality may be contaminated by noise, resulting in lower confidence, and in such cases, the model needs to rely on the other modality for support. It requires the fusion network to dynamically adjust the weights of the modalities, assigning a higher weight to the more reliable modality. Therefore, a fusion network needs to be designed to adaptively adjust the weights based on the fluctuating quality of each modality, ensuring robustness in environments where the quality of modalities alternates.

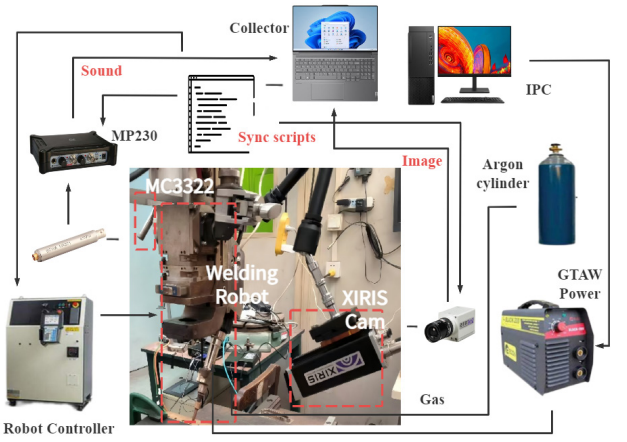


Fig. 1. Schematic of laser welding platform.

In this article, a dynamically balanced multimodal fusion welding state detection method is proposed, which aims to make full use of image and audio modalities to improve the detection performance. For modality feature extraction, an unsupervised learning-based model is employed, where welding sample pairs are fine-tuned to enhance the semantic representation of welding states. To address noise interference across different modalities during the welding process, a cross-attention module is designed. Through nonlinear mapping and cross-attention-based feature alignment, the model suppresses noise while extracting discriminative welding-related features with higher reliability. Additionally, to address the issue of imbalanced quality between modalities at different times, a dynamic weight-balancing network has been designed. It ensures that robustness is maintained in the model even when one modality exhibits low quality.

Moreover, existing publicly available datasets for welding state recognition are predominantly single-modal, mainly consisting of color or grayscale welding images or audio segments of welding states. However, multimodal models require a sufficient number of aligned image–audio pairs for effective training. As shown in Fig. 1, an integrated image–audio data acquisition platform was developed to support the experimental validation of the proposed framework. The dataset will be made publicly available in the future, enabling the broader research community to explore and advance multimodal welding state detection. To summarize, our contributions are as follows.

1) A novel welding state detection method is proposed, combining image and audio modalities to enhance the model's accuracy and robustness.
2) An unsupervised ViT extractor is developed to capture contour, texture, and local details from welding images and audio.
3) To address the noise issues in harsh environments, a multimodal attention module with self- and cross-attention is designed to align complementary cues under noise.
4) To tackle the issue of data quality imbalance, a dynamic weight adjustment model is proposed, balancing the contribution of data from modalities with varying quality.

II. Method

In this study, a welding state detection method is proposed, combining image and audio feature information to enhance both accuracy and robustness during the detection process. As shown in Fig. 2, the method consists of three crucial stages: 1) feature extraction from welding modalities through an unsupervised encoder; 2) feature refinement via a multimodal attention mechanism; and 3) dynamic weighting for the computation of classification confidence.

A. Self-Distilled Extractor

We adopt ViT-Small as the backbone encoder for multimodal welding data. Compared with larger variants, ViT-Small provides sufficient capacity to capture semantic features while maintaining a low memory footprint, which is essential in welding scenarios where long video sequences with thousands of frames must be processed under strict GPU memory constraints.

Unlike conventional ViT-Small models trained in a supervised manner with annotated data, our encoder is adapted through an unsupervised self-distillation strategy. Similar self-distillation approaches have been shown to learn powerful representations in generic vision tasks [13], [27], [28]. In the welding domain, manual annotation is extremely difficult due to intense arc light, smoke, and spatter, which obscure object boundaries and contaminate acoustic signals. Therefore, instead of relying on labels, the encoder learns by enforcing consistency across different augmented views of the same sample.

As shown in Fig. 3, both teacher and student networks share the ViT-Small architecture. The student parameters θ are updated by backpropagation, while the teacher parameters ϕ are updated as an exponential moving average (EMA) of the student

$$\phi \leftarrow m\,\phi + (1-m)\,\theta \tag{1}$$

where m is a momentum coefficient. Although the teacher has no external supervision, the EMA update acts as a temporal filter, producing outputs that are more stable than the student's instantaneous predictions [13]. The stability, rather than absolute correctness, is sufficient to serve as a learning target under noisy welding conditions.

During training, each welding image or audio spectrogram is augmented into multiple views. Global views $\mathcal{G}$ capture holistic seam geometry and arc distribution, while local views $\mathcal{L}$ emphasize subtle details such as molten pool ripples, spatter particles, or smoke trails. Passing a view x through the encoder and projection head $h(\cdot)$ yields

$$\mathbf{z}^s = h(E_\theta(x)), \qquad \mathbf{z}^t = h\big(E_\phi(x)\big) \tag{2}$$

where $\mathbf{z}^s$ and $\mathbf{z}^t$ denote the student and teacher representations, respectively. Then, the representations are converted into predictive distributions using temperature-scaled Softmax with a centering vector $\mathbf{c}$ that prevents collapse by compensating for overly dominant signals such as bright arcs or strong acoustic frequencies

$$\mathbf{p}^s(x) = \mathrm{softmax}\left(\frac{\mathbf{z}^s}{\tau_s}\right), \qquad \mathbf{p}^t(x) = \mathrm{softmax}\left(\frac{\mathbf{z}^t-\mathbf{c}}{\tau_t}\right). \tag{3}$$

For welding images, the loss aligns student predictions from local crops $\mathcal{L}^{im}$ with teacher predictions from global crops $\mathcal{G}^{im}$

$$\mathcal{L}_{im} = -\frac{1}{|\mathcal{G}^{im}|\,|\mathcal{L}^{im}|}\sum_{x'\in\mathcal{G}^{im}}\sum_{x\in\mathcal{L}^{im}}\sum_{k=1}^{K}\mathbf{p}_k^t(x')\log\mathbf{p}_k^s(x). \tag{4}$$

Here, $\mathcal{G}^{im}$ encodes the seam trajectory and electrode region, while $\mathcal{L}^{im}$ isolates local phenomena such as porosity, spatter bursts, or molten pool boundaries.

Similarly, for welding audio spectrograms, global crops $\mathcal{G}^{au}$ preserve long-term arc stability, while local crops $\mathcal{L}^{au}$ focus on transient events such as droplet transfer or spatter crackling. The corresponding loss is

$$\mathcal{L}_{au} = -\frac{1}{|\mathcal{G}^{au}|\,|\mathcal{L}^{au}|}\sum_{x'\in\mathcal{G}^{au}}\sum_{x\in\mathcal{L}^{au}}\sum_{k=1}^{K}\mathbf{p}_k^t(x')\log\mathbf{p}_k^s(x). \tag{5}$$

The training paradigm is feasible despite the absence of labels because it leverages a bootstrapping mechanism [13], [28]. The teacher, being an EMA of the student, provides temporally stabilized targets, while the cross-view consistency loss forces the student to reconcile local noisy details with global welding semantics. In practice, training is deemed successful once the learned representations form discriminative clusters corresponding to different welding states, as validated by downstream recognition tasks.

