

Thermal-history and morphology-conditioned multimodal fusion for robotic MAG weld-quality classification

Bin Zhong

Shanghai Jiao Tong University School of Materials Science and Engineering

Kanghong Zhu

Shanghai Jiao Tong University School of Materials Science and Engineering

Runquan Xiao

ITG Institute of new quality productive forces, Xiamen CCRE Group Co., Ltd, Xiamen

Yiyang Zhao

Shanghai Jiao Tong University School of Materials Science and Engineering <https://orcid.org/0000-0003-1221-1414>

Yingcai Wan

Shanghai Jiao Tong University School of Materials Science and Engineering

Huabin Chen

hbchen@sjtu.edu.cn

Shanghai Jiao Tong University School of Materials Science and Engineering <https://orcid.org/0000-0002-0962-6232>

Research Article

Keywords: robotic MAG welding, weld quality classification, nominal thermal history, molten-pool morphology, multimodal fusion

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11218666/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Thermal-history and morphology-conditioned multimodal fusion for robotic MAG weld-quality classification

Bin Zhong¹, Kanghong Zhu¹, Runquan Xiao², Yiyang Zhao^{1*},
Yingcai Wan¹, Huabin Chen^{1*}

¹School of Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China.

²ITG Institute of New Quality Productive Forces, Xiamen CCRE Group Co., Ltd, Xiamen, 361004, China.

*Corresponding author(s). E-mail(s): hbchen@sjtu.edu.cn (Huabin Chen, primary contact); 804311109@sjtu.edu.cn (Yiyang Zhao)

Abstract

In robotic metal active gas (MAG) welding, weld formation reflects both current molten-pool behaviour and preceding thermal history. Short-window multimodal sensing captures local process dynamics but does not explicitly describe the accumulated thermal response. We propose a thermal-history and morphology-conditioned fusion network (HMCF-Net) for four-class weld-quality classification. The network combines global-local visual features from every recorded camera frame in a 200-ms window with electrical and acoustic features. A finite-plate calculation driven by recorded power provides a six-dimensional nominal thermal-history descriptor, including the period before the window. An eight-dimensional descriptor expresses pool geometry through fixed image-space definitions using masks from the same images. Both descriptors jointly condition a nonlinear residual update, with a bypass retaining the observation representation. The dataset was split at the weld level into 26 training records and 10 validation records. On 5,078 validation windows, HMCF-Net achieved 94.72% accuracy, 93.04% balanced accuracy and 93.46% macro-F1. Joint conditioning improved macro-F1 by 1.69 percentage points over a matched model using the same multimodal observation encoder without the two descriptors. It also outperformed variants using either descriptor alone. These results show that jointly conditioning multimodal observations on calculated thermal history and explicit pool geometry

improves overall weld-quality classification performance without requiring direct temperature measurement or a longer camera observation window.

Keywords: robotic MAG welding, weld quality classification, nominal thermal history, molten-pool morphology, multimodal fusion

1 Introduction

In-process weld-quality recognition supports online assessment of robotic metal active gas (MAG) welding by distinguishing normal welding, penetration defects and weld deviation. Multi-source studies in robotic MAG and adaptive gas metal arc welding (GMAW) have demonstrated the value of sensing for quality evaluation and control [1, 2]. However, actual welding behaviour can differ from nominal settings [3], and the relationship between measured signals and penetration state depends on process conditions [4]. Joint gap, groove geometry and molten-pool appearance further influence the relationship between process observations and the resulting bead [5, 6]. Reliable recognition therefore involves interpreting local sensor observations in the context of the conditions under which a weld is formed.

Molten-pool images reveal visible geometry and position, whereas electrical and acoustic signals capture complementary arc responses. Image-voltage fusion has been used to recognise burn-through and weld deviation in metal inert gas lap welding [7]. Acoustic-electrical features have supported state and penetration recognition in cold metal transfer and keyhole tungsten inert gas welding [8, 9]. Multisensor systems have combined image, current and temperature measurements for laser-MAG monitoring [10], or image, acoustic and electrical sequences for robotic gas tungsten arc welding [11]. These studies establish the utility of multimodal sensing. However, a short observation window does not itself include the recorded power history preceding that window.

Recent fusion models represent interactions and dynamics among observed welding signals. Attention-based methods have addressed welding-state recognition, penetration monitoring and bead-width prediction [12–14]. Transfer learning and image-based statistical monitoring have addressed cross-condition recognition and additive-manufacturing anomalies [15, 16]. Sequence models and edge-history representations exploit molten-pool motion and process-signal evolution [17–22]. Their temporal information is drawn primarily from the observed modalities. By contrast, a calculated response driven by recorded power can describe the process history preceding the current observation without extending the camera window.

Thermal context can be obtained by measurement or calculation. Infrared-visible fusion combines observed thermal radiation with surface appearance [23]. Alternatively, a time-varying double-ellipsoidal source in an analytical finite-plate model yields a nominal response governed by input power, source motion and plate boundaries [24]. Calculated process quantities and physical constraints have also informed learning for porosity, penetration and bead geometry [25–28]. Incorporating such a calculated

history into multimodal quality recognition would make preceding power input available alongside current observations, without treating the response as a measured temperature field.

Learned visual features represent pool appearance and spatial context, whereas descriptors derived from masks of the same images express pool extent, position and dispersion through fixed image-space definitions. Such descriptors make these quantities explicitly available alongside learned features. Pool contours have supported robotic MAG morphology classification [29], while pool width and leading-edge position have informed adaptive GMAW control [2]. These findings motivate combining explicit geometric descriptors with learned visual features and assessing their added value for quality recognition.

The remaining challenge is to integrate accumulated thermal history and explicit pool geometry with short-window multimodal observations in a unified recognition model. A conditioning mechanism must connect these descriptors to the learned observation representation, while matched comparisons establish their individual and combined contributions.

To address this challenge, we present a thermal-history and morphology-conditioned fusion network (HMCF-Net) for four-class quality recognition in robotic MAG welding. The network uses nominal thermal history and explicit molten-pool morphology to jointly condition a residual update of the multimodal observation representation. The update is computed from both the joint condition and the current observation state through a nonlinear mapping, while a bypass retains the original representation. The six-dimensional thermal descriptor summarises a finite-plate response driven by recorded power, while the eight-dimensional morphology descriptor quantifies visible pool geometry. The observation encoder retains every recorded camera frame in a 200-ms window and couples global–local visual features with temporally available electrical and acoustic features. Matched ablations quantify the individual and joint contributions of the descriptors with the observation encoder retained. Comparisons with alternative temporal and fusion architectures, together with class- and weld-level analyses, establish the performance and scope of the approach.

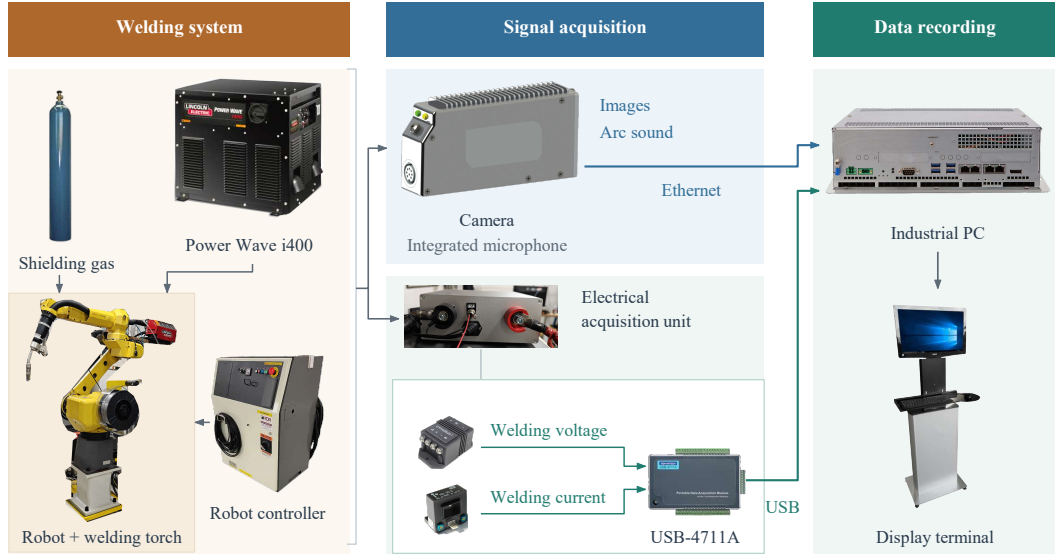
2 Experimental setup and dataset

2.1 Welding setup and process conditions

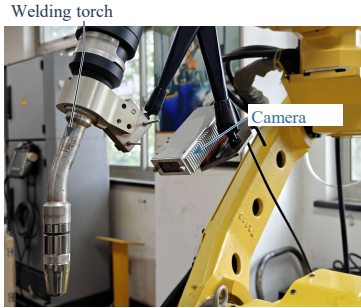
We investigated single-pass robotic MAG welding of Q235 steel V-groove butt joints under varied process conditions. Molten-pool images, electrical signals and arc sound were linked to post-weld quality labels for each weld record.

The welding platform comprised a FANUC M-10iA/12 robot, a Lincoln Electric Power Wave i400 welding power source and multimodal acquisition equipment (Fig. 1). Each joint consisted of two Q235 steel plates measuring $300 \times 150 \times D$ mm, which formed a 60° V-shaped butt groove with 30° on each side and no root face. The base material, welding wire, shielding gas and process conditions are summarised in Table 1. The dataset comprised 36 weld records spanning variations in current and voltage setpoints, welding speed, plate thickness and assembly gap, with both weaving and non-weaving conditions represented.

(a) Experimental system



(b) Sensor mounting



(c) Butt-joint geometry

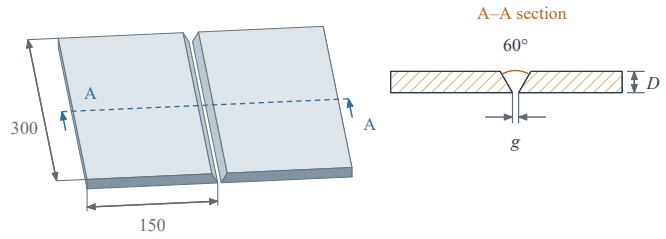


Fig. 1 Experimental setup and butt-joint geometry: (a) welding and data-acquisition system, (b) torch-mounted camera, and (c) plate dimensions and V-groove section

Table 1 Welding materials, joint geometry and process conditions

Item	Condition or range
Base material and joint	Q235 steel; 60° V-shaped butt groove; no root face
Welding task	Single-pass welding, targeting full penetration
Welding wire	1.2-mm steel solid wire
Shielding gas	80% Ar + 20% CO ₂
Current setpoint	120–300 A
Voltage setpoint	16–30 V
Welding speed	11–42 cm/min
Plate thickness	12, 16, 20 and 25 mm
Recorded assembly gap	0–4 mm

2.2 Multimodal data acquisition and preprocessing

Visual, acoustic and electrical measurements captured the visible molten-pool appearance and corresponding process responses. An 8-bit grayscale molten-pool camera was installed approximately 45° obliquely above and behind the welding torch to view the rear part of the molten pool. Its raw frames had a resolution of 640×480 pixels. A fixed 400-pixel-wide region was cropped horizontally and resized to 320×384 pixels for visual encoding. Morphology descriptors were calculated in the original-image coordinates of this fixed context.

