Fusiform-Aware Network for online laser welding penetration state monitoring by the utilization of vision measurement system

Songlin Li [a], Ting Yuan [b], Jiawei Fan [b], Haonan Zhang [c], Zhuguo Li [c], Uwe D. Hanebeck [d], Jiuchao Qian [b,*]

[a] School of Integrated Circuits (School of Information Science and Electronic Engineering), Shanghai Jiao Tong University, 200240 Shanghai, China
[b] School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, 200240 Shanghai, China
[c] School of Materials Science and Engineering, Shanghai Jiao Tong University, 200240 Shanghai, China
[d] Intelligent Sensor-Actuator-Systems Laboratory (ISAS), Karlsruhe Institute of Technology, D-76021 Karlsruhe, Germany

ARTICLE INFO

Keywords:
Laser welding
Online penetration state monitoring
Vision-transformer based network
Vision-based measurement
Molten pool analysis

ABSTRACT

Accurate online monitoring of the penetration state is critical for quality assurance in laser welding, yet it remains a challenge due to the harsh process environment. To address the challenge, a novel transformer-based monitoring algorithm, named Fusiform-Aware Network (FANet), is proposed to process visual signals of the laser welding process. Firstly, the core of FANet is the Triplet Multi-head Linear Attention (TMLA) mechanism, a novel architecture specifically designed to align its receptive fields with the distinct, elongated fusiform geometry of the molten pool. Moreover, an overlapping partition strategy is proposed to further improve the alignment between the receptive field and molten pool. Furthermore, Scalar Gated Attention (SGA) is designed to ensure a high recognition speed suitable for online industrial applications. Experiments are conducted under a dedicated laser welding platform to validate the performance of the monitoring algorithm. Results demonstrate that the proposed method achieves a penetration state recognition accuracy of 93.24%, significantly outperforming general-purpose vision algorithms. This work provides an effective solution for online monitoring of the penetration state in industrial laser welding scenarios.

1. Introduction

Laser welding has become an indispensable manufacturing technique in modern industries, particularly in high-end sectors such as automotive, shipbuilding, aerospace, micro-electronics and nuclear industries [1]. However, the quality assurance and process stability of laser welding remain susceptible to multiple operational parameters, including laser power, workpiece geometrical characteristics, material surface characteristics, and ambient welding conditions [2,3]. Parametric instabilities frequently induce welding defects. These defects can significantly compromise mechanical properties and adversely affect product reliability.[4]. Among various quality indicators, the weld penetration state is a critical physical phenomenon that directly reflects the result of the weld formation [5]. Therefore, developing a method for accurate and real-time monitoring of the penetration state is a key research field in advanced manufacturing.

Various sensing modalities have been explored to monitor the penetration state indirectly, including acoustic sensing that analyzes sound emissions [6–8], light spectrum analysis that examines plasma plume characteristics [9,10], and thermal imaging that captures temperature distributions [11,12]. These methods provide valuable information regarding the welding process. However, they often offer only indirect indications of the penetration state and can be sensitive to ambient noise or instrument instabilities. In contrast, vision-based methods are considered the most effective approach. Visual features captured by cameras, such as the morphology of the molten pool, keyhole, plasma plume, and spatters, contain rich information. These features are directly correlated with the physical characteristics of the molten pool and the penetration state [13,14]. The challenge is designing a robust and accurate vision-based monitoring algorithm to analyze complex visual signals.

Early vision-based methods rely on hand-crafted image feature extraction methods and traditional image processing techniques [15]. These approaches typically extract the geometric features of the area, size and edge information of keyhole or the molten pool, and then build the relationship between visual features and penetration state by machine learning algorithms [14,16,17]. While traditional methods are

* Corresponding author.
E-mail addresses: finale007@sjtu.edu.cn (S. Li), tyuan@sjtu.edu.cn (T. Yuan), david0703@sjtu.edu.cn (J. Fan), zhn0706@sjtu.edu.cn (H. Zhang), lizg@sjtu.edu.cn (Z. Li), uwe.hanebeck@kit.edu (U.D. Hanebeck), jcqian@sjtu.edu.cn (J. Qian).