After the unsupervised self-distillation training converges, both the image-based and audio-based ViT-Small encoders are frozen and no longer updated. Each encoder is domain-adapted to its modality through the pretraining process and subsequently serves as a fixed feature extractor tailored to welding data. For a given welding image I or audio spectrogram S, the frozen encoders directly output the feature matrices $\mathbf{v}$ and $\mathbf{a}$. The representations are then forwarded to downstream fusion modules without further fine-tuning of the backbone. Formally, the frozen extractors are written as follows:

$$\mathbf{v} = E_{\theta^{im}}^{\text{frozen}}(I), \qquad \mathbf{a} = E_{\theta^{au}}^{\text{frozen}}(S) \tag{6}$$

where $E_{\theta^{im}}^{\text{frozen}}$ and $E_{\theta^{au}}^{\text{frozen}}$ denote the ViT-Small encoders for the image and audio modalities, respectively. Freezing the encoders ensures that welding-specific knowledge captured during the unsupervised self-distillation stage is preserved, while reducing computational overhead and stabilizing downstream training.

B. Attention Model

Following the feature extraction stage, the representations from video and audio are refined and aligned using stacked self-attention and cross-attention layers. The design is particularly beneficial for welding, where visual frames may be degraded by arc flare or smoke, and audio spectrograms may be corrupted by background noise. By allowing each modality to leverage complementary cues from the other, robust multimodal embeddings are obtained. As shown in Fig. 4, for a modality $m \in \{im, au\}$, given its input sequence $\mathbf{X}^m$, we define the query, key, and value projections as follows:

$$\mathbf{Q}^m = \mathbf{X}^m W_q^m, \quad \mathbf{K}^m = \mathbf{X}^m W_k^m, \quad \mathbf{V}^m = \mathbf{X}^m W_v^m. \tag{7}$$

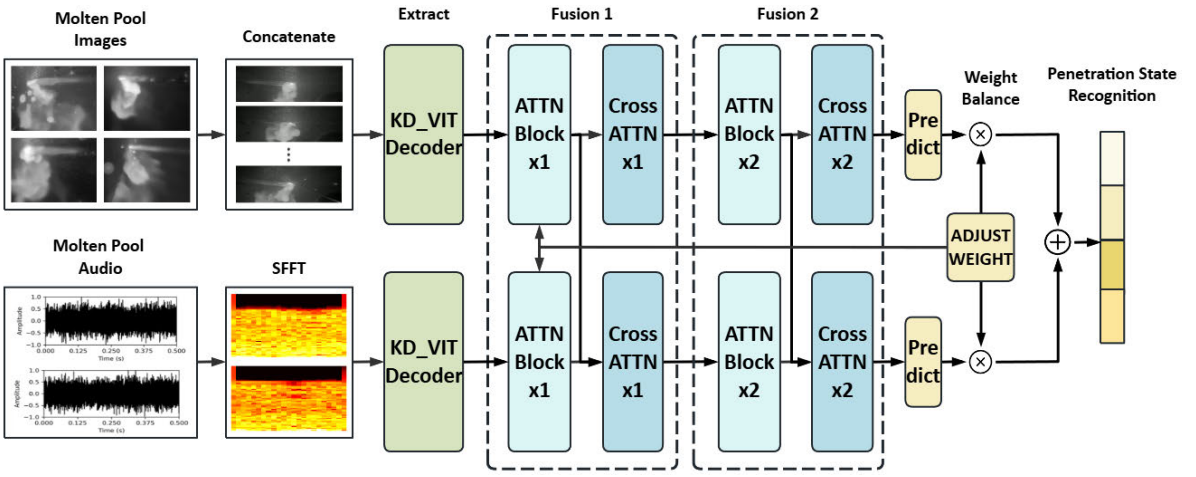


Fig. 2. Working principles of the proposed multimodal fusion framework. We encode multimodal inputs using a ViT-based encoder, refine them through attention mechanisms, and leverage a dynamic weight adjustment model to ensure robust, accurate welding state detection.

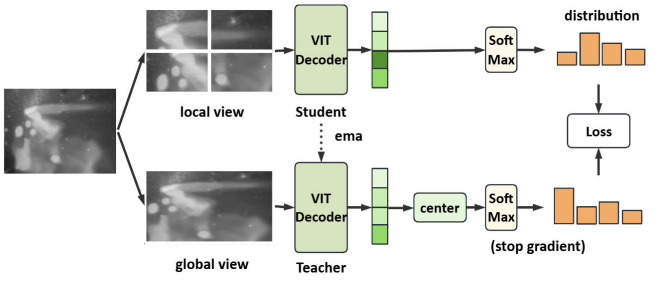


Fig. 3. Unsupervised ViT training with student–teacher self-distillation, where local and global welding views are aligned through EMA and view-consistent learning.

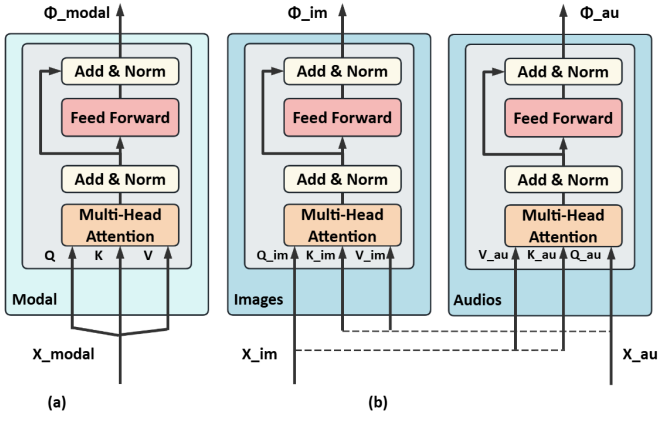


Fig. 4. Architecture of the attention modules. (a) Intramodal attention block for modeling dependencies within a single modality. (b) Cross-attention block for aligning image and audio features through multihead attention.

Self-attention is first applied within each modality to capture long-range intramodal dependencies

$$\boldsymbol{\phi}^{m} = \mathrm{softmax}\left(\frac{\mathbf{Q}^{m}\left(\mathbf{K}^{m}\right)^{\top}}{\sqrt{d_k}}\right)\mathbf{V}^{m}. \tag{8}$$

In welding, the step enables the visual encoder to model global weld pool morphology, while the audio encoder emphasizes consistent harmonic patterns corresponding to stable arc burning. Cross-attention is then introduced to enable one modality to borrow complementary cues from the other. When video attends to audio, the update is given by the following equation:

$$\boldsymbol{\phi}^{im\leftarrow au} = \mathrm{softmax}\left(\frac{\mathbf{Q}^{im}\left(\mathbf{K}^{au}\right)^{\top}}{\sqrt{d_k}}\right)\mathbf{V}^{au} \tag{9}$$

where $\mathbf{K}^{au}$ establishes alignment and $\mathbf{V}^{au}$ injects semantic information from audio into video. This is especially useful when the camera is saturated by arc light: the video queries still carry temporal context, which align with stable arc sound patterns in the audio keys, and the values provide corrective features that compensate for missing visual details.

Symmetrically, when audio attends to video

$$\boldsymbol{\phi}^{au\leftarrow im} = \mathrm{softmax}\left(\frac{\mathbf{Q}^{au}\left(\mathbf{K}^{im}\right)^{\top}}{\sqrt{d_k}}\right)\mathbf{V}^{im} \tag{10}$$

the audio features incorporate spatial seam context from video, which allows the model to avoid misinterpreting transient mechanical noise or spatter crackling as abnormal welding states, since the injected video values provide reliable geometric cues.