Current and voltage were acquired using an Advantech USB-4711A at a sampling rate of 10 kHz. Arc sound was processed at a sampling rate of 36,621 Hz. The multimodal data were temporally aligned using recorded timestamps.

Current and voltage followed two processing paths. The observation encoder used training-standardised signed acquisition-count channels to represent short-term process fluctuations. Calibrated conversion coefficients provided current and voltage in amperes and volts for constructing the nominal power history. Conversion details and the finite-plate response are given in Section 3.3.

2.3 Weld-quality labelling

Quality labels were assigned from post-weld visual inspection at the level of the weld record. Each record was classified as normal, incomplete penetration, burn-through or weld deviation. Normal welding was identified by a continuous root-side bead; incomplete penetration mainly by a visible joint line or discontinuous root-side bead; burn-through by local perforation; and weld deviation by displacement of the face-side bead relative to the groove region. Figure 2 presents representative process images and post-weld appearances of these classes. The process images define model inputs, whereas post-weld appearances provide the visual basis for the quality labels. The frames in this figure were gamma-adjusted to 0.60 for display only.

Each observation window inherited the quality label of its parent weld record, thereby defining a supervised mapping from short-term multimodal observations to weld-quality classes. The labels were assigned to weld records rather than to defect locations along the weld path. Accordingly, the evaluation in Section 4 combines window-level classification metrics with the distribution of predictions within each weld record to describe both overall recognition and weld-to-weld variation.

2.4 Window construction and dataset partitioning

To reduce the influence of start and end transients, a 0.5-s buffer was imposed at both ends of the retained main-welding interval. Using an original camera frame as the temporal anchor, a candidate window was constructed only when the preceding 200 ms lay fully within the retained interval, the required images were present, and the acoustic and electrical signals were complete and valid. The 0.5-s buffer was applied uniformly to the interval selection.

Each window contained all real images in the preceding 200 ms up to the anchor frame, together with the corresponding acoustic and electrical signals (Fig. 3). The images retained their original order; the sequence typically contained 6–7 frames.

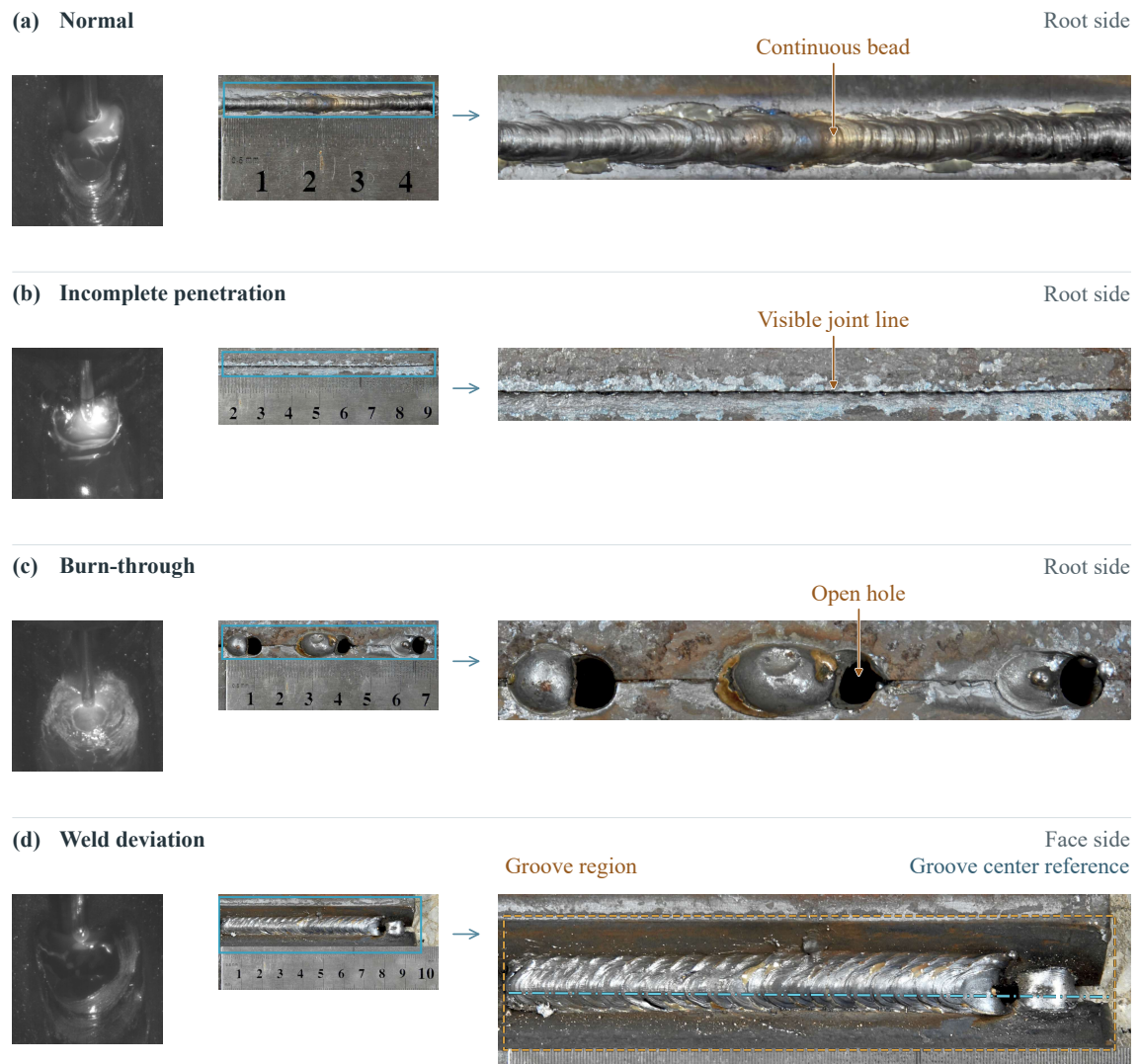


Fig. 2 Representative process frames, post-weld views and magnified regions for (a) normal welding, (b) incomplete penetration, (c) burn-through and (d) weld deviation. Post-weld views show the root side except in (d), which shows the face side and a qualitative groove-centre reference.

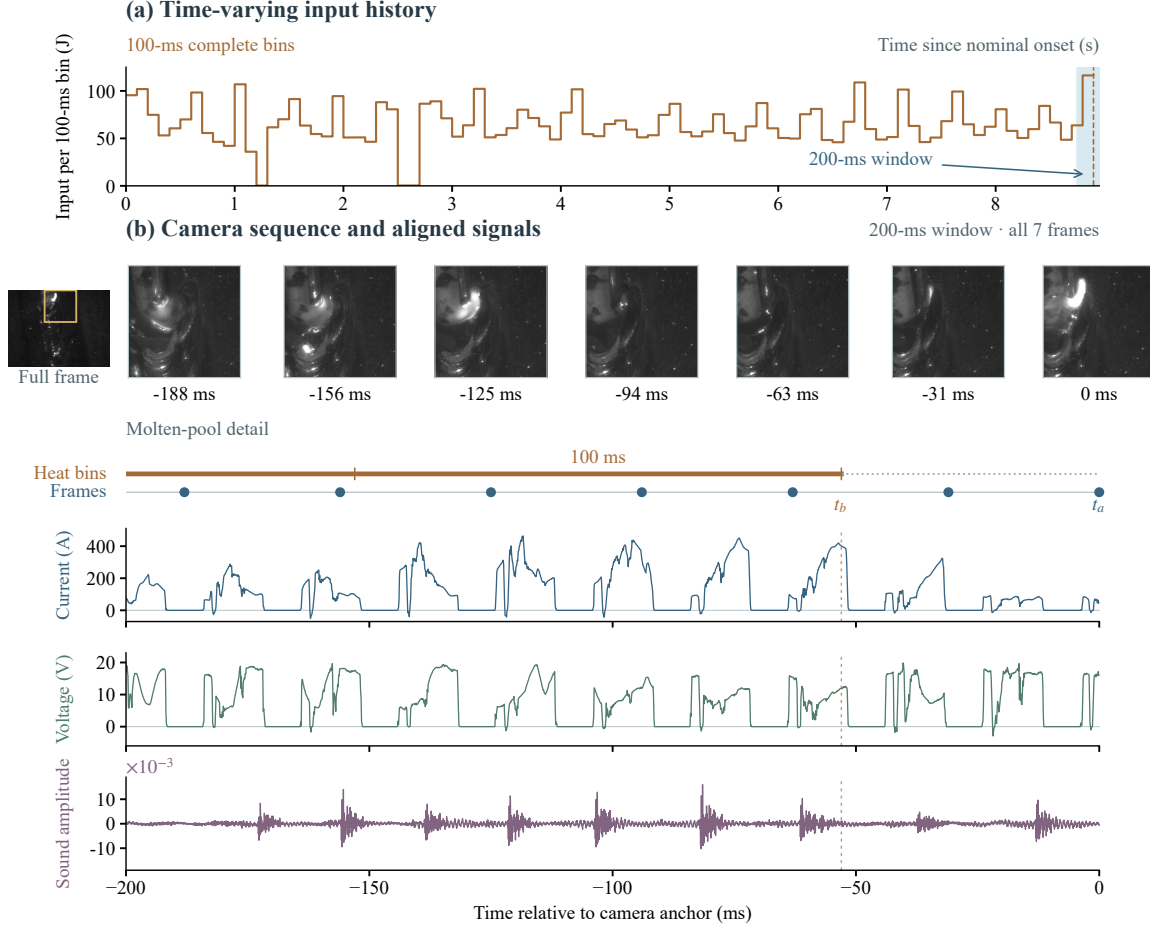


Fig. 3 The 200-ms observation window and preceding input history. (a) Energy in complete 100-ms bins from nominal arc onset. (b) Seven recorded frames with aligned current, voltage and arc sound. The latest complete heat bin ends at t_b , before the camera anchor t_a

Unlike the short-term observation window, the thermal history was accumulated from the nominal starting point of the corresponding record to the current time and was not reinitialised for each window. Thus, local process observations and accumulated thermal history represented separate time scales.

We obtained 19,946 valid windows. The training set contained 26 weld records and 14,868 windows, whereas the validation set contained 10 weld records and 5,078 windows (Table 2). All comparison methods used the same weld-level split and window collections.

Because neighbouring windows within a weld record overlapped, the weld record was used as the partitioning unit and all windows from the same record remained in one split.

Table 2 Distribution of weld records and windows by quality state

Quality state	Training welds	Training windows	Validation welds	Validation windows
Normal	6	4,454	2	1,820
Incomplete penetration	10	5,826	2	1,403
Burn-through	8	2,720	4	609
Weld deviation	2	1,868	2	1,246
Total	26	14,868	10	5,078

Data-driven normalisation statistics were estimated from the training windows only. Valid observations were first aggregated within each record and then combined with equal weight across records; the resulting statistics were fixed when processing the validation set. The visual input used fixed image-normalisation constants. Training and evaluation are described in Section 3.6.