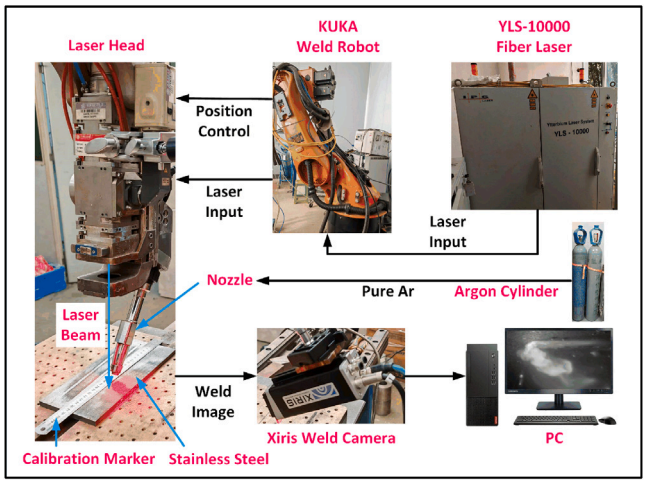

Fig. 1. Schematic diagram of vision-based measurement system for laser welding penetration state. A welding camera captures images of the molten pool and then are processed by FANet to determine the penetration state in real time.

effective under specific working conditions, their performance degrades when faced with changed welding parameters, workpiece surface conditions and variations of light. The traditional methods depends on the quality of the manually extracted features, which limits the adaptability to a variety of welding conditions.

Deep learning approaches have revolutionized vision-based penetration state monitoring tasks by building end-to-end models that learn intrinsic features from raw pixel data automatically and adaptively [18]. Among the deep learning models, Convolution Neural Network (CNN) based models have been used as a major type. Studies have demonstrated the power of CNN-based methods by directly predicting the actual penetration state or segmenting the molten pool and keyhole area with high accuracy [19–22]. Other architectures are introduced to model long-range dependencies of molten pool image [23,24], or perform time-sequence analysis after the primary feature extraction of the molten pool image [25,26]. Above studies have demonstrated the effectiveness and adaptability of deep learning methods, but a fundamental limitation still persists. The receptive fields of standard CNNs (square kernels) and Vision Transformers (square windows) are isotropic, which leads to an inherent geometric mismatch with the elongated, fusiform geometry of the molten pool, the primary physical information carrier in laser welding process. This architectural mismatch weakens the ability of the model to pay sufficient attention to comprehensive morphology of the pool. There exists a research gap for a specialized deep learning algorithm that explicitly incorporates the physical geometric prior of the molten pool into its structural design.

To address this gap, this paper introduces a novel monitoring algorithm for laser welding penetration state, named Fusiform-Aware Network (FANet). In this work, we call the method "physically-informed" because the network architecture is explicitly constructed based on the physical prior knowledge that the molten pool exhibits a distinct fusiform shape determined by welding dynamics. Molten pool images are captured by a welding camera in a laser welding platform shown in Fig. 1, and then the penetration state is identified by FANet. The core innovation is the Triplet Multi-head Linear Attention (TMLA) mechanism, an attention architecture that aligns the receptive field with the fusiform-shaped molten pool. The TMLA mechanism simultaneously employs local window attention and strip-shaped axial attention, enabling perception on both the local details and overall morphology of the molten pool. Moreover, the TMLA mechanism is adaptively modified to improve the alignment between molten pool area and receptive field by introducing an overlapping attention window partition strategy. The above two design enables FANet to perceive the molten pool as a complete entity just as how a human expert would assess it. Furthermore, Scalar Gated Attention (SGA) is integrated into the basic computation process of the TMLA mechanism to ensure the high inference speed required for online industrial applications. Our contributions can be summarized as follows.

1. A physically-informed visual signal processing algorithm FANet is proposed, featuring the TMLA module to align the receptive field with the elongated shape of the molten pool.
2. An overlapping partition strategy is designed to further improve the adaptability between the TMLA mechanism and geometry feature of molten pool, thus improving accuracy and robustness.
3. The SGA mechanism is implemented to achieve an optimal balance between the accuracy of the measurement and the speed of the inference.
4. Experiments are conducted to validate the performance of the proposed algorithm on a custom laser welding platform, demonstrating its effectiveness over general-purpose vision methods for penetration state online monitoring task.

The overall structure is as follows. The structure of proposed FANet and the principle of components are described in Section 2. Section 3.1 introduces the measurement platform and the acquired dataset. Section 3.2 presents the implementation of the FANet. Sections 3.3 to 3.5 present experimental results and a discussion on effectiveness of the proposed structure. Finally, Section 4 concludes the study.

2. Method

2.1. Overall framework

The core of our proposed monitoring method is named Fusiform Aware Network (FANet), designed to process the visual signals captured by the welding camera. As illustrated in Fig. 2, the overall framework of FANet adopts a four-stage hierarchical structure to form a pyramidal feature representation. Unlike generic vision backbones, FANet explicitly aligns the receptive field with the fusiform-shaped molten pool, which is the most critical visual indicator of the penetration state.