By alternating self-attention and cross-attention layers, the network progressively enhances both intramodal consistency and intermodal complementarity. In each cross-attention block, one modality is treated as the query that attends to the key–value representations of the other, enabling mutual guidance between visual and acoustic feature streams during feature refinement. The bidirectional interaction aligns temporally correlated cues across modalities, thereby reinforcing complementary information and improving feature robustness. The final embeddings $\boldsymbol{\phi}^{im\leftarrow au}$ and $\boldsymbol{\phi}^{au\leftarrow im}$ thus capture global weld pool evolution and transient acoustic events simultaneously, leading to accurate and noise-resilient weld state recognition.

C. Fusion Method

Both visual and auditory features are highly susceptible to variability caused by environmental interferences. To overcome these challenges, an adaptive fusion model is proposed,

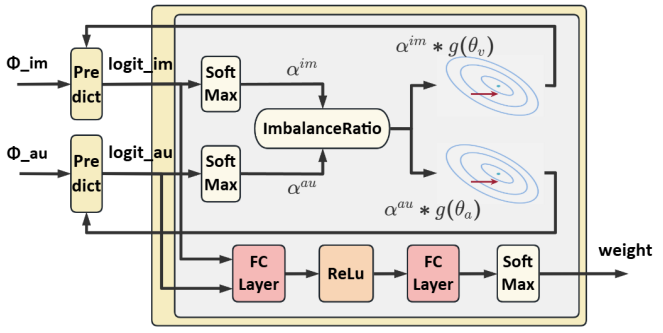


Fig. 5. Architecture of the proposed imbalance-aware multimodal fusion module.

as shown in Fig. 5, which dynamically adjusts the contributions of the image and audio modalities based on their reliability in a given welding scenario. Initially, each modality is trained independently to ensure that both the audio and visual streams possess predictive capabilities. During training, feature representations from both modalities are jointly optimized, where the back-propagated gradients are adaptively scaled according to each modality's confidence to enhance intermodal consistency. During inference, the adaptive weighting module automatically adjusts each modality's contribution based on its reliability, ensuring robust prediction under varying sensing conditions. Let $\boldsymbol{\phi}^{ai} \in \mathbb{R}^{T_a \times d}$ and $\boldsymbol{\phi}^{vi} \in \mathbb{R}^{T_v \times d}$ represent the high-level features extracted from the audio and visual modalities. Each modality's prediction is computed independently via the respective classification head $f_{\text{cls}}(\cdot)$

$$\hat{y}_{ai} = f_{\text{cls}}^{au}\left(\boldsymbol{\phi}^{ai}\right), \quad \hat{y}_{vi} = f_{\text{cls}}^{im}\left(\boldsymbol{\phi}^{vi}\right). \tag{11}$$

To enable the network to dynamically prioritize the more informative modality based on the current welding conditions, a dynamic weighting module is introduced. First, the module performs a pooling operation $M(\cdot)$ to aggregate the global context of each modality. Then, a two-layer perceptron is employed to project the concatenated pooled features into a higher-dimensional space, allowing the model to capture more complex relationships between the modalities. The output is subsequently passed through a series of nonlinear activations to further capture the intricate interactions between the modalities. Finally, category-specific weights are estimated through the following learnable projection:

$$\begin{aligned}\text{weight} = \text{softmax}\big(W_2 \cdot \text{ReLU}\big(W_1 \cdot \big[M\left(\boldsymbol{\phi}_{ai}\right) \\ M\left(\boldsymbol{\phi}_{vi}\right)\big] + \boldsymbol{b}_1\big) + \boldsymbol{b}_2\big)\end{aligned} \tag{12}$$

where W_1 and W_2 are trainable weight matrices, $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ represent the bias vectors, and ReLU$(\cdot)$ introduces a nonlinearity that helps capture more complex patterns in the feature interactions. The Softmax operation is then applied to the final output to ensure that the weights are normalized across categories. The multilayered approach improves the model's ability to dynamically adjust modality contributions based on the current task context.

The fused prediction is computed using a category-wise Hadamard product between the weights and the modality predictions

$$\hat{y}_i = \left[\text{weight}\ 1 - \text{weight}\right]\begin{bmatrix}\hat{y}_{ai} \\ \hat{y}_{vi}\end{bmatrix}. \tag{13}$$

To enforce effective learning at both the modality and fusion levels, a composite loss function is defined that supervises the predictions of the audio, visual, and fused outputs

$$\begin{aligned}\mathcal{L}_{\text{total}} = \lambda_1 \cdot \mathcal{L}_{CE}\left(\hat{y}_{ai}, y_i\right) + \lambda_2 \cdot \mathcal{L}_{CE}\left(\hat{y}_{vi}, y_i\right) \\ + \lambda_3 \cdot \mathcal{L}_{CE}\left(\hat{y}_i, y_i\right)\end{aligned} \tag{14}$$

where $\mathcal{L}_{CE}(\cdot,\cdot)$ denotes the standard cross-entropy loss, y_i is the ground truth label, and λ_1, λ_2, and λ_3 are hyperparameters controlling the relative importance of each term.

Moreover, in the training process of multimodal models, high-quality modalities usually generate higher confidence in the training process, indicating that the model predicts the modality more accurately, and then is more inclined to assign more weight to high-quality modalities while ignoring low-quality modalities. To overcome the problem, a dynamic weight adjustment algorithm is proposed, which actively increases the weight of low-quality modes during training to ensure that low-quality modes receive sufficient attention and perform appropriate parameter updates

$$y^{\text{train}} = \begin{bmatrix}p^a \\ p^v\end{bmatrix}\begin{bmatrix}W_a & W_v\end{bmatrix}\begin{bmatrix}\hat{y}_{ai} \\ \hat{y}_{vi}\end{bmatrix} \tag{15}$$

where $\mathbf{p}$ is the weight value of adaptive adjustment in the training process, which dynamically adjusts according to the confidence level of the two types of modes in the training process. It is calculated using the following method:

$$\text{con}_i^m = \sum_{j=1}^{M} 1_{j=y_i} \cdot \text{softmax}\left(f_{\text{cls}}^m\left(\phi_{mi}\left(\theta^m, x_i^m\right)\right)\right)_j. \tag{16}$$

con^m is defined as the total confidence of a specific class predicted by image or audio modality during one training cycle. Based on the different confidence of the two modalities, the imbalance ratio of the modality quality in each training round can be defined as α_t^m

$$\alpha_t^v = \frac{\sum_{i \in \text{cls}_t} \text{con}_i^a}{\sum_{i \in \text{cls}_t} \text{con}_i^v}. \tag{17}$$

The ratio for the audio modality is $\alpha_t^a = 1/\alpha_t^v$. By continuously monitoring the contributions of different modalities, the imbalance can be observed during the training process, and then the gradient descent information can be adaptively adjusted to improve it

$$\begin{cases} p_t^m = \beta \cdot \sigma\left(\alpha_t^m\right) = \dfrac{\beta}{1 + e^{-\alpha_t^m}} \\ \theta_{t+1}^m = \theta_t^m - p_t^m \cdot \eta \tilde{g}\left(\theta_t^m\right). \end{cases} \tag{18}$$

The gradient descent is employed as the primary optimization algorithm, where $\tilde{g}(\theta_t^m)$ represents the gradient corresponding to the modality and p_t^m is an additional parameter used to adjust the backpropagated gradient. The parameter is computed from the imbalance factor of the modality, and its adjustment range is controlled by the sigmoid function $\sigma(x)$, constrained within (0, 1).