3 Methodology

3.1 HMCF-Net architecture

HMCF-Net combines a short-window visual–acoustic–electrical observation with two complementary condition descriptors: nominal thermal history and explicit molten-pool morphology. A conditioned residual mapping integrates these descriptions within a single classification path (Fig. 4).

For a camera anchor time t_a , HMCF-Net receives every real image and the corresponding current, voltage and arc-sound sequences from the preceding 200 ms. Global–local visual features first interact with temporally available signal features. The resulting sequence is then aggregated in camera-frame order into a 96-dimensional window observation $\mathbf{z}$. The thermal-history branch converts the time-varying input accumulated from nominal arc onset into a six-dimensional descriptor $\mathbf{H}$. The morphology branch aggregates geometric quantities from valid segmentations into an eight-dimensional descriptor $\mathbf{g}$. These descriptors are embedded separately, summed as a joint condition, and used to update $\mathbf{z}$ before classification into normal, incomplete penetration, burn-through and weld deviation.

3.2 Multimodal observation encoding

The observation encoder first couples global–local visual features to acoustic and electrical features at each camera frame, then aggregates the real-frame sequence over the window (Fig. 5).

3.2.1 Global and local visual representations

Visual recognition requires both the molten pool’s position relative to its surroundings and details within and near its boundary. We therefore used global and local visual

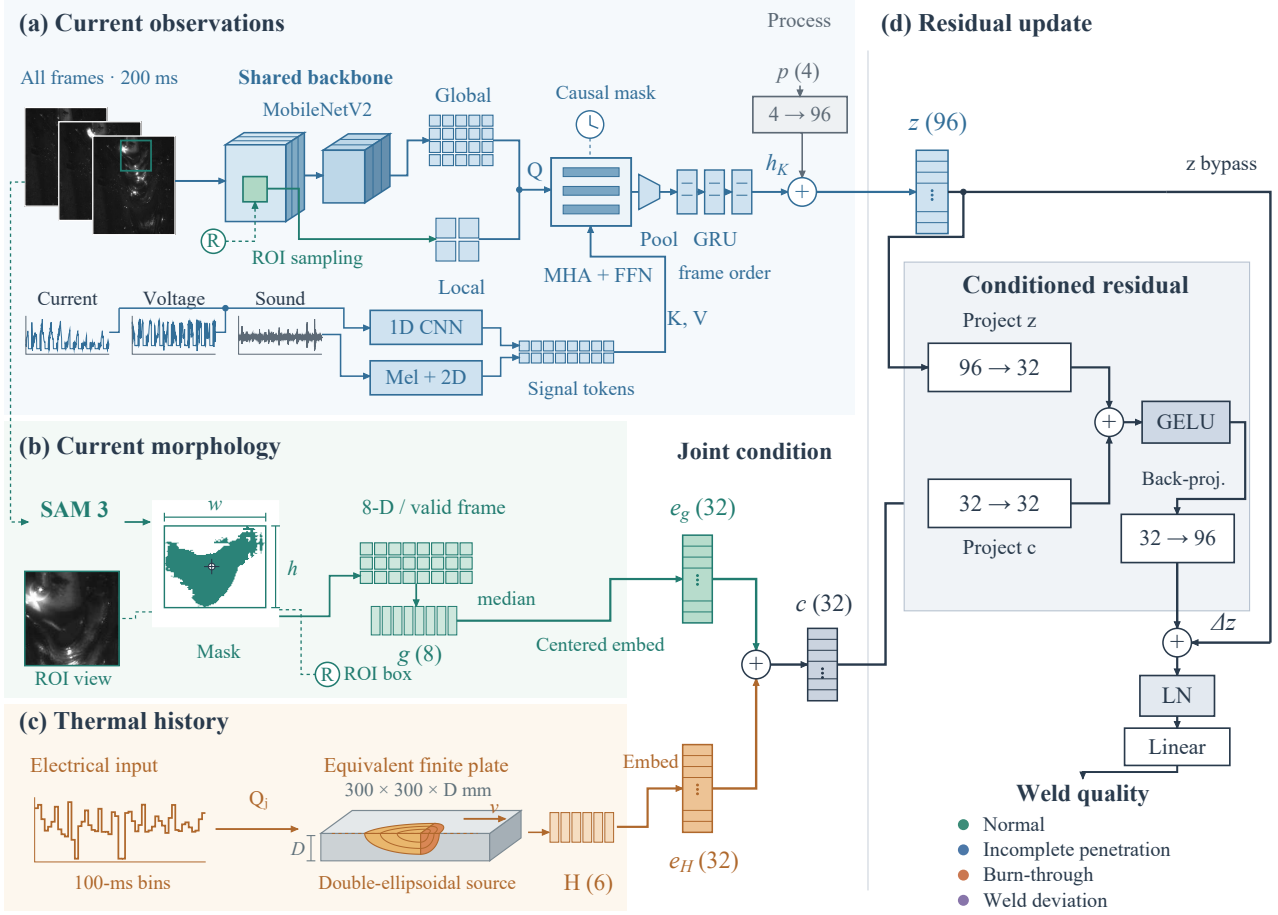
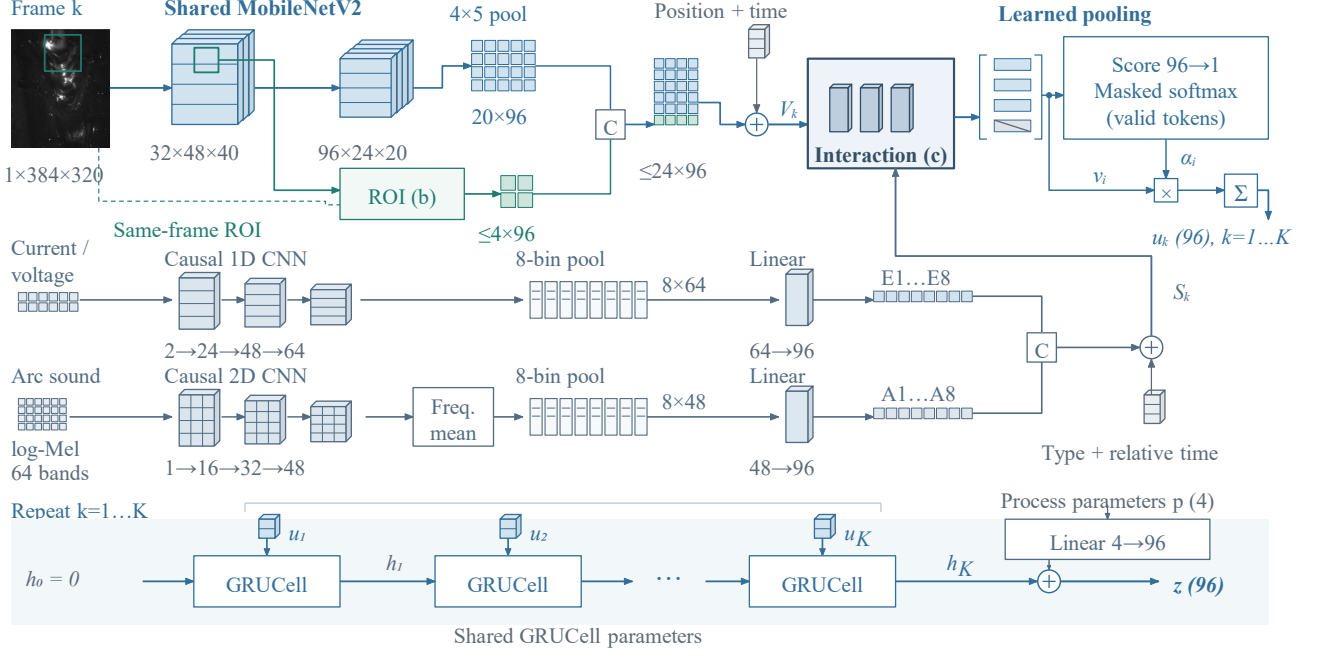
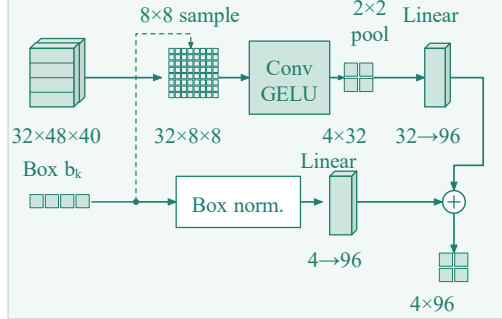


Fig. 4 HMCF-Net architecture: (a) multimodal observation encoding, (b) molten-pool morphology, (c) finite-plate thermal history, and (d) conditioned residual classification

(a) Shared encoding and framewise aggregation



(b) Shared-feature ROI encoding



(c) Time-constrained cross-modal interaction

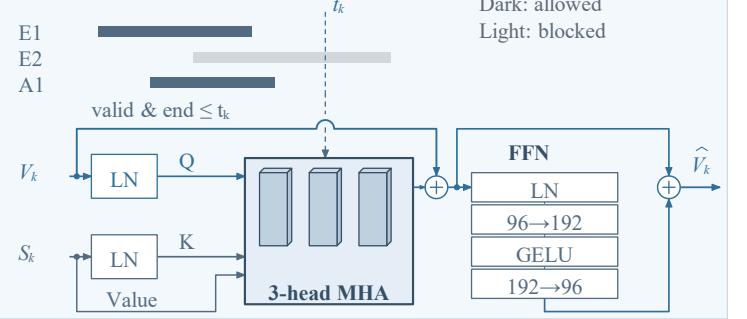


Fig. 5 Global-local encoding and time-constrained fusion: (a) shared visual and signal encoders with temporal aggregation, (b) same-frame local feature sampling, and (c) the causal validity mask for cross-modal attention

paths on a shared MobileNetV2 backbone [30]. Each cropped grayscale image was replicated to three channels and processed with fixed image normalisation through the truncated backbone. The backbone produced a 32-channel intermediate feature map with 48×40 spatial positions and a 96-channel deeper map with 24×20 positions. Adaptive average pooling with a 4×5 output grid converted the deep feature map into 20 global vectors of dimension 96.

The local path used a same-frame Segment Anything Model 3 (SAM 3) molten-pool segmentation [31] to define the region of interest. On the intermediate feature map, 8×8 bilinear grid sampling was followed by convolution, Gaussian error linear unit (GELU) activation, 2×2 adaptive average pooling and linear projection. This path produced 4 local vectors of dimension 96. A linear embedding of the normalised region-box coordinates was added to these vectors. Each frame therefore contained 20 global visual tokens and up to 4 local visual tokens. When the local region was unavailable, the corresponding tokens were masked while the global path remained active.

The local sampling box was obtained by expanding the same-frame mask bounding box into a square. Its side length was 1.5 times the longer side of the bounding box. The expanded box was checked against the context boundary and the supporting feature grid. The two visual paths used different feature levels of the shared backbone. Both paths produced a common token width for cross-modal interaction. These local tokens form the learned visual representation; the explicit geometric descriptor is introduced separately in Section 3.4.