As shown in Fig. 2, the proposed FANet takes a molten pool image $I \in \mathbb{R}^{H\times W\times 3}$ as input. A convolution-style embedding module (two consecutive convolutions of kernel 3 stride 2 followed by normalization with a GELU activation between them) is applied to convert the original image into $H/4 \times W/4$ tokens with C channels. The core of the FANet is composed of four sequential stages, each built upon our newly designed Triplet multi-head Linear Attention (TMLA) Block. These TMLA blocks are the cornerstone of FANet, responsible for capturing both fine-grained local details and long-range dependencies inherent in the molten pool. The TMLA block will be discussed in the next section. The downsample layers (a convolution of kernel size 3 and stride 2) are positioned between stages to reduce the spatial resolution of the feature map by half while doubling the channel dimension. Finally, the feature representation is passed to a classification head, making predictions of the penetration state.

2.2. TMLA: Triplet multi-head linear attention for morphology-aware feature extraction

A key challenge in a vision-based algorithm is to effectively extract features that are most representative of the penetration state. The standard self-attention mechanism excels at global dependencies modeling, but leads to unacceptable computation cost in industrial vision tasks. Swin Transformer is a well-balanced trade-off by restricting self-attention to local windowed areas. While the windowed attention mechanism excels at modeling fine-grained details, it is insufficient

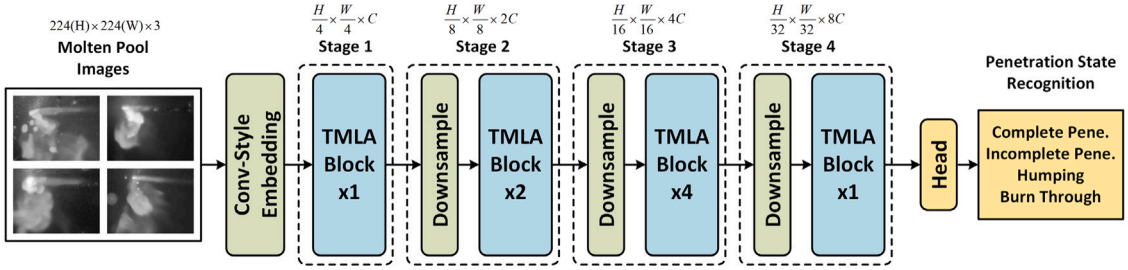

Fig. 2. The overall architecture of the proposed FANet. The model takes molten pool images as input and processes them through four hierarchical stages with several TMLA blocks. Downsampling layers between stages enables a hierarchical feature representation. The final head predicts one of the four penetration states.

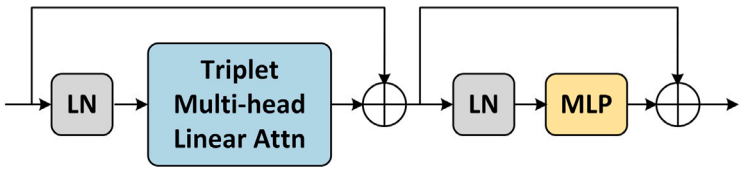

Fig. 3. The architecture of one TMLA block. It follows the structure of an attention block, but replaces the self-attention mechanism with triplet multi-head linear attention mechanism.

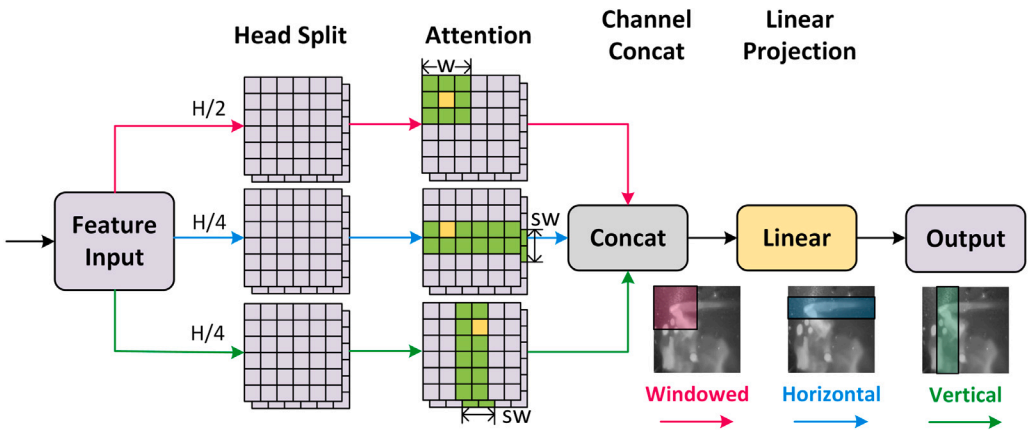

Fig. 4. Standard TMLA mechanism. The attention heads are divided into three branches, one-half for windowed attention and one-fourth each for horizontal and vertical axial attention. The calculation are independent between branches. Attention calculations are performed in window or strip windows, respectively, and the results are concatenated together and projected by a linear layer. (c) The difference when attention branches are partially enabled. The allocation of heads changes accordingly.

for capturing long-range spatial dependencies across the entire molten pool.