TABLE I
WELDING EXPERIMENTAL PARAMETERS

No	Power (kw)	Speed (m/min)	Thickness (mm)	Welding State
#1	7.0	1.2	6.0	Full Pene.
#2	7.0	1.2	10.0	Incomplete Pene.
#3	7.0	1.2	8.0	Hump Defect
#4	6.5	1.2	4.0	Burn Through

Other Params: Shielding Gas Flow: 20L/min, 25L/min for laser power $\geq$ 7kw

TABLE II
DISTRIBUTION OF VISUAL AND ACOUSTIC DATA ACROSS FOUR WELDING STATES

No	Welding State	Images	Audio Duration (s)	Ratio (%)
#1	Complete Penetration	3,287	86.7	25.3
#2	Incomplete Penetration	3,276	87.3	25.2
#3	Humping	3,189	84.1	24.5
#4	Burn Through	3,248	92.9	25.0

Frame Rate: 37 fps, Audio Sampling Rate: 44.1 kHz

III. DATASET AND EXPERIMENT

A. Experimental Setup

The experimental setup of the welding platform is illustrated in Fig. 1, which consists of three main modules: the welding unit, the data acquisition system, and the monitoring system. The welding unit integrates a FANUC robot, a pulsed MIG power source, a WF007A wire feeder, a water-cooling system, and 99.99% pure argon as the shielding gas. The data acquisition module employs an XIRIS camera and an MPA231 microphone, with analog signals digitized by a DAQ card and filtered using an MC3322 signal conditioner. Image data are recorded at 37 frames/s, and audio signals are sampled at 44.1 kHz. The microphone is mounted 35 cm from the torch at a 40° angle, while the XIRIS camera is positioned 30° from the side and 30 cm from the weld pool to minimize interference.

Q1100 high-strength steel is used as the base material. To evaluate the influence of welding parameters on multimodal sensing performance, four typical welding states were designed by varying laser power, welding speed, and material thickness. The representative parameter settings are summarized in Table I. Multiple trials were conducted under different parameter combinations to ensure statistical reliability and data diversity.

B. Dataset

A total of 13 000 synchronized image–audio samples were collected at 37 frames/s and 44.1 kHz. As shown in Table II, the dataset is nearly balanced, with each class providing approximately 3200 images and 84–93 s of audio. The video data were recorded at 37 frames/s, and every 20 consecutive frames were grouped together and combined with the corresponding audio segment from the same time period to form a complete sample pair. To classify the welding states, the dataset was divided into four categories based on the geometric characteristics of the weld seam: incomplete penetration, complete penetration, humping, and burn through. As illustrated in Fig. 6, both image and audio samples were visualized for each welding state. The top portion of the figure displays images capturing the molten pool from a side view, which helps reduce interference from smoke and other visual noise while clearly presenting the pool geometry. In the state of incomplete penetration, the molten pool fusion is not obvious, typically showing the widest transverse width. The humping condition produces a depression on the weld surface, whereas the burn through condition leads to a penetration defect at the weld bottom.

To achieve millimeter-level precision in data labeling, post-welding images were recorded for each process. A ruler was employed to measure the length of each segment corresponding to a specific welding state. To avoid inaccuracies caused by transitions between states, we excluded the 0.5 s before and after each state transition from the dataset. To ensure accurate synchronization and labeling of audio data, a unified sensor control script was developed, allowing for the synchronization of audio and image data via timestamp alignment. The labeling of the audio data was then carried out based on the labeled timestamps from the image data.

The data processing and experiments are performed on a workstation equipped with an NVIDIA[1] GeForce RTX 4080 GPU (16 GB VRAM), running PyTorch 2.1.0 with CUDA 12.4. The proposed dynamic multimodal attention-based weighting (DMAW) model is trained from scratch using the Adam optimizer (learning rate 2×10^{-5}, weight decay 1×10^{-5}), a batch size of 32, and the cross-entropy loss as the objective function.

C. Classification

To validate the superiority of the proposed DMAW model in terms of accuracy and robustness, we conducted comprehensive comparisons against recently proposed unimodal and multimodal approaches. As summarized in Table III, DMAW outperforms all baselines, including unimodal models [7], [8], simple concatenation-based multimodal models [22], late-fusion methods [20], and hybrid fusion frameworks [24]. Specifically, DMAW achieves the highest accuracy (96.1%), precision (96.4%), recall (96.3%), and $F1$-score (96.8%), demonstrating its strong feature extraction and fusion capability.

In multimodal settings, the complete DMAW architecture that integrates both image and audio features further improves overall performance. Experimental results show that DMAW surpasses existing multimodal baselines owing to its adaptive attention-weighting mechanism, which efficiently captures complementary information across modalities. Unlike prior multimodal frameworks relying on simple concatenation or tightly coupled fusion, which may either dilute discriminative cues or increase model complexity, DMAW dynamically adjusts the weight of each modality based on the estimated feature quality, allowing more reliable information to exert

[1]Registered trademark.