3.2.2 Time-constrained visual–acoustic–electrical interaction

The electrical signals were encoded in the time domain, whereas arc sound was encoded in the time–frequency domain. Both representations were then mapped to a common token width. The standardised current and voltage channels were processed by three-layer causal one-dimensional convolutions. Arc sound was first transformed into a 64-band log-Mel representation. Three two-dimensional convolutions, causal along time, then processed this representation. The encoded representation was averaged along the frequency dimension. Both signal encodings were aggregated into eight time bins within the current window and projected to 96 dimensions. This procedure yielded at most 16 signal tokens.

Cross-modal interaction used the camera frame as the query unit. The accessible signal information was restricted by the recorded-time axis. Each frame could access only valid signal tokens whose temporal support had ended by that frame. Signal tokens received modality-type and relative-time embeddings. Visual tokens received positional and inter-frame temporal information. Let $\mathbf{V}_k \in \mathbb{R}^{N_k \times 96}$ denote the valid visual tokens for frame k , where $20 \leq N_k \leq 24$, and let $\mathbf{S}_k \in \mathbb{R}^{L_k \times 96}$ denote the $L_k \leq 16$ temporally available acoustic and electrical tokens. Their multi-head attention (MHA) interaction [32], using layer normalisation (LN) and a feed-forward network (FFN), was

$$\tilde{\mathbf{V}}_k = \mathbf{V}_k + \text{MHA}(\text{LN}(\mathbf{V}_k), \text{LN}(\mathbf{S}_k), \mathbf{S}_k; \mathbf{M}_k), \quad \hat{\mathbf{V}}_k = \tilde{\mathbf{V}}_k + \text{FFN}(\tilde{\mathbf{V}}_k). \quad (1)$$

In Eq. (1), LN acts on the visual queries and acoustic-electrical keys; the unnormalised signal tokens provide the values. The mask $\mathbf{M}_k$ excludes invalid or not-yet-available signal tokens. MHA uses three heads at feature width 96. The FFN comprises layer normalisation, a $96 \rightarrow 192$ linear mapping, GELU activation and a $192 \rightarrow 96$ linear mapping. If no valid signal is available, the attention term is omitted while the visual and feed-forward paths are retained. This temporal constraint is determined from recorded timestamps.

3.2.3 Window-level temporal aggregation

Frame-wise interaction associated each visual frame with signal features available up to that frame. Window-level aggregation then organised the resulting sequence. For every frame, the network learned a scalar weight for each valid visual token. A softmax-weighted sum produced the frame representation $\mathbf{u}_k$. The original camera-frame sequence was subsequently processed by a gated recurrent unit (GRU) [33]:

$$\mathbf{h}_k = \text{GRUCell}(\mathbf{u}_k, \mathbf{h}_{k-1}), \quad \mathbf{z} = \mathbf{h}_K + \mathbf{W}_p \tilde{\mathbf{p}}, \quad \mathbf{h}_0 = \mathbf{0}. \quad (2)$$

Here, K is the number of real camera frames in the window, $\mathbf{u}_k \in \mathbb{R}^{96}$ is the weighted visual-token representation of frame k , and $\mathbf{h}_k \in \mathbb{R}^{96}$ is the recurrent state after that frame. Padding frames used for batch processing do not update the state. When different original frames share a timestamp, their camera-sequence order is retained. The vector $\tilde{\mathbf{p}} \in \mathbb{R}^4$ contains the normalised current setpoint, voltage setpoint, welding speed and plate thickness. The matrix $\mathbf{W}_p \in \mathbb{R}^{96 \times 4}$ maps these conditions without a bias into the window representation $\mathbf{z} \in \mathbb{R}^{96}$. The setpoints remain distinct from the measured fluctuations captured by the acoustic and electrical observations.

The truncated visual backbone comprised the first 14 MobileNetV2 feature modules, initialised from ImageNet weights [34] with batch-normalisation running statistics fixed. For reproducibility, the electrical convolutional layers used output widths of 24, 48 and 64, kernel sizes of 31, 15 and 9, and strides of 5, 2 and 2. The 64-band log-Mel transform used a 916-sample periodic Hann window, a 366-sample hop and a 1024-point fast Fourier transform (FFT). Mel-band power P_{Mel} was transformed as $\ln(P_{\text{Mel}} + 10^{-6})$, and the acoustic convolutional layers used output widths of 16, 32 and 48. All data-derived normalisation statistics were estimated from the training records and fixed for validation.

3.3 Time-varying finite-plate thermal history

3.3.1 Finite-plate response driven by time-varying power

Whereas the 200-ms window describes current process behaviour, the thermal branch represents the preceding power-driven response. We combined a double-ellipsoidal heat source [35] with the analytical finite-plate response for time-varying heat sources of Zhu *et al.* [24]. Its nominal response incorporates changing input power, source motion and plate boundaries before reduction to a fixed-dimensional classification condition.

The detected start of the main-arc interval in each record defines the nominal arc-onset time t_0 . The 0.5-s buffer in Section 2.4 constrains window selection only. Thermal history is still accumulated from t_0 .

Power is binned with $\Delta = 0.1$ s. For an anchor t_a , only complete bins with right endpoints no later than the anchor are used. The calculation time is $t_n = n\Delta$ after arc onset, where

$$n = \left\lfloor \frac{t_a - t_0}{\Delta} \right\rfloor,$$

$\lfloor \cdot \rfloor$ denotes the floor function. The efficiency-corrected mean input power in bin j is

$$Q_j = \frac{\eta}{\Delta} \int_{j\Delta}^{(j+1)\Delta} I_+(t_0 + s) U_+(t_0 + s) ds, \quad j = 0, \dots, n-1. \quad (3)$$

Here, $I(t)$ and $U(t)$ denote the converted current and voltage waveforms, respectively, and $I_+ = \max(I, 0)$, $U_+ = \max(U, 0)$. The 12-bit acquisition counts were converted after subtracting the zero point of 2048, using calibrated coefficients of approximately 0.275783 A/count for current and 0.0456873 V/count for voltage. These factors were obtained by experimental calibration against reference measurements. With nominal thermal efficiency $\eta = 0.80$, the point-wise power product was averaged within each bin; the current and voltage were not averaged separately before multiplication. Thus, Q_j is an efficiency-weighted mean power in watts, while $Q_j\Delta$ is the energy delivered in that bin in joules. For a record-specific nominal travel speed v in mm/s, Q_j/v is the corresponding nominal line heat input in J/mm.

Let x , y and z denote the welding, transverse and thickness directions, respectively, with t measured from nominal arc onset. The heat-source centre moves along $x = vt$. The two plates, each $300 \times 150 \times D$ mm, are represented in the thermal-response calculation by an equivalent $300 \times 300 \times D$ mm plate domain. The mirror-boundary parameters are $B = B_l = 150$ mm. Plate thickness D and welding speed v are set for each record. The temperature rise is written as

$$\theta(\mathbf{r}, t) = T(\mathbf{r}, t) - T_0 = \int_0^t Q(s) G_{\text{fp}}(\mathbf{r}, t, s; \boldsymbol{\psi}) ds. \quad (4)$$

Here, $\mathbf{r} = (x, y, z)$, T is calculated temperature, T_0 is the nominal initial temperature, and G_{fp} is a finite-plate response kernel containing the moving double-ellipsoidal distribution and image terms. The parameter vector $\boldsymbol{\psi}$ comprises the material properties, heat-source scales and plate-domain parameters. Thermal efficiency is included in Q , whose units are watts; G_{fp} maps power integrated over time to temperature rise. The integral is evaluated piecewise within each power bin. Subintervals are no longer than 0.05 s and use two-point Gaussian quadrature.

Table 3 lists the nominal parameters. Q235 properties were fixed at representative 500 °C values from Ref. [36]; efficiency was set to 0.80 with reference to GMAW calorimetry [37]. Millimetre-scale heat-source semiaxes were selected with reference to a related welding study [38] and fixed at 3, 4, 3 and 9 mm. The equivalent rectangular domain and straight source path abstract the groove, weaving and pool flow. Their calculated response provides a thermal-history descriptor rather than a measured pool temperature.

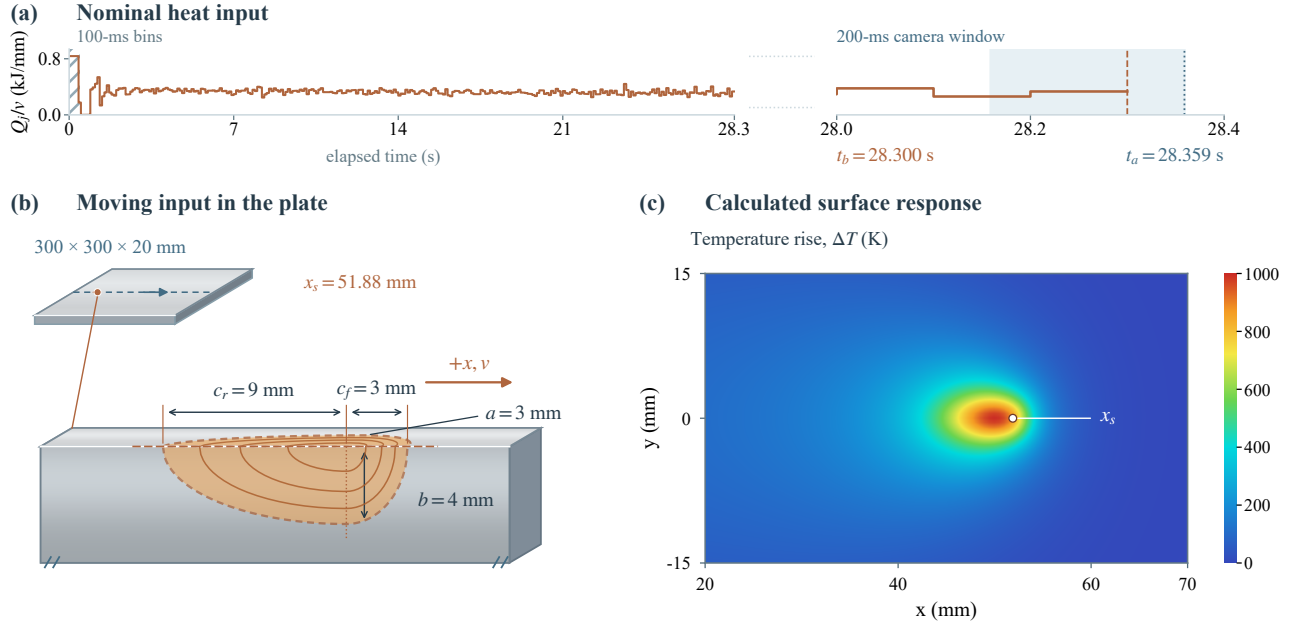


Fig. 6 Nominal heat input and finite-plate response: (a) binned line energy, (b) moving double-ellipsoidal source in the equivalent plate, and (c) calculated surface temperature rise. The orange envelope in (b) depicts the prescribed heat-source distribution