Conversely, the axial attention mechanism is better at capturing long-distance dependencies. Its attention range coincides with the elongated fusiform geometry of the molten pool. However, axial attention mechanism pays insufficient attention to localized features, especially in oblique directions.

To resolve the above trade-off, we propose the Triplet multi-head Linear Attention (TMLA) block, a novel hybrid architecture designed to concurrently model features at multiple orientations. By integrating windowed attention with both horizontal and vertical axial attention in parallel order, TMLA is able to form a comprehensive understanding of the molten pool.

As shown in Fig. 3, the TMLA block follows the standard architecture common in various vision transformer models. For an input feature map $X_{l-1} \in \mathbb{R}^{H\times W\times C}$ from the preceding layer, the computation process is defined as follows:

$$\begin{aligned} X' &= \mathrm{TMLA}(\mathrm{LN}(X_{l-1})) + X_{l-1} \\ X_l &= \mathrm{MLP}(\mathrm{LN}(X')) + X' \end{aligned} \tag{1}$$

where LN denotes Layer Normalization operation and MLP represents a linear-GELU-linear structure.

As illustrated in Fig. 4, the key of TMLA is to partition the attention heads into three parallel groups. The windowed attention heads are responsible for local feature extraction, while the horizontal and vertical axial attention heads are responsible for long-range horizontal and vertical feature modeling. Considering that the direction of the molten pool is not limited to one direction in industrial scenarios, both horizontal and vertical attention mechanisms are applied to enhance the robustness of the recognition. In our typical configuration, we allocate half of the heads to windowed attention and a quarter to each of the axial directions:

$$\begin{aligned} N_w &= N/2 \\ N_h &= N_v = N/4 \end{aligned} \tag{2}$$

where N represents number of heads and N_w, N_h, N_v are the heads distributed for windowed, horizontal and vertical branches. The number of channels is evenly distributed across all heads, so the number of channels per head is C/N. The attention for each branch is computed independently.

For windowed attention, the feature map X is divided into non-overlapping windows of size $w\times w$, represented as $[X_w^1, X_w^2, \ldots, X_w^{\frac{W\times H}{w^2}}]$.

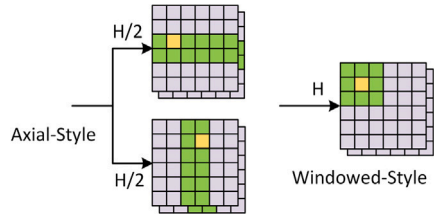

Fig. 5. The difference when attention branches are partially enabled. The allocation of heads changes accordingly.

Self-attention is limited within each window. For self-attention calculation in ith window on nth head, a light-weight self-attention algorithm is executed on X_w^i and the output Y_{wn}^i can be expressed as:

$$Y_{wn}^i = \text{Attention}(X_w^i) \tag{3}$$

The Attention$(\cdot)$ function is the standard scaled dot-product attention, which will be further detailed as a linear-complexity variant in the subsequent section. The attention result on nth head is then reconstructed from all the Y_{wn}^i.

$$\text{W\_Attn}_n(X) = [Y_{wn}^1, Y_{wn}^2, \dots, Y_{wn}^{\frac{W\times H}{w^2}}] \tag{4}$$

For horizontal attention, the feature map is partitioned into non-overlapping horizontal strips of size $sw\times W$, where sw is the strip width and W is the full width of the feature map. This allows contextual modeling across the entire horizontal axis. The feature map X is denoted as $[X_h^1, X_h^2, \dots, X_h^{H/sw}]$. Self-attention is limited within each strip-shaped window. For self-attention calculation in ith strip on nth head, the attention calculation is similar:

$$\begin{aligned} Y_{hn}^i &= \text{Attention}(X_h^i) \\ \text{H\_Attn}_n(X) &= [Y_{hn}^1, Y_{hn}^2, \dots, Y_{hn}^{H/sw}] \end{aligned} \tag{5}$$

Vertical attention is calculated in the same way as horizontal attention.

$$\begin{aligned} X &= [X_v^1, X_v^2, \dots, X_v^{W/sw}] \\ Y_{vn}^i &= \text{Attention}(X_v^i) \\ \text{V\_Attn}_n(X) &= [Y_{vn}^1, Y_{vn}^2, \dots, Y_{vn}^{W/sw}] \end{aligned} \tag{6}$$