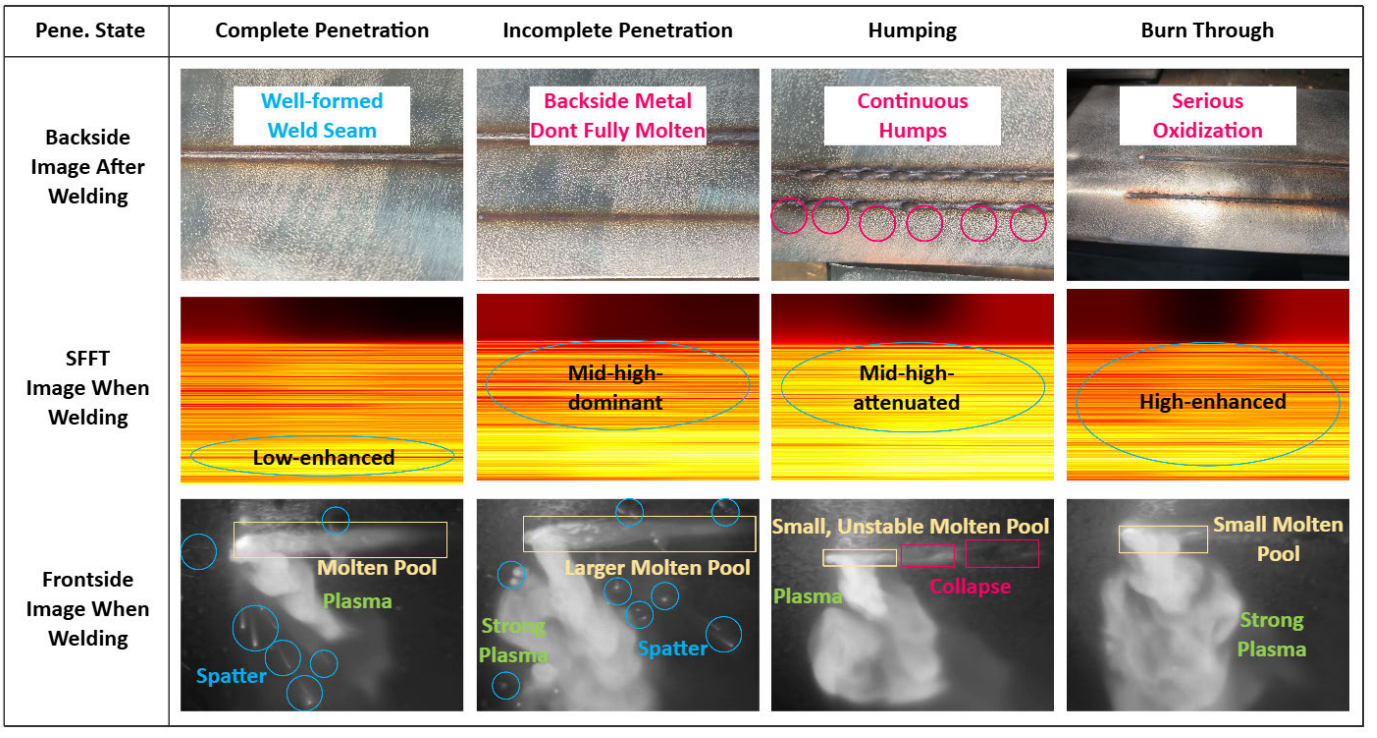


Fig. 6. Multimodal comparison of four welding penetration states—complete penetration, incomplete penetration, humping, and burn through—through backside morphology, SFFT spectrograms, and frontside molten-pool images. Distinct frequency features (low-enhanced, mid–high-dominant, mid–high-attenuated, and high-enhanced) correspond to visible differences in weld formation and plasma behavior.

TABLE III

COMPARISON OF METHODS WITH DIFFERENT MODALITIES. THE BEST RESULT PER COLUMN IS IN BOLD. THE LAST FOUR COLUMNS REPORT PER-CLASS ACCURACIES (%)

Method	Image	Audio	Image +Audio	Accuracy	Precision	Recall	F1 Score	Complete Pene.	Incomplete Pene.	Humping	Burn Through
ResNet [29]	✓	–	–	82.3	82.6	83.8	82.9	78.2	81.7	82.1	83.7
DPWNet [8]	✓	–	–	88.3	86.6	86.8	87.9	85.2	85.7	85.1	88.7
ViTNet [30]	✓	–	–	90.2	91.2	90.9	91.0	84.4	90.5	91.3	91.7
ResNet [29]	–	✓	–	75.6	76.8	77.0	77.5	74.9	75.4	71.8	84.2
TFNet [7]	–	✓	–	77.4	78.3	78.5	79.2	76.3	77.1	73.5	86.9
ViTNet [30]	–	✓	–	84.4	84.1	85.7	84.4	82.2	85.5	80.7	92.6
MRMNet [22]	–	–	✓	83.8	87.7	81.3	83.2	84.2	83.5	82.1	84.7
SMMTNet [20]	–	–	✓	87.1	88.7	86.5	87.4	88.9	86.8	85.4	87.5
AVINet [24]	–	–	✓	91.4	92.5	91.9	92.2	91.1	91.8	90.5	91.7
DMAW$_{fusion}$	✓	✓	✓	**96.1**	**96.4**	**96.3**	**96.8**	**95.1**	**95.8**	**98.3**	**98.6**

greater influence on the final decision. The design enhances both accuracy and generalization, confirming the effectiveness of the proposed approach.

As shown in Fig. 7, the detection accuracy of each category is further analyzed, and the results show that the "burn-through" category achieves the highest accuracy. The trend aligns with the welding image characteristics observed, where an increase in laser power results in shorter molten pool tails and more pronounced burn-through depressions, making the category easier to identify. Additionally, audio features contribute to distinguishing welding states. In the "hump" condition, changes in the audio signal, such as higher frequency components, provide complementary information to the visual cues. The synergy between the audio and visual modalities enhances the model's capacity to distinguish between different welding states, resulting in higher precision. Moreover, we analyzed the parameters and computational complexity of the proposed network, as summarized in Table IV. In addition, inference experiments on an NVIDIA RTX 4080 GPU achieved a latency of 26.2 ms, demonstrating that the model satisfies the real-time requirements of industrial monitoring applications.

D. Unsupervised Encoder

To better understand the role of the unsupervised ViT encoder in welding state recognition, we visualize the attention correlation maps between the classification token (`[CLS]`) and the patch tokens under different heads. This analysis provides insights into how different heads capture complementary aspects of weld pool images across various welding states.

Formally, for a given input sequence $X \in \mathbb{R}^{(N+1)\times d}$ consisting of one classification token and N patch tokens, the

TABLE IV
PARAMETERS AND FLOPS OF THE PROPOSED MULTIMODAL FUSION TRANSFORMER NETWORK

Module	Input/Output	Params	FLOPs
Patch Embedding (Video)	224×224×3 → 196×384	0.30M	0.116G
Patch Embedding (Audio)	224×224×1 → 196×384	0.10M	0.038G
KD_ViT Decoder (Video, 12 layers)	196×384 → 196×384	21.23M	9.04G
KD_ViT Decoder (Audio, 12 layers)	196×384 → 196×384	21.23M	9.04G
Fusion Self-Attn (6 layers)	196×384 → 196×384	10.61M	4.52G
Fusion Cross-Attn (3× bi-dir = 6 single)	196×384 ↔ 196×384	10.62M	4.52G
Dynamic Weight MLP (2 layers)	768 → 256 → 4	0.20M	0.0004G
Classifiers (2 heads)	384 → 4	0.003M	∼0G
Total	–	**64.3M**	**27.27G**

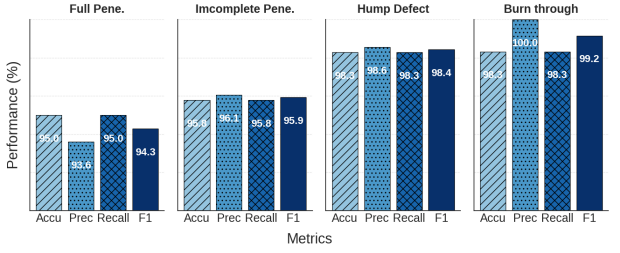


Fig. 7. Performance evaluation across four welding states: complete penetration, incomplete penetration, humping, and burn-through.

query, key, and value projections for head h are defined as follows:

$$Q_h = XW_q^h, \quad K_h = XW_k^h, \quad V_h = XW_v^h. \tag{19}$$

The attention weight matrix is computed via scaled dot-product attention

$$A_h = \text{softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_k}}\right). \tag{20}$$

We extract the correlation between the `[CLS]` token and all other tokens by removing the row corresponding to `[CLS]`, yielding the relevance vector

$$r_h = A_h^{[\text{CLS}],\,1:N}. \tag{21}$$

For each head h, we reshape r_h into a 2-D map aligned with the spatial patch layout, thereby obtaining the attention heatmap. The final averaged heatmap is obtained by thresholding and averaging across heads

$$r_{\text{avg}} = \frac{1}{H}\sum_{h=1}^{H} \mathbb{I}(r_h > \tau) \cdot r_h \tag{22}$$

where H denotes the number of heads and τ is a visualization threshold.

As shown in Fig. 8, different attention heads demonstrate clear preferences in feature extraction. Head1 and Head2 mainly capture the global contour of the weld seam and molten pool, providing a coarse but structurally consistent representation of the welding region. Head3 and Head5 tend to emphasize fine-grained texture features, such as surface ripples or local variations in brightness and roughness. Head4 and Head6 focus more on small and localized details, including spatter regions and subtle irregularities. The diverse attention patterns suggest that each head develops its own specialization, contributing complementary perspectives on the same input.