Table 3 Nominal parameters for the time-varying thermal-history calculation

Parameter	Value	Adopted property
Density ρ	7800 kg/m ³	Q235 literature value [36]
Specific heat c_p , thermal conductivity λ	530 J/(kg · K), 33 W/(m · K)	Representative values at 500 °C, used as constants [36]
Initial temperature T_0	293.15 K	Uniform nominal room temperature
Thermal efficiency η	0.80	Engineering nominal choice with reference to calorimetry [37]
Transverse and depth semi-axes a, b	3, 4 mm	Heat-source scale guided by Ref. [38]
Front and rear semi-axes c_f, c_r	3, 9 mm	Uniform setting with a rear-to-front length ratio of 3:1
Front and rear weights f_f, f_r	0.5, 1.5	Satisfying $f_f + f_r = 2$ and $f_f/c_f = f_r/c_r$
Plate-domain length and width	300, 300 mm	Equivalent calculation dimensions for two 300×150 -mm plates
Plate thickness D , welding speed v	Record-specific process values	Set according to each record

Table 4 Six-dimensional thermal-history features and their meanings

Feature	Definition	Information represented
H_1 Peak rise along the transverse sampling line	$\max_i \theta_i^y$	Local response intensity, K
H_2 Weighted transverse spread	$\sqrt{\sum_i w_i y_i^2 - (\sum_i w_i y_i)^2},$ $w_i = \theta_i^y / \sum_j \theta_j^y$	Response-weighted transverse distribution scale, mm
H_3 Longitudinal asymmetry	$(R - F)/(R + F),$ $R = \sum_{\xi_i < 0} \theta_i^x, F = \sum_{\xi_i > 0} \theta_i^x$	Relative rear–front response, dimensionless
H_4 Backside-to-surface response ratio	$\theta_b(t)/\theta_c(t)$	Relative through-thickness response, dimensionless
H_5 Rate of change of the moving-centre response	$[\theta_c(t) - \theta_c(t - \Delta)]/\Delta$	Change in centre response between adjacent calculation times, K/s
H_6 Cumulative input energy	$\sum_{j=0}^{n-1} Q_j \Delta$	Efficiency-weighted energy accumulated since arc onset, J

3.3.2 Six-dimensional thermal-history descriptor for recognition

Six features summarise the intensity, spatial distribution and accumulation of the calculated response illustrated in Fig. 6. Table 4 gives their definitions.

At $t = t_n$, a transverse sampling line is placed at $(vt, y, 0)$. It contains 41 equally spaced points over $y \in [-20, 20]$ mm. Their response rises are denoted by θ_i^y . Along the welding direction, 41 points are sampled at $(vt + \xi, 0, 0)$, with $\xi \in [-r, r]$ and

$$r = \min\{20 \text{ mm}, vt, 300 \text{ mm} - vt\},$$

where vt is expressed in millimetres. The corresponding response rises are denoted by θ_i^x . We further define the moving-centre surface response $\theta_c(t) = \theta(vt, 0, 0, t)$ and the backside response $\theta_b(t) = \theta(vt, 0, D, t)$.

The six-dimensional vector $\mathbf{H} = [H_1, \dots, H_6]^T$ is defined in Table 4. In the table, w_i is the response-normalised weight of transverse point i ; R and F sum the rear and front longitudinal responses, respectively, and are distinct from the plate-boundary parameters.

H_1 is the maximum along the transverse sampling line; H_4 is a through-thickness response ratio in the equivalent plate domain. H_5 compares responses at the moving source position at successive calculation times, using only the power history available at each time. Feature calculation requires at least 0.2 s of history and a source centre within the plate length. After validity checks, the six features are standardised using training-record statistics and supplied to the fusion module. Invalid responses are excluded rather than coded as zero.

3.4 Explicit molten-pool morphology

Explicit morphology characterises visible molten-pool geometry in the current window, complementing the preceding thermal history. Width, position and contour parameters have previously supported molten-pool monitoring [2, 29]. We extracted an eight-dimensional, dimensionless descriptor of region extent, position, occupancy and

dispersion from same-frame SAM 3 masks [31] and aggregated it over the 200-ms window. The learned local path and explicit descriptor share the same visual source but expose different information: the former learns appearance and spatial context, whereas the latter supplies fixed-definition geometric quantities.

3.4.1 Prompted extraction of the molten-pool region

The target was the visible molten-pool body, excluding the trailing solidified bead and groove walls. We ran pretrained SAM 3 [31] offline with a rectangular prompt in original-image coordinates, adding foreground or background points where needed. A reused localisation box came only from the same or an earlier frame; segmentation was always performed on the current image. SAM 3 weights remained fixed, and prompting defined the target rather than training the segmentation model.

Each inference yielded three candidate masks. Candidate masks were selected by visual inspection based on pool coverage and exclusion of bead and groove background. Where disconnected regions were present, the largest eight-connected component was retained. Accepted masks were used without hole filling or contour smoothing.

Each accepted mask remained linked to its original frame; frames without a usable mask were omitted from morphology aggregation. Figure 7 shows the extraction and geometric description.

3.4.2 Normalised description of mask geometry

Morphology features were computed from the accepted binary mask in the original 640×480 -pixel image coordinates. A fixed horizontal context $[120, 520]$ was used. Its width and height were $W_c = 400$ and $H_c = 480$, respectively. Let A denote the number of foreground pixels in a valid mask. Let w_b and h_b denote the width and height of its axis-aligned bounding box. Let (x_i, y_i) denote pixel-centre coordinates relative to the upper-left corner of the context. The centroid is denoted by $(\bar{x}, \bar{y})$. Pixel centres are defined as integer row and column indices plus 0.5. The bounding box uses a half-open interval. An empty mask or a mask extending beyond the horizontal context is treated as invalid. The descriptor for frame k is

$$\mathbf{g}_k = \left[\frac{A}{W_c H_c}, \frac{w_b}{W_c}, \frac{h_b}{H_c}, \frac{\bar{x}}{W_c}, \frac{\bar{y}}{H_c}, \frac{A}{w_b h_b}, \frac{\text{Var}(x)}{W_c^2}, \frac{\text{Var}(y)}{H_c^2} \right]^\top. \quad (5)$$

The first three components describe foreground area and bounding-box extent; the next two locate the centroid in the fixed field of view. The sixth is foreground occupancy within the bounding box, and the final two are population variances of foreground-pixel coordinates about the centroid. Together, these quantities distinguish the pool’s occupied area, extent, location and spatial distribution (Fig. 7).

All quantities are normalised by the corresponding length, area or squared-length scale and remain dimensionless. Calculations use the accepted foreground mask, including its holes, or the retained component when component processing was applied. The descriptor represents geometry in the fixed image field, not calibrated physical size.

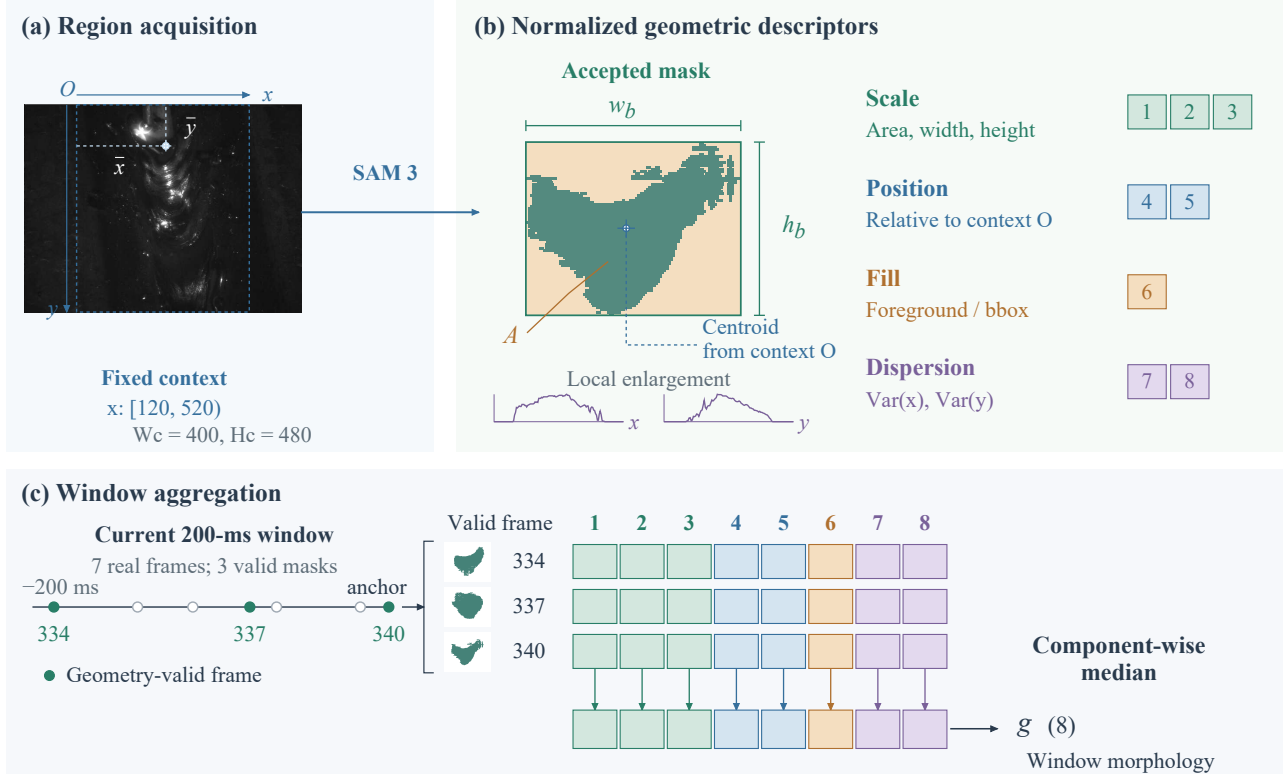


Fig. 7 Molten-pool morphology extraction and aggregation. (a) Same-frame mask acquisition within the fixed image context. (b) Eight dimensionless descriptors of scale, position, occupancy and dispersion; marginal traces show peak-scaled foreground-pixel counts. (c) Component-wise median across valid frames in the 200-ms window

3.4.3 Window-level aggregation and condition input

For anchor t_a , the component-wise median of valid frame descriptors $\mathbf{g}_k$ within $[t_a - 0.2, t_a]$ defined the window descriptor $\mathbf{g}$. Masks were matched by original-frame identity; for duplicate timestamps, only frames up to the anchor’s sequence index were eligible. This summary captured typical geometry, while frame order and process dynamics remained in the observation encoder (Section 3.2).

A window without valid segmentation was retained with an invalid morphology flag. Local visual sampling and explicit morphology used separate validity checks, because an accepted mask could still yield an unusable expanded local-feature box.

Valid morphology descriptors were standardised with training-record statistics. For an invalid descriptor, the standardised input was set to zero under the validity flag; the zero-input-centred embedding in Section 3.5 then removed its contribution without suppressing the observation or thermal branches. The matched ablation retained the learned local visual path.

3.5 Joint conditioning and classification

Thermal history and explicit morphology describe the preceding thermal process and current image-space geometry, respectively. HMCF-Net applies joint conditioning through a residual update of the observation representation. The two descriptors are first mapped separately to 32 dimensions and added to form a joint condition. This condition and the 96-dimensional window observation then generate a nonlinear residual. A bypass retains the original observation, as shown in the HMCF-Net fusion module in Fig. 8. The condition descriptors therefore update the visual–acoustic–electrical representation through the residual path.