The outputs from three parallel branches are then concatenated and fused by a linear layer to restore the original feature dimension. Finally the output of TMLA can be expressed as follows.

$$\begin{aligned} \text{attn}_n &= \begin{cases} \text{W\_Attn}_n(X), 1 \le n \le N_w \\ \text{H\_Attn}_n(X), N_w < n \le N_w + N_h \\ \text{V\_Attn}_n(X), N_w + N_h < n \le N \end{cases} \\ \text{TMLA(X)} &= \text{Channel\_Concat}(attn_1, attn_2, \dots, attn_n)W_P \end{aligned} \tag{7}$$

where W_P is the weight matrix of the final projection layer.

To provide a clearer understanding of the information flow within the TMLA block, the detailed procedure is summarized in Algorithm 1. The input are split into three groups along the channel dimension, corresponding to three branches. Each branch performs attention calculation independently within its specific partition strategy. Finally, the outputs are concatenated along the channel dimension and passed through a linear projection layer to fuse the multi-view features.

This algorithm allows the TMLA block to be configured into simplified variants, as shown in Fig. 5. By allocating all heads to the axial branches, it becomes a purely axial-style attention block. Also, allocating all heads to windowed branch results in a standard windowed attention. These variants can be used in architectural implementations to emphasize long-range or localized dependency modeling.

Algorithm 1 Pseudocode for TMLA Mechanism

Input: Feature map x (B, C, H, W)
Output: Refined feature map y

{Split heads for three branches}
$x_w, x_h, x_v = \text{split}([N_w, N_h, N_v], \text{dim} = 1)$
{Branch 1: Window Attention}
for x_{wi} in WindowPartition(x_w)
    $o_w = \text{Attention}(x_{wi})$
$o_w = \text{WindowReverse}(o_{wi})$
{Branch 2: Horizontal Strip Attention}
for x_{hi} in HStripPartition(x_h)
    $o_{hi} = \text{Attention}(x_{hi})$
$o_h = \text{HStripReverse}(o_{hi})$
{Branch 3: Vertical Strip Attention}
for x_{vi} in VStripPartition(x_v)
    $o_{vi} = \text{Attention}(x_{vi})$
$o_v = \text{VStripReverse}(o_{vi})$
{Fusion}
$y = \text{Concat}([o_w, o_h, o_v], \text{dim} = 1)$
$y = \text{LinearProj}(y)$
return y

2.3. Overlapping partition strategy for strip-shaped windows

The strip-shaped windowed attention suits the elongated fusiform geometry of the molten pool well. However, a limitation remains regarding robustness: the strip width is typically constrained to be divisible by the feature map dimensions. The size of the molten pool is a dynamic physical characteristic, not an artifact of the network architecture. An overlapping strip-shaped window partition strategy is applied in TMLA to address the inconsistency.

The partition of strip-shaped windows is no longer constrained by the resolution of feature maps by introducing the concept of sliding windows. A new parameter stride is introduced to control the distance between two adjacent strip-shaped windows. Given a feature map X, horizontal strip width sw and stride s, the partition result can be described as follows.

$$X_i = \begin{cases} \text{row}_X(s(i-1)+1, s(i-1)+sw+1) \\ \text{row}_X(H-sw+1, H), \text{if no enough space} \end{cases} \tag{8}$$

where X_i stands for ith strip-shaped windows and $\text{row}_X(x, y)$ refers to all rows between the x-th row and the y-th row of X, inclusive.

A counter tensor D is applied to count how many times each pixel is covered by a patch. The actual output X' is finally given by an element-wise normalization $X' = X_{sum}/D$ to resolve the magnitude inconsistency issue caused by overlapping in reconstruction. The same procedure is applied for vertical strips.

The selection of strip width sw and stride s is determined by the physical characteristics of the molten pool in our dataset. Statistical analysis of the training data reveals that the width of the molten pool typically occupies between 9% and 13% of the image height. Taking Stage 3 (resolution 14×14) as an example, a standard non-overlapping partition with $sw = 2$ theoretically covers the pool. However, such a rigid division is prone to cutting the molten pool into separate strips, especially when the orientation of the pool is not parallel to the grid.

To address this, we adopt the overlapping strategy with a wider receptive field. We set the strip width to 4 and stride to 2. This design ensures: (1) At least one strip can fully cover the molten pool without truncation; (2) The strip is not overly wide, thereby avoiding the inclusion of excessive background noise which could distract the attention score, as illustrated in Fig. 6. This configuration offers a balanced trade-off between feature completeness and noise suppression compared to the non-overlapping approach used in architectures like CSwin Transformer.