When the results of multiple heads are aggregated, the averaged heatmap effectively integrates global contours, texture features, and local details, which produces a more comprehensive and stable representation of the weld pool. The observations indicate that the unsupervised ViT encoder, without reliance on labeled supervision, can naturally organize its heads to specialize in different feature patterns. As a result, the extracted representations exhibit stronger robustness and a finer sensitivity to important details, which are essential for reliable welding state characterization.

To assess the effectiveness of the proposed attention-based architecture in multimodal feature extraction, an ablation study was conducted by removing the attention mechanisms from both modalities. To visualize the differences in feature representations, t-distributed stochastic neighbor embedding (t-SNE) was applied to the final-layer feature embeddings. t-SNE is a nonlinear dimensionality reduction technique that preserves the relative distances of high-dimensional data, making it particularly useful for visualizing complex feature distributions. As shown in Fig. 9(a), the embeddings without attention exhibit noticeable overlap among classes, whereas Fig. 9(b) shows compact and well-separated clusters after attention is applied.

The t-SNE results offer valuable insights into the model's ability to separate different features, emphasizing the effectiveness of the module in feature differentiation. For instance, the "burn-through" defect, characterized by distinct visual cues, forms a clearly isolated cluster in both Fig. 9(a) and (b). In contrast, distinguishing between the "complete penetration" and "incomplete penetration" states is more challenging. As shown in Fig. 9(a), removing the attention mechanism leads to significant feature overlap, whereas Fig. 9(b) demonstrates that incorporating attention enables each class to form compact and well-separated clusters. On the other hand, feature overlap suggests a reduced ability to distinguish between classes, ultimately impairing the model's detection accuracy.

To quantitatively evaluate the effect of the attention mechanism on feature separation, we calculate the *Silhouette score* for both conditions. The *Silhouette score* measures how closely an object resembles its own cluster in relation to other clusters. The score is computed as follows:

$$S(i) = \frac{1}{N}\sum_{i=1}^{N} \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{23}$$

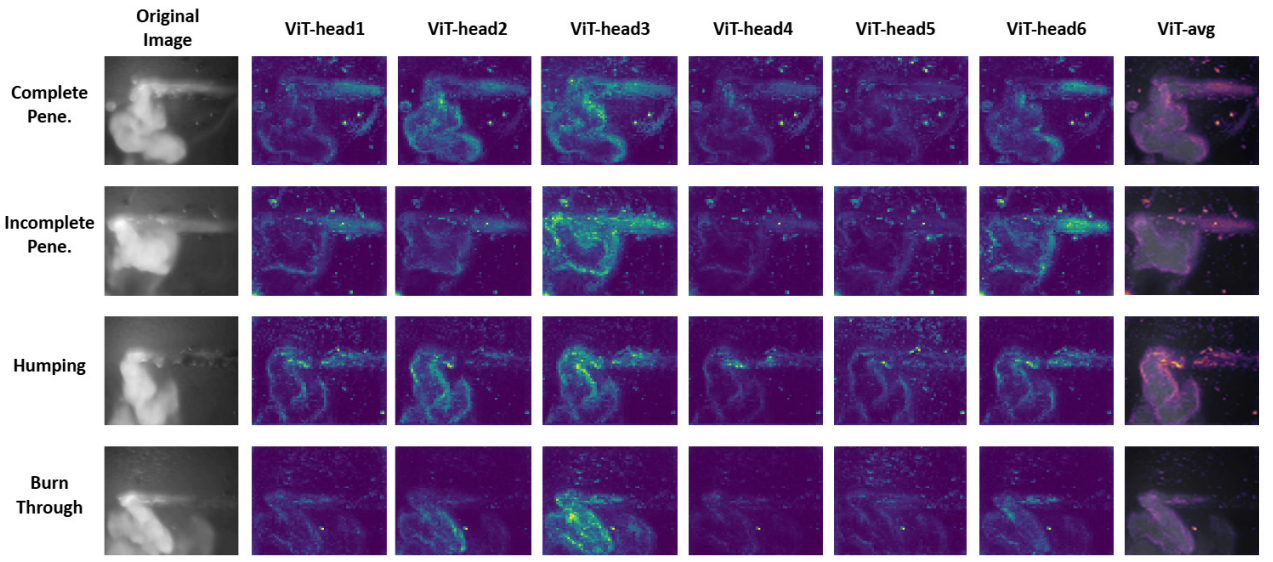


Fig. 8. Visualization of feature maps from different heads and aggregation across welding states.

where $a(i)$ represents the mean distance between point i and all other points within the same cluster and $b(i)$ denotes the mean distance between point i and the closest cluster it does not belong to. A higher Silhouette score reflects better-separated clusters and, consequently, a more effective model. Without the attention mechanism, the Silhouette score is 0.47, reflecting weak clustering performance and significant overlap between classes. However, when the attention mechanism is applied, the Silhouette score increases to 0.52, suggesting improved feature separation and more distinct clusters.

In the context of welding state detection, the attention model plays a crucial role in filtering modality-specific noise and emphasizing features directly relevant to the task. Image data may contain distracting background elements, while audio data may include ambient noise unrelated to the welding process. By selectively enhancing relevant features, such as dynamic molten pool characteristics in images or acoustic patterns indicative of welding conditions, the attention mechanism improves feature separability, as reflected in the t-SNE visualization. The enhanced feature discrimination contributes to more robust classification performance, demonstrating the necessity of an attention model in multimodal learning.

E. Dynamic Fusion

To quantify the performance impact of our adaptive fusion mechanism, we compared it with several baseline fusion methods, such as FiLM [32], Gate [31], and Gradient Control [33], to analyze how each fusion method better integrates information from different modalities and improves overall model performance.

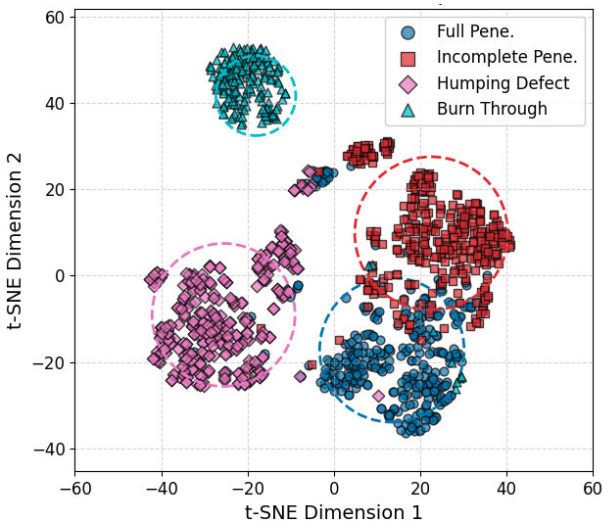
Most general multimodal fusion methods assume that both modalities provide high-quality information and are designed to enhance overall fusion accuracy, often assuming that the quality of both modalities remains constant. However, in welding scenarios, the assumption does not hold. Welding-specific challenges, such as noise interference, can cause one modality to be overwhelmed by noise, particularly the audio or visual streams, leading to a degradation in model accuracy. As shown in Table V, the results underscore the advantages of dynamically adjusting modality contributions. Unlike static methods that treat all modalities equally, dynamic weight fusion provides a flexible, adaptive approach that ensures the model focuses on the most reliable modality, particularly in complex environments such as welding state detection.