Let $\tilde{\mathbf{H}}$ and $\tilde{\mathbf{g}}$ denote the standardised features. Thermal history uses a $6 \rightarrow 32$ linear mapping. Morphology uses an $8 \rightarrow 84 \rightarrow 32$ nonlinear mapping with the zero-input response as its reference:

$$\mathbf{e}_H = \mathbf{W}_H \tilde{\mathbf{H}} + \mathbf{b}_H, \quad f_g(\mathbf{x}) = \mathbf{W}_{g2} \text{GELU}(\mathbf{W}_{g1} \mathbf{x} + \mathbf{b}_{g1}) + \mathbf{b}_{g2},$$

$$\mathbf{g}^* = \begin{cases} \tilde{\mathbf{g}}, & \text{if morphology is valid,} \\ \mathbf{0}, & \text{if morphology is invalid,} \end{cases} \quad \mathbf{e}_g = f_g(\mathbf{g}^*) - f_g(\mathbf{0}), \quad \mathbf{c} = \mathbf{e}_H + \mathbf{e}_g. \quad (6)$$

Here, $\mathbf{e}_H, \mathbf{e}_g, \mathbf{c} \in \mathbb{R}^{32}$ are the thermal, morphological and joint condition embeddings. The matrices $\mathbf{W}_H$, $\mathbf{W}_{g1}$ and $\mathbf{W}_{g2}$, with their corresponding bias vectors, define the stated $6 \rightarrow 32$ and $8 \rightarrow 84 \rightarrow 32$ mappings. For an invalid morphology input, $\mathbf{g}^* = \mathbf{0}$ and hence $\mathbf{e}_g = \mathbf{0}$ despite the mapping biases. The observation and thermal-history branches remain available.

The joint condition $\mathbf{c} \in \mathbb{R}^{32}$ and observation state $\mathbf{z}$ are linearly projected. The projections are added before a nonlinear mapping and form a residual:

$$\Delta \mathbf{z} = \mathbf{W}_o \text{GELU}(\mathbf{W}_z \mathbf{z} + \mathbf{b}_z + \mathbf{W}_c \mathbf{c}) + \mathbf{b}_o, \quad \mathbf{z}' = \text{LN}(\mathbf{z} + \Delta \mathbf{z}). \quad (7)$$

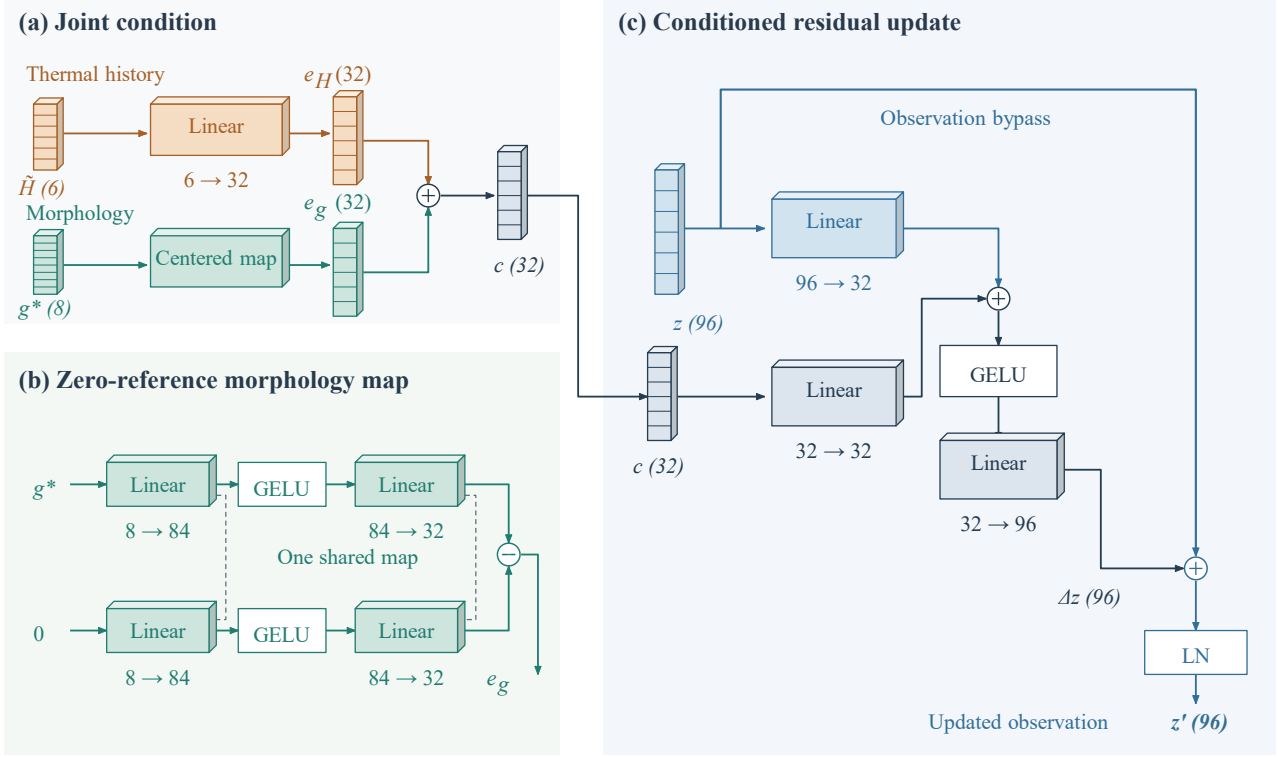


Fig. 8 Thermal-history and morphology conditioning: (a) joint condition formation, (b) zero-reference morphology mapping, and (c) residual update of the observation representation

Here, $\mathbf{W}_z \in \mathbb{R}^{32 \times 96}$, $\mathbf{W}_c \in \mathbb{R}^{32 \times 32}$ and $\mathbf{W}_o \in \mathbb{R}^{96 \times 32}$; the associated $\mathbf{b}_z$ and $\mathbf{b}_o$ are 32- and 96-dimensional biases. The conditioned projection $\mathbf{W}_c$ has no bias. The observation term $\mathbf{W}_z \mathbf{z}$ and condition term $\mathbf{W}_c \mathbf{c}$ are combined before the nonlinear transformation. The residual $\Delta \mathbf{z} \in \mathbb{R}^{96}$ therefore depends jointly on the current observation and the two conditions. Layer normalisation yields $\mathbf{z}' \in \mathbb{R}^{96}$, which is passed to a linear classifier:

$$\ell = \mathbf{W}_y \mathbf{z}' + \mathbf{b}_y, \quad \hat{y} = \arg \max_{j \in \{1, 2, 3, 4\}} \ell_j. \quad (8)$$

The classifier in Eq. (8) has $\mathbf{W}_y \in \mathbb{R}^{4 \times 96}$ and $\mathbf{b}_y \in \mathbb{R}^4$; $\ell \in \mathbb{R}^4$ denotes class logits, and $\hat{y}$ is the highest-logit class index in the fixed class order stated in Section 2.

3.6 Training and evaluation protocol

Masks, geometric descriptors and thermal responses were computed offline before training. HMCf-Net used four-class cross-entropy; its observation encoder, condition mappings and classifier were optimised jointly. All comparisons used the weld-level split described in Section 2 and were evaluated on the validation set.

The complete model and three matched ablations used the same visual–acoustic–electrical observation encoder and process inputs; each configuration was trained separately with the specified descriptors present or omitted. The three alternative architectures shared the visual front end and received the same windowed images, electrical and acoustic observations and four process inputs, but not the nominal thermal-history or explicit morphology descriptors. The convolutional neural network (CNN)–long short-term memory (LSTM) comparator combined frame-level visual features with the electrical and acoustic features available at each frame and aggregated the sequence using an LSTM cell. CNN–Transformer used the same frame-level feature construction but replaced recurrence with a two-layer, four-head temporal Transformer with time embeddings and a causal attention mask. Cross-Attention Fusion used visual frame queries to attend separately to time-available electrical and acoustic tokens, followed by feed-forward processing and gated recurrent unit aggregation. Cross-Attention Fusion follows a general cross-modal interaction pattern adapted from [39]. The comparison models were included to evaluate HMCF-Net against alternative temporal modelling and cross-modal fusion architectures, while the matched ablations assessed the individual and joint contributions of the morphology and thermal-history descriptors.

All seven models used the same windows, class order, record-balanced sampling and random seed (42), and were optimised with AdamW [40]. The learning rate and weight decay were both 10^{-4} ; the pretrained visual backbone used one tenth of the main learning rate, and gradients were clipped at a norm of 5. All seven models used a logical batch size of 16. Training was limited to 20 epochs, with early stopping after four consecutive non-improving epochs; the checkpoint with the highest validation macro-F1 was retained. Accuracy was calculated over validation windows. Balanced accuracy (BA) and macro-F1 are the arithmetic means of the four class recalls and F1 scores, respectively. Confusion matrices and per-weld window accuracies provided descriptive class- and record-level analyses.

4 Results and discussion

4.1 Overall performance and model comparison

HMCF-Net achieved the highest accuracy, balanced accuracy and macro-F1 among the evaluated methods, reaching 94.72%, 93.04% and 93.46%, respectively, across 5,078 validation windows (Table 5). Among the alternative architectures, Cross-Attention Fusion attained the highest macro-F1, at 90.94%, which HMCF-Net exceeded by 2.52 percentage points. The matched ablations in Section 4.2 isolate the contributions of the two condition descriptors.

Class-wise results show how the overall performance was distributed across welding states (Fig. 9). HMCF-Net attained recalls of 95.49%, 95.30%, 82.43% and 98.96% for normal welding, incomplete penetration, burn-through and weld deviation, respectively. For weld deviation, F1 was 99.48% across 1,246 windows from 2 validation welds. The model therefore combined high recognition of normal and incomplete-penetration states with accurate discrimination of weld deviation.