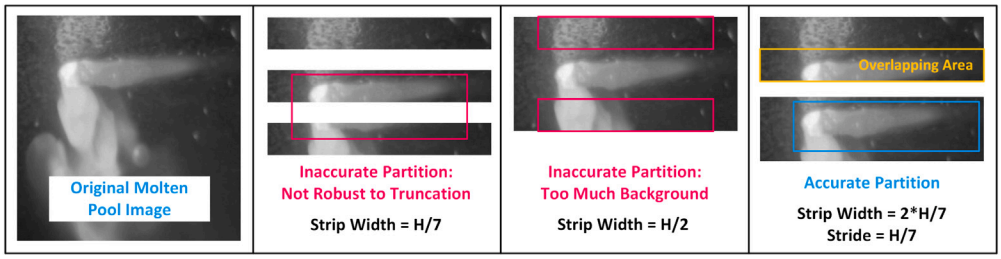

Fig. 6. An example of strip partition on stage 3. Reasonable strip width under the original partition method is H/7 or H/2, but these divisions are inaccurate for molten pool images. After introducing overlapping partition, the strip width can be set to any positive integer and can be designed specifically for the molten pool target.

The overlapping partition strategy allows free selection of strip width and overlap ratios at each stage of the network, adjusting the receptive field to match the molten pool size. Also the overlap ensures that features at the boundaries of one strip are also processed in an adjacent strip, forming a more robust and continuous representation of the molten pool structure. The primary cost of this approach is a slight increase in computational burden, which is controllable under the design of linear attention mechanism, which will be discussed in the following section.

2.4. SGA: Scalar Gated Attention

An online monitoring system should be efficient in computation for industrial application. The standard self-attention mechanism suffers from quadratic complexity with respect to the number of input tokens. Although the windowed and axial attention mechanisms limit the input token into a localized area, they are still not efficient enough for online monitoring tasks, especially when large window size and strip width are required for wide receptive fields. To overcome this challenge, an efficient linear complexity attention mechanism named Scalar Gated Attention (SGA), is integrated into TMLA blocks. This design uses approximate attention calculations to significantly reduce the amount of computation required without significantly affecting overall performance.

Consider an input tensor $X \in \mathbb{R}^{B\times N\times d}$ representing a group of tokens from a standard window or a strip-shaped window, where B is the batch size, N is the number of tokens, and d is the feature dimension on a single head. X is projected into three distinct representations: a concentrated information tensor $I \in \mathbb{R}^{B\times N\times 1}$, a key tensor $K \in \mathbb{R}^{B\times N\times d}$, and a value tensor $V \in \mathbb{R}^{B\times N\times d}$.

$$I = XW_I, \quad K = XW_K, \quad V = XW_V \tag{9}$$

where W_I, W_K, W_V are the learnable weight matrix. Here, the Information Tensor I serves as a compressed abstraction of the input features. Intuitively, I acts as a global summary vector that gates the value features based on aggregated context, allowing the network to selectively emphasize informative features while suppressing irrelevant background. Unlike standard self-attention where queries interact with keys pixel-by-pixel, I is designed to aggregate global context information into a compact representation. A softmax function is then applied to the information tensor, which is augmented with a learnable relative position encoding bias to embed spatial information in pixels.

$$S = \text{softmax}(I^T + B) \tag{10}$$

where $S \in \mathbb{R}^{N\times 1}$ represents the attention score for each pixel. These score vector are then used to compute a single, aggregated context vector by taking a weighted sum of K. The result of SGA is obtained through an element-wise Hadamard product between this context vector and V. The complete process can be formulated as:

$$\begin{aligned} &C_v = S \otimes K, C_v \in \mathbb{R}^{1\times C} \\ &\text{SGA\_Result} = C_v \odot V \end{aligned} \tag{11}$$

where C_v denotes the context vector, $\otimes$ denotes matrix multiplication and $\odot$ denotes element-wise Hadamard product. The resulting context vector C_v acts as a global scalar gate. It dynamically modulates the feature values V based on the global information. This mechanism allows the network to selectively emphasize informative features while suppressing irrelevant background, all with linear computational complexity.

The primary advantage of SGA mechanism is the computational efficiency. The scalar-gated mechanism avoids generating an $N \times N$ query matrix and performing the corresponding matrix multiplication. Consequently, the overall complexity is reduced from $HW(6C^2+4NC)$ to $HW(4C^2+5C)$, where N represents the tokens in one window. The elimination of quadratic complexity item not only significantly reduces the total computation burden, but also achieves a linear relationship between complexity and total pixels. It allows the use of large strip-shaped attention regions without extra cost, which are essential for modeling the complete geometry of the molten pool. The design of SGA mechanism achieves an optimal balance between high identification accuracy and high inference speed.