TABLE V
COMPARATIVE PERFORMANCE OF VARIOUS FUSION STRATEGIES FOR THE PROPOSED DMAW FRAMEWORK ON WELDING STATE RECOGNITION

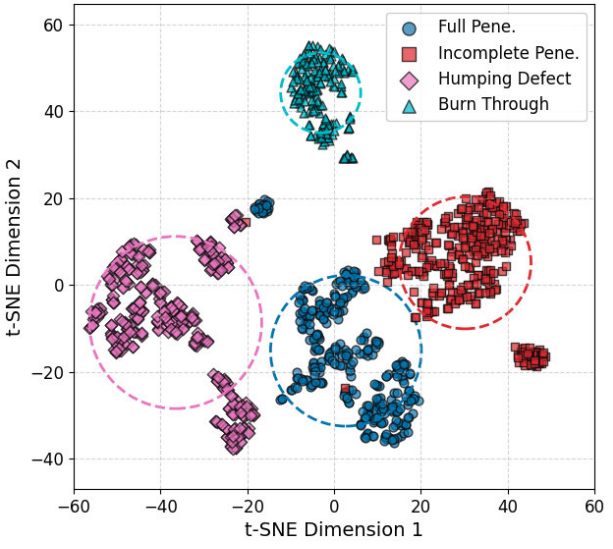
Fusion Method	Accuracy	Precision	Recall	F1 Score
Feature Sum	90.5	91.7	91.7	91.6
Feature Gate [31]	92.6	93.2	92.9	93.0
Feature FiLM [32]	93.7	94.0	93.8	93.8
Gradient Control [33]	93.4	94.5	94.3	94.3
Dynamic Fusion	**96.1**	**96.4**	**96.3**	**96.8**

To further validate the effectiveness of the dynamic balancing (DB) training strategy, we conducted an ablation study that tracks validation accuracy throughout training and compares it with a baseline model trained without this strategy. As illustrated in Fig. 10, the lighter curves depict the raw validation accuracy recorded at each epoch, capturing the instantaneous fluctuations of the model, whereas the darker curves represent a 7-epoch moving average that highlights the underlying convergence trend.

Notably, the DB method exhibits occasional decreases in validation accuracy relative to the baseline at certain epochs. These transient fluctuations are consistent with the strategy's ability to encourage exploration and mitigate the risk of entrapment in shallow local optima. When the results are examined using a 7-epoch moving average, the DB curve consistently surpasses the baseline by approximately 0.4%–0.8% and ultimately converges to a validation accuracy of about 96%. The consistent improvement demonstrates that the proposed DB strategy not only alleviates premature



(a)



(b)

Fig. 9. t-SNE visualization of feature embeddings from the DMAW framework. (a) Without attention modules, the feature clusters show noticeable overlap. (b) With attention enabled, the clusters become compact and well separated, highlighting improved interclass discriminability.

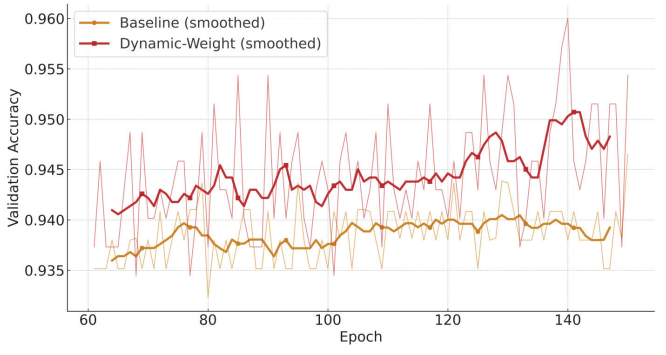


Fig. 10. Influence of dynamic weight training strategies on the accuracy of the validation set during the training process.

convergence to suboptimal solutions but also achieves a higher overall level of performance.

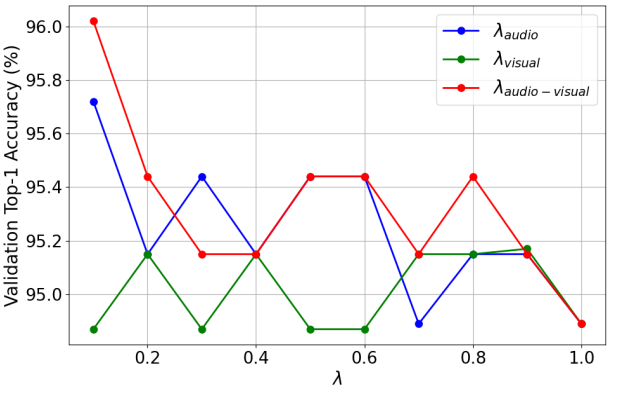


Fig. 11. Sensitivity analysis of hyperparameters λ on validation accuracy.

Besides, to assess the impact of the hyperparameters λ_{audio}, λ_{visual}, and $\lambda_{\text{audio-visual}}$ on model performance, we conducted a control variable analysis. The initial values of all hyperparameters were set to 1. Each hyperparameter was then varied from 0.1 to 1.0 in increments of 0.1, while keeping the other two parameters fixed at 1.0. As shown in Fig. 11, the results demonstrate the influence of each hyperparameter on validation accuracy, supporting the rationale for their final selection. The result reveals that the impact of the λ hyperparameters on validation accuracy is relatively minimal, with performance fluctuations within a narrow range. Based on this analysis, we selected the following λ values: $\lambda_{\text{audio}} = 0.1$, $\lambda_{\text{visual}} = 0.9$, and $\lambda_{\text{audio-visual}} = 0.1$.

IV. Conclusion

In this article, a dynamic multimodal fusion method is proposed that integrates image and audio data from the welding process to address challenges posed by complex industrial environments in welding state detection. A key challenge in multimodal fusion is noise interference across modalities. To address the problem, a multimodal attention model is introduced, which effectively extracts key features and is validated through t-SNE visualization. Additionally, to handle modality quality variations, a feature-based dynamic weight adjustment algorithm is proposed that ensures optimal feature learning and improved robustness. Comparative experiments demonstrate that our multimodal fusion approach achieves superior detection accuracy in both single-modal and multimodal models, validating its effectiveness and reliability for industrial welding state detection under challenging conditions.

References

[1] Y. He, W. Li, Y. Zhang, K. Xu, H. Wan, and Z. Chen, "A data-driven multiscale convolutional adaptive network for welding robot operating state recognition," *IEEE Sensors J.*, vol. 25, no. 3, pp. 5231–5240, Feb. 2025.

[2] R. Xiao, Y. Xu, F. Xu, Z. Hou, H. Zhang, and S. Chen, "LSFP-tracker: An autonomous laser stripe feature point extraction algorithm based on Siamese network for robotic welding seam tracking," *IEEE Trans. Ind. Electron.*, vol. 71, no. 1, pp. 1037–1048, Jan. 2024.

[3] J. Zhang, J. Pan, H. Qu, L. Li, and S. Yang, "Deep learning-based crack detection in spiral-welded pipelines: A novel adaptive convolutional feature extraction method," *IEEE Sensors J.*, vol. 25, no. 12, pp. 22895–22906, Jun. 2025.

[4] Z. Zhao, N. Lv, R. Xiao, and S. Chen, "A novel penetration state recognition method based on LSTM with auditory attention during pulsed GTAW," *IEEE Trans. Ind. Informat.*, vol. 19, no. 9, pp. 9565–9575, Sep. 2023.