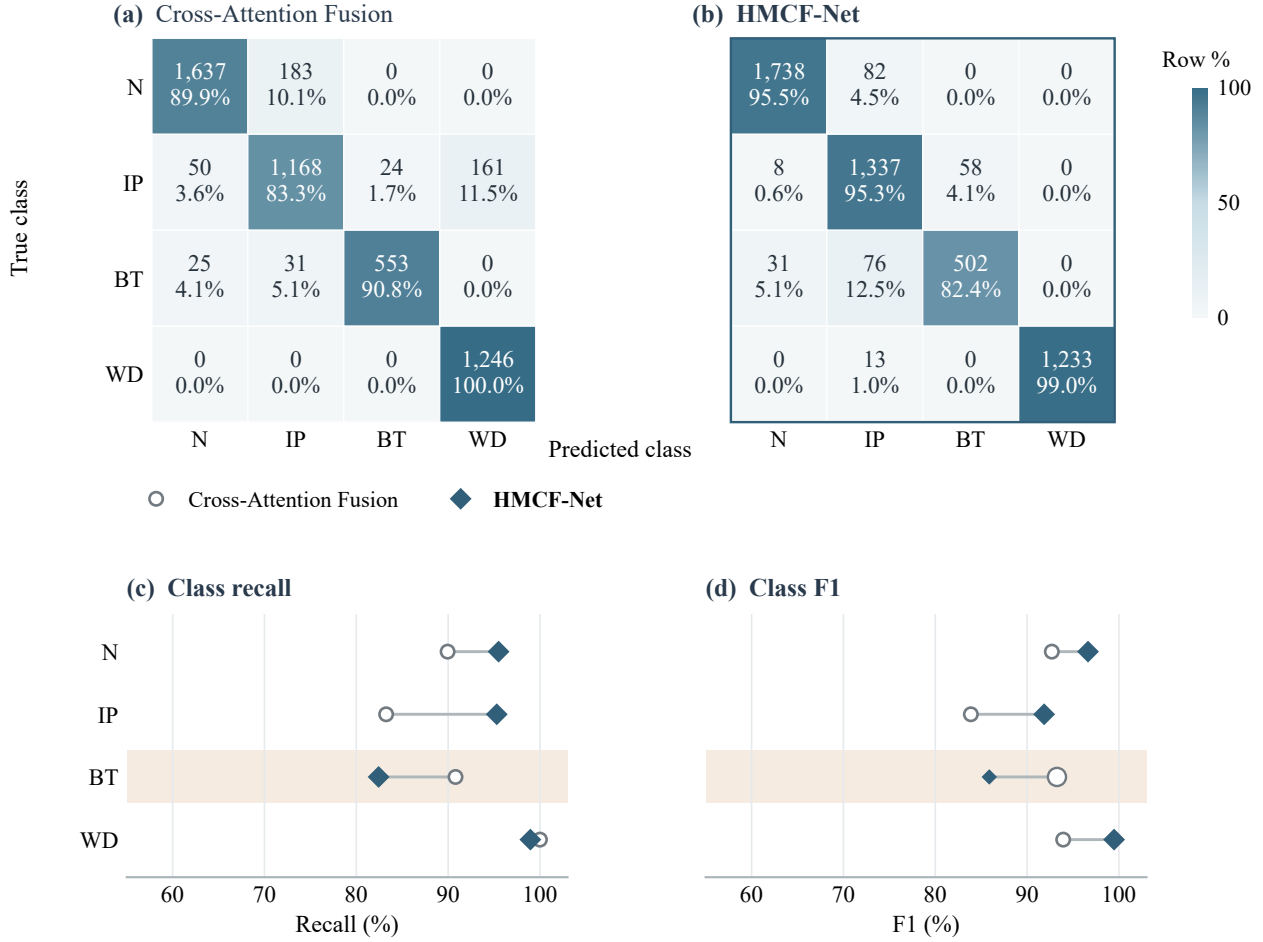


Fig. 9 Class-specific validation results for Cross-Attention Fusion (CAF) and HMCF-Net. (a, b) Confusion matrices show counts and true-class row percentages on a shared scale. (c, d) Recall and F1; open circles indicate CAF, filled diamonds HMCF-Net, and shading burn-through. N, normal; IP, incomplete penetration; BT, burn-through; WD, weld deviation

Table 5 Overall window-level performance on the validation set (%)

Method	Accuracy	BA	Macro-F1
CNN-LSTM	89.33	88.63	87.41
CNN-Transformer	85.78	85.81	84.94
Cross-Attention Fusion	90.67	91.00	90.94
HMCF-Net observation-only	92.26	91.57	91.78
Complete HMCF-Net	94.72	93.04	93.46

Compared with Cross-Attention Fusion, HMCF-Net increased F1 for normal welding from 92.70% to 96.64%, incomplete penetration from 83.88% to 91.86%, and weld deviation from 93.93% to 99.48%. Burn-through remained more difficult: its F1 was 85.89% for HMCF-Net and 93.25% for Cross-Attention Fusion. The class- and weld-level analysis in Section 4.3 examines the remaining penetration-state confusions.

4.2 Matched ablation of thermal history and explicit morphology

The matched ablations evaluated the contribution of each descriptor while retaining the shared observation encoder (Table 6). Observation-only achieved 91.78% macro-F1; adding explicit morphology increased it to 92.20%, and adding nominal thermal history increased it to 92.39%. Joint conditioning achieved 93.46% macro-F1, exceeding observation-only, observation plus morphology and observation plus thermal history by 1.69, 1.27 and 1.07 percentage points, respectively. The two descriptors each improved recognition, and their joint use provided the highest macro-F1 in the matched comparison.

Table 6 Matched ablation of the two condition descriptors with the observation encoder retained in every configuration (%). A check mark indicates that the descriptor is included; a dash indicates that it is omitted

Configuration	Morphology	Thermal history	Accuracy	BA	Macro-F1
Observation only	–	–	92.26	91.57	91.78
Observation + morphology	✓	–	92.91	91.70	92.20
Observation + thermal history	–	✓	93.66	92.08	92.39
Complete HMCF-Net	✓	✓	94.72	93.04	93.46

The clearest class-level gain over observation-only was in normal-weld recognition: recall increased from 88.19% to 95.49%, while incomplete-penetration precision increased from 80.94% to 88.66%. This combination indicates fewer normal windows being assigned to incomplete penetration. Weld-deviation recall also increased, from 98.07% to 98.96%. The improvement was not uniform: burn-through recall was 82.43%,

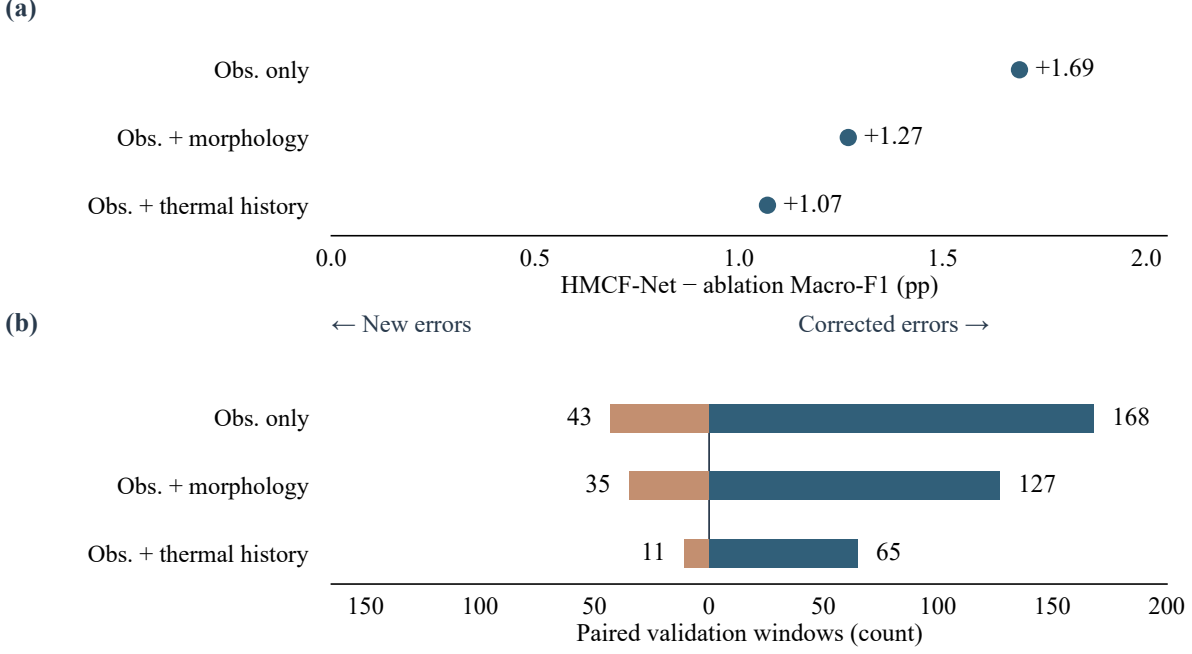


Fig. 10 Matched ablation results: (a) macro-F1 gains and (b) paired errors corrected and introduced by HMCf-Net relative to each variant. Obs., observation encoder

compared with 84.07% for observation-only. Paired predictions showed that HMCf-Net corrected 168 observation-only errors and introduced 43 new ones, reducing the total by 125 (Fig. 10). The corresponding net reductions were 92 and 54 relative to the observation-plus-morphology and observation-plus-thermal-history variants, respectively.

The ablations support conditioning the learned observation representation with two descriptors that have distinct roles. Nominal thermal history represents the accumulated response from arc onset, including the period before the observation window. Morphology makes current pool geometry explicitly available to the conditioning path. The morphology gain was obtained with the learned local visual path retained, indicating value in the fixed-definition geometric representation alongside learned appearance features. This result extends the use of pool geometry for monitoring or control [2, 29] to conditioning a multimodal sequence representation [20, 21]. HMCf-Net thus combines process-history context and explicit geometry without extending the visual observation window.

4.3 Class-boundary errors and weld-level variation

Weld-level accuracies show how the aggregate improvement was distributed across validation records. In one normal weld, window accuracy increased from 71.06%

for observation-only to 88.96% for HMCF-Net; the other normal weld remained at 100%. The two weld-deviation records achieved 100% and 98.16% accuracy, with the latter improving from 96.61% for observation-only. These results locate substantial recognition gains within individual weld records.

Incomplete penetration and burn-through remained the largest bidirectional confusion pair, with 76 burn-through windows assigned to incomplete penetration and 58 windows confused in the reverse direction. Burn-through recall was 82.43% across 4 welds, whose window accuracies ranged from 70.85% to 100%; normal-to-incomplete-penetration errors also remained visible in Fig. 9. Such variation is consistent with short local observations incompletely describing the process and joint conditions that determine the final penetration state. Previous studies have linked penetration behaviour to process signals, joint gap or groove geometry and molten-pool response [4–6].

Taken together, the model comparisons and matched ablations support augmenting immediate multimodal observations with explicit process-history and geometric context. HMCF-Net achieved the strongest overall recognition by bringing these descriptions into a shared conditioned representation. Its contribution lies in combining information from different descriptive scales within a short observation window, while the residual penetration-state ambiguity provides a specific target for further improvement.

5 Conclusions

This study developed HMCF-Net for robotic MAG weld-quality classification by integrating accumulated thermal history and explicit molten-pool geometry with short-window multimodal observations. The main conclusions are as follows.

1. Nominal thermal history describes the accumulated response, including the period before the observation window, while explicit morphology represents geometry from the same camera images. The joint condition and current observation state generate a nonlinear residual update, with a bypass retaining the learned multimodal representation. This integrates process history and explicit geometry without direct temperature measurement or a longer camera window.
2. HMCF-Net achieved 94.72% accuracy, 93.04% balanced accuracy and 93.46% macro-F1 across 5,078 windows from 10 validation welds, yielding the highest overall scores among the evaluated models. Its macro-F1 exceeded the strongest alternative model by 2.52 percentage points.
3. Matched ablations showed that each descriptor improved macro-F1 with the observation encoder retained, and their joint use produced the strongest result. Joint conditioning improved macro-F1 by 1.69 percentage points over the matched model without either descriptor. These results support the complementary value of calculated thermal history and explicit image-space geometry alongside learned multimodal features.