3. Experiment

3.1. Experimental setup

Experiment Platform: To implement and validate the proposed vision based monitoring system, a laser welding experiment platform is constructed, as shown in Fig. 1. The platform consists of a welding platform and a data acquisition system. The welding platform includes a YLS-10000 fiber laser capable of generating laser pulses with a maximum power of 10 kW and a KUKA welding robot for precise control of the welding position. A 99.99% argon cylinder delivers shielding gas through a nozzle in the laser head to prevent oxidation of the weld seam in the welding process. Q235 carbon structural steel is the material used in welding experiments. Data acquisition is done using an Xiris welding camera that captures videos of the molten pool with a resolution of 1280 × 1024 at a maximum frame rate of 40 fps. A 660 nm bandpass filter is installed in front of the camera. The welding camera is set at a distance of 260–300 mm from the molten pool with a pitch angle of 30–35 ° through a fixture. The camera setup can capture a side view of the molten pool and reduce the effect of the occlusion caused by strong plasma plume in the welding process.

Data Acquisition: Data acquisition is performed at a frame rate of 40 fps with a raw resolution of 1280 × 1024. The raw images are cropped to 800 × 640 to focus on the region of interest. The experiments are designed to cover various penetration states by adjusting key welding parameters, primarily laser power and material thickness. Due to the stochastic nature of the laser welding process, identical parameter combinations can yield different penetration results. Therefore, 137 welding experiments are conducted to ensure the diversity of welding penetration states.

Data Labeling and Description: As illustrated in Fig. 7, four welding penetration states are defined based on the morphology of the weld

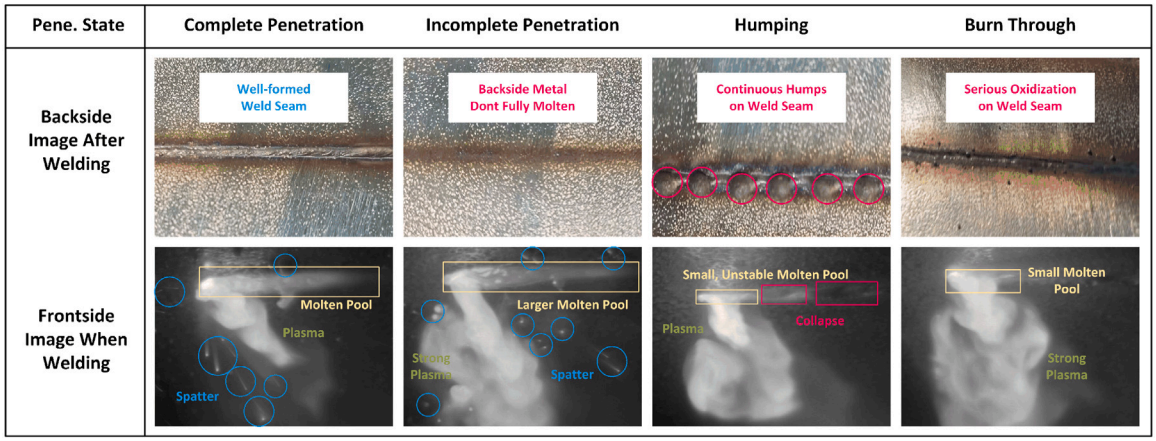

Fig. 7. Comparison of weld seams on the backside between different penetration states (first row). Corresponding molten pool images are listed in the second row. Penetration states include complete penetration (no defect), incomplete penetration, humping and burn through. Backside weld seams in different penetration states have distinctive morphology characteristics, while corresponding molten pool images have features such as molten pool (yellow box), plasma plume coverage, spatters (blue circle) and weld seam forming (magenta box).

Table 1
Typical combinations of laser welding parameters of four penetration states. The power includes values within a range of plus or minus 200 W. A smaller index means a more frequent occurrence in one penetration state.

Index	Laser power (kW)	Material thickness (mm)	Pene. state
1	7.0	6	Complete Pene.
2	7.5	8	Complete Pene.
3	6.5	6	Complete Pene.
4	7.0	8	Incomplete pene.
5	6.5	8	Incomplete pene.
6	6.0	6	Incomplete pene.
7	7.0	8	Humping
8	6.5	8	Humping
9	6.5	4	Burn Through
10	9	6	Burn Through

Welding Speed: 1.2 m/min Shielding Gas Flow: 20 L/min, 25 L/min for power ≥ 7 kW Laser Angle: 5° – 7° Defocus: 0 mm.

Table 2
Data distribution of numbers of images among four penetration states.