[5] W. Yan et al., "Multisource multimodal feature fusion for small leak detection in gas pipelines," *IEEE Sensors J.*, vol. 24, no. 2, pp. 1857–1865, Jan. 2024.

[6] H. Xu, R. Lv, and B. Zi, "A multimodal information fusion network combining CNN and adaptive mamba for fault monitoring of selective laser melting ceramic printing instrument," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–16, 2025.

[7] W. Ren, G. Wen, B. Xu, and Z. Zhang, "A novel convolutional neural network based on time–frequency spectrogram of arc sound and its application on GTAW penetration classification," *IEEE Trans. Ind. Informat.*, vol. 17, no. 2, pp. 809–819, Feb. 2021.

[8] Y. Feng, Z. Chen, D. Wang, J. Chen, and Z. Feng, "DeepWelding: A deep learning enhanced approach to GTAW using multisource sensing images," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 465–474, Jan. 2020.

[9] J. Lee, I. Noh, J. Lee, and S. W. Lee, "Development of an explainable fault diagnosis framework based on sensor data imagification: A case study of the robotic spot-welding process," *IEEE Trans. Ind. Informat.*, vol. 18, no. 10, pp. 6895–6904, Oct. 2022.

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.

[11] X. Zhang, Y. Ma, D. Chang, and J. Du, "SS-FDPNet: A self-supervised denoising network for radiographic images of ship welds," *IEEE Sensors J.*, vol. 25, no. 8, pp. 14235–14251, Apr. 2025.

[12] Q. Wan, L. Gao, X. Li, and L. Wen, "Unsupervised image anomaly detection and segmentation based on pretrained feature mapping," *IEEE Trans. Ind. Informat.*, vol. 19, no. 3, pp. 2330–2339, Mar. 2023.

[13] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9650–9660.

[14] W. Cui et al., "A rapid screening method for suspected defects in steel pipe welds by combining correspondence mechanism and normalizing flow," *IEEE Trans. Ind. Informat.*, vol. 20, no. 9, pp. 11171–11180, Sep. 2024.

[15] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[16] Y. Ma et al., "An efficient and robust complex weld seam feature point extraction method for seam tracking and posture adjustment," *IEEE Trans. Ind. Informat.*, vol. 19, no. 11, pp. 10704–10715, Nov. 2023.

[17] T. Zhan, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 10078–10093.

[18] H. Wang, X. Wu, Z. Huang, and E. P. Xing, "High-frequency component helps explain the generalization of convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8684–8694.

[19] W. Huang, A. Han, Y. Chen, Y. Cao, Z. Xu, and T. Suzuki, "On the comparison between multi-modal and single-modal contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 81549–81605.

[20] L. Guo et al., "Transformer-based spiking neural networks for multimodal audiovisual classification," *IEEE Trans. Cognit. Develop. Syst.*, vol. 16, no. 3, pp. 1077–1086, Jun. 2024.

[21] S. Mo and P. Morgado, "Unveiling the power of audio-visual early fusion transformers with dense interactions through masked modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 27186–27196.

[22] B. Shen et al., "Multimodal-based weld reinforcement monitoring system for wire arc additive manufacturing," *J. Mater. Res. Technol.*, vol. 20, pp. 561–571, Sep. 2022.

[23] Z. Yuan, Q. Shen, B. Zheng, Y. Liu, L. Jiang, and G. Guo, "Video and audio are images: A cross-modal mixer for original data on video–audio retrieval," *Knowl.-Based Syst.*, vol. 299, Sep. 2024, Art. no. 112076.

[24] W. Pian, S. Mo, Y. Guo, and Y. Tian, "Audio-visual class-incremental learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 7799–7811.

[25] Q. Zhang et al., "Provable dynamic fusion for low-quality multimodal data," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 41753–41769.

[26] H. Chi, Y. Wei, B. Yuan, X. Wang, Q. Sun, and L. Shu, "Bearing health state representation and fault identification based on multisource sensor signal feature information fusion," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–16, 2025.

[27] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," 2023, *arXiv:2304.07193*.

[28] O. Siméoni et al., "DINOv3," 2025, *arXiv:2508.10104*.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[30] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

[31] D. Kiela, É. Grave, A. Joulin, and T. Mikolov, "Efficient large-scale multi-modal classification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018, vol. 32, no. 1, pp. 5198–5204.

[32] E. Perez, F. Strub, H. D. Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018, vol. 32, no. 1, pp. 3942–3951.

[33] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8228–8237.

Jiawei Fan received the B.S. degree from Hangzhou Dianzi University, Hangzhou, China, in 2023. He is currently pursuing the M.S. degree with the School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai, China.

His research interests include multimodal fusion, machine vision, and industrial defect detection.

Ting Yuan received the B.S. and M.S. degrees in electrical engineering from the University of Electronic Science and Technology of China, Hefei, China, in 2003 and 2006, respectively, and the Ph.D. degree from the University of Connecticut, Storrs, CT, USA, in 2013.

He is currently a Professor at the School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai, China. His research focuses on detection, classification, and tracking using information from camera, radar, and lidar sensors, as well as data fusion for multisensor systems.

Haonan Zhang received the B.S. degree from Wuhan University of Technology, Wuhan, China, in 2023. He is currently pursuing the Ph.D. degree with the Institute of Welding and Laser Manufacturing, School of Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai, China.

His research interests include vision-based laser welding defect monitoring, penetration estimation, and seam tracking.

Songlin Li received the B.S. degree in electronic information engineering from Beijing Institute of Technology, Beijing, China, in 2022. He is currently pursuing the M.S. degree with the School of Integrated Circuits, Shanghai Jiao Tong University, Shanghai, China.

His research focuses on intelligent welding defect monitoring and online inspection for laser welding processes.

Uwe D. Hanebeck (Fellow, IEEE) received the Ph.D. and Habilitation degrees in electrical engineering from the Technical University of Munich, Munich, Germany, in 1997 and 2003, respectively.

He is currently a Chaired Professor of Computer Science at Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, and the Director of the Intelligent Sensor-Actuator Systems Laboratory (ISAS), Karlsruhe. His research interests include information fusion, nonlinear state estimation, stochastic modeling, and system identification.

Zhuguo Li received the B.S. and M.S. degrees from Shanghai Jiao Tong University, Shanghai, China, and the Ph.D. degree from Osaka University, Japan, in 1997 and 2004, respectively.

He is currently a Distinguished Professor with the School of Materials Science and Engineering, Shanghai Jiao Tong University, and the Director of Shanghai Key Laboratory of Laser Manufacturing and Materials Modification. His research focuses on advanced laser manufacturing, process control, and welding defect characterization.

Edmond Q. Wu (Senior Member, IEEE) received the Ph.D. degree in control theory and engineering from Southeast University, Nanjing, China, in 2009.

He is a Professor with the Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai Jiao Tong University, Shanghai, China. His research interests include deep learning, fatigue recognition, and human–machine interaction.

Dr. Wu serves as an Associate Editor for IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS and IEEE TRANSACTIONS ON INTELLIGENT VEHICLES.

Jiuchao Qian received the B.S. degree from Shandong University, Jinan, China, in 2008, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2016.

He is currently an Assistant Professor at the School of Automation and Intelligent Sensing, Shanghai Jiao Tong University. His research interests include smart manufacturing, video analysis, deep learning, signal processing, and ubiquitous computing.