The findings support condition-guided fusion for connecting immediate welding observations with accumulated process history and explicit pool geometry. Class-wise

gains were uneven, with burn-through discrimination remaining a priority for improvement. Future work will focus on distinguishing penetration states and evaluating transferability across independent plates and broader joint conditions.

Statements and Declarations

Funding. This work was supported by the National Science and Technology Major Project (Grant No. 2026ZD1613300) and the National Natural Science Foundation of China (Grant No. 52505411).

Author contributions. Bin Zhong conducted the experiments, developed the code, prepared the figures and drafted the manuscript. Kanghong Zhu provided the heat-source model. Runquan Xiao and Yingcai Wan contributed to research discussions. Yiyang Zhao provided guidance on manuscript development. Huabin Chen acquired funding and reviewed the manuscript.

Competing interests. The authors declare no competing interests.

Data availability. The data supporting the findings of this study are available from the corresponding authors upon reasonable request.

Code availability. The code used in this study is available from the corresponding authors upon reasonable request.

References

- [1] Zhu K, Wang Q, Chen W, et al (2025) Robotic MAG welding defects and quality assessment with a defect threshold decision model-driven method. *Mechanical Systems and Signal Processing* 224:112056. <https://doi.org/10.1016/j.ymssp.2024.112056>
- [2] Biber A, Sharma R, Reisgen U (2024) Robotic welding system for adaptive process control in gas metal arc welding. *Welding in the World* 68(9):2311–2320. <https://doi.org/10.1007/s40194-024-01756-y>
- [3] Cheng Y, Yu R, Zhou Q, et al (2021) Real-time sensing of gas metal arc welding process – a literature review and analysis. *Journal of Manufacturing Processes* 70:452–469. <https://doi.org/10.1016/j.jmapro.2021.08.058>
- [4] Yu R, Cao Y, Chen H, et al (2023) Deep learning based real-time and in-situ monitoring of weld penetration: Where we are and what are needed revolutionary solutions? *Journal of Manufacturing Processes* 93:15–46. <https://doi.org/10.1016/j.jmapro.2023.03.011>
- [5] Sikström F, Öberg AE (2017) Prediction of penetration in one-sided fillet welds by in-process joint gap monitoring—an experimental study. *Welding in the World* 61(3):529–537. <https://doi.org/10.1007/s40194-017-0448-7>
- [6] Gao P, Wu Z, Wang Y, et al (2024) Method for monitoring and controlling penetration of complex groove welding based on online multi-modal data.

- [7] Wang J, Zhang Z, Bai Z, et al (2023) On-line defect recognition of MIG lap welding for stainless steel sheet based on weld image and CMT voltage: Feature fusion and attention weights visualization. *Journal of Manufacturing Processes* 108:430–444. <https://doi.org/10.1016/j.jmapro.2023.10.081>
- [8] Liu L, Chen H, Chen S (2022) Quality analysis of CMT lap welding based on welding electronic parameters and welding sound. *Journal of Manufacturing Processes* 74:1–13. <https://doi.org/10.1016/j.jmapro.2021.11.055>
- [9] Cui Y, Shi Y, Zhu T, et al (2020) Welding penetration recognition based on arc sound and electrical signals in K-TIG welding. *Measurement* 163:107966. <https://doi.org/10.1016/j.measurement.2020.107966>
- [10] Lin M, Xu Y, Yu C, et al (2025) Online monitoring of typical welding defects in robotic laser-MAG hybrid welding based on multi-source information fusion and stacking-PSO-LightGBM. *Journal of Manufacturing Processes* 156:780–800. <https://doi.org/10.1016/j.jmapro.2025.10.114>
- [11] Bilal M, Hussain S, Jawad M, et al (2026) A multi-sensor-based intelligent monitoring system for surface and internal defects in robotic GTAW process. *Welding in the World* 70(8):3307–3331. <https://doi.org/10.1007/s40194-025-02270-5>
- [12] Cao L, Li J, Zhang L, et al (2023) Cross-attention-based multi-sensing signals fusion for penetration state monitoring during laser welding of aluminum alloy. *Knowledge-Based Systems* 261:110212. <https://doi.org/10.1016/j.knosys.2022.110212>
- [13] Xu Y, Liu Q, Xu J, et al (2025) A multi-spectral channel attention mechanism for prediction of welding state during pulsed GTAW. *Journal of Manufacturing Processes* 134:1021–1033. <https://doi.org/10.1016/j.jmapro.2025.01.023>
- [14] Xu Y, Liu Q, Xu J, et al (2025) An effective penetration depth and width prediction method in pulsed GTA welding based on multimodal transformer-serial fusion network. *Journal of Manufacturing Processes* 156:252–267. <https://doi.org/10.1016/j.jmapro.2025.10.116>
- [15] Lin Z, Xu Y, Wang X, et al (2026) Transfer learning based on multimodal feature fusion: A study on intelligent detection of robotic MAG welding defects under cross-condition and small-sample scenarios. *Journal of Manufacturing Processes* 162:269–284. <https://doi.org/10.1016/j.jmapro.2026.02.014>
- [16] Mattera G, Cao Y, Zhang Y, et al (2026) Vision-based online quality tracking in wire arc additive manufacturing via hybrid unsupervised deep learning–statistical process monitoring. *Welding in the World* 70(4):1261–1281. <https://doi.org/10.1007/s40194-025-02270-5>

- [17] Chen C, Xiao R, Chen H, et al (2021) Prediction of welding quality characteristics during pulsed GTAW process of aluminum alloy by multisensory fusion and hybrid network model. *Journal of Manufacturing Processes* 68:209–224. <https://doi.org/10.1016/j.jmapro.2020.08.028>
- [18] Gao P, Su X, Wu Z, et al (2024) Online penetration prediction based on multimodal continuous signals fusion of CMT for full penetration. *Journal of Manufacturing Processes* 115:431–440. <https://doi.org/10.1016/j.jmapro.2024.02.033>
- [19] Yang B, Tan C, Chen G, et al (2024) Online monitoring system for welding states of bottom-locking joints in high-speed trains via multi-information fusion and 3DCNN. *Journal of Manufacturing Processes* 113:105–116. <https://doi.org/10.1016/j.jmapro.2024.01.055>
- [20] Yuan H, Wei S, Yi J, et al (2026) High-speed monitoring and control of forming defects in coarse-wire MAG welding based on multi-frame spatio-temporal neural networks. *Welding in the World* 70(8):3533–3551. <https://doi.org/10.1007/s40194-026-02471-6>
- [21] El Assal O, Mateo-Agullo C, Corolleur A, et al (2026) A comparative study of deep learning models for welding process analysis using sequential models. *Welding in the World* <https://doi.org/10.1007/s40194-026-02376-4>, published online 13 February 2026
- [22] Hong Y, Yang M, Yuan R, et al (2024) A novel quality monitoring approach based on multigranularity spatiotemporal attentive representation learning during climbing GTAW. *IEEE Transactions on Industrial Informatics* 20(6):8218–8228. <https://doi.org/10.1109/TII.2024.3369235>
- [23] Jiang R, Xiao R, Chen S (2021) Prediction of penetration based on infrared thermal and visual images during pulsed GTAW process. *Journal of Manufacturing Processes* 69:261–272. <https://doi.org/10.1016/j.jmapro.2021.07.046>
- [24] Zhu K, Liu W, Qi B, et al (2026) Analytical model for weld penetration prediction in finite plates with time-varying heat sources. *International Journal of Heat and Mass Transfer* 256:127921. <https://doi.org/10.1016/j.ijheatmasstransfer.2025.127921>
- [25] Meng X, Bachmann M, Yang F, et al (2026) Porosity prediction in laser beam welding with a multimodal physics-informed machine learning framework. *Advanced Engineering Informatics* 74:104611. <https://doi.org/10.1016/j.aei.2026.104611>

- [26] Li S, Liu X, Xu X, et al (2026) A welding penetration prediction model for laser welding process based on self-supervised learning using physics-informed neural networks. *Journal of Manufacturing Processes* 160:642–662. <https://doi.org/10.1016/j.jmapro.2026.01.035>
- [27] Rashid A, Vatandoust F, Kota A, et al (2025) Bead geometry prediction in wire arc directed energy deposition using physics-informed machine learning and low-fidelity data. *Additive Manufacturing* 109:104881. <https://doi.org/10.1016/j.addma.2025.104881>
- [28] Cheng S, Yu R, Shen A, et al (2025) MSDF-TL: A transfer learning and physics-constrained multi-source data fusion framework for weld state prediction. *Journal of Manufacturing Processes* 154:504–523. <https://doi.org/10.1016/j.jmapro.2025.09.070>
- [29] Zhou H, Chen H, He Y, et al (2025) Characteristic classification and extraction of robotic multi-layer multi-pass MAG welding pool—an extended UNet network implementation based on transfer learning. *Journal of Manufacturing Processes* 141:517–535. <https://doi.org/10.1016/j.jmapro.2024.12.016>
- [30] Sandler M, Howard A, Zhu M, et al (2018) MobileNetV2: Inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4510–4520, <https://doi.org/10.1109/CVPR.2018.00474>
- [31] Carion N, Gustafson L, Hu YT, et al (2026) SAM 3: Segment anything with concepts. In: International Conference on Learning Representations, pp 138846–138923, URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e0982cbc81401df3430ee1ff780dc7a2-Abstract-Conference.html
- [32] Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: *Advances in Neural Information Processing Systems*, vol 30. Curran Associates, Inc., pp 5998–6008, URL <https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- [33] Cho K, van Merriënboer B, Gulcehre C, et al (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pp 1724–1734, <https://doi.org/10.3115/v1/D14-1179>
- [34] Deng J, Dong W, Socher R, et al (2009) ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 248–255, <https://doi.org/10.1109/CVPR.2009.5206848>
- [35] Goldak J, Chakravarti A, Bibby M (1984) A new finite element model for welding heat sources. *Metallurgical Transactions B* 15(2):299–305. <https://doi.org/10.1007/BF02667333>

- [36] Tong Y, Ge Z, Zhuo X, et al (2018) Research on welding deformation for box girder of bridge crane based on thermal elasto-plastic theory. *Advances in Mechanical Engineering* 10(5):1687814018775885. <https://doi.org/10.1177/1687814018775885>, URL <https://journals.sagepub.com/doi/10.1177/1687814018775885>
- [37] Pépe N, Egerland S, Colegrove PA, et al (2011) Measuring the process efficiency of controlled gas metal arc welding processes. *Science and Technology of Welding and Joining* 16(5):412–417. <https://doi.org/10.1179/1362171810Y.0000000029>
- [38] Machado TdS, da Silveira T, Isoldi LA, et al (2025) Computational modeling of the temperature distribution in a butt weld of AISI 304L stainless steel using a volumetric heat source. *Metals* 15(12):1371. <https://doi.org/10.3390/met15121371>
- [39] Tsai YHH, Bai S, Liang PP, et al (2019) Multimodal transformer for unaligned multimodal language sequences. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp 6558–6569, <https://doi.org/10.18653/v1/P19-1656>
- [40] Loshchilov I, Hutter F (2019) Decoupled weight decay regularization. In: *International Conference on Learning Representations*, URL <https://openreview.net/forum?id=Bkg6RiCqY7>