Pene. state	Training set	Test set	Sum
Complete Pene.	8181	2905	11 086
Incomplete Pene.	7172	2519	9691
Humping	2767	1092	3859
Burn Through	7705	2848	10 553
Sum	25 825	9364	35 189

seam on the backside of the material after the welding process. Different penetration states are achieved by precise control of key welding parameters including laser power and material thickness. However, identical combinations of welding parameters can produce different penetration states because of inherent uncertainty in laser welding process. Therefore, multiple laser welding experiments are conducted under different parameter combinations to ensure that the obtained data can cover a sufficient number of situations. Typical combinations of welding parameters applied in welding experiments for different penetration states are listed in Table 1.

The actual penetration state is determined by the formation of the weld seam on the backside after the welding process. Complete penetration refers to a well-formed weld seam on the backside without any defects. For incomplete penetration, the backside of weld seam is not covered by molten metal, indicating that the heat input is insufficient. The humping state refers to continuous small metal humps formed on the weld seam and is an intermediate state between incomplete penetration and complete penetration. The burn through state is recognized by severe oxidization around the weld seam, suggesting excessive heat input.

The frontside molten pool images also contain specialized features, including the shape, length and width of the molten pool, spatter, and the formation of weld seam right after the molten pool. The long-range semantic information in welding images includes the morphology of the molten pool and the formation weld seam, while the detailed information includes features such as the edge morphology of the molten pool, the shape of the tail of the molten pool, details on weld seam, and spatters. Both long-range semantic features and local details contain unique information corresponding to the welding state, and therefore can be utilized to identify the penetration state.

Dataset Splitting: To ensure a fair evaluation and prevent data leakage caused by the high similarity between consecutive frames, we adopt a temporal block splitting strategy rather than random frame shuffling. For each welding process, the first 20 frames and last 10 frames are discarded first. The video is then segmented based on the transition points of the penetration states. A total of 348 video clips are generated. Each video segment is divided into a training block and a testing block. Typically, 60%–70% of the continuous sequence are assigned to the training set, 10% are discarded as a buffer block, and the remaining portion is assigned to the test set. The order of training and testing blocks is randomly swapped to avoid bias. Finally, 25,825 images are split into the training set and 9364 into the test set. The data distribution of the training and test sets is shown in Table 2. 20% of the training set is randomly selected as the validation set before training, which is not shown in the table.

Dataset Imbalance: The amount of the Humping defect is lower than that of other states. This is because of the physical reality of the welding process: Humping is a unstable transition state. It typically occurs randomly and persists for short durations, resulting in naturally fewer captured frames compared to other states. Though the amount of Humping defect is imbalanced, we find that the overall performance fluctuations are negligible (< 0.3%) by applying data resampling techniques or weighted loss function, as long as the amount of minority class is enough (> 1500 images for training). Consequently, the standard Cross-Entropy Loss is employed without additional balancing weights, as the distinct visual features of the defects are sufficient for the model to learn robust feature boundaries.

Table 3
Detailed structural and dimensional hyperparameter configuration of FANet. The strip width and stride in Stage 3 are determined by statistical analysis showing the molten pool width typically occupies 9%–13% of the image height, ensuring the receptive field fully encapsulates the pool geometry.

Parameter	Stage 1	Stage 2	Stage 3	Stag e 4
Feature Dimension	64	128	256	512
Depth	1	2	4	1
Attention Style	window	window-full	window-full	full
Window Size	7	7	7	7
Strip (width, stride)	/	(2,2)	(4,2)	(4,3)
MLP expansion	1.0	1.0	1.0	1.0

3.2. Implementation details

The core identification algorithm FANet is implemented in the Python and Pytorch framework. All experiments are conducted on an NVIDIA® GeForce RTX™ 4080 GPU with 16 GB VRAM. The software environment is based on Pytorch version 2.4.0 and CUDA version 12.4. For all experiments, our proposed FANet model is trained from scratch for 75 epochs. The AdamW optimizer is applied with an initial learning rate of 2×10^{-4} and a weight decay of 0.05. The learning rate is scheduled by a cosine annealing scheduler with a minimum learning rate of 1×10^{-6}. A batch size of 16 is used. The standard Cross-Entropy Loss is employed as the objective function.

During training, all input images are resized to 224×224 pixels. A set of data augmentation techniques are employed, including random horizontal and vertical flipping with a probability of 0.5, random rotation within a range of $[-20, 20]$ degrees, and random brightness and contrast adjustment with a maximum magnitude of 0.2. These data enhancement techniques enrich the data without compromising the integrity of the molten pool, and reduce the bias that might be introduced by the setup of the data acquisition platform.

To effectively balance local feature aggregation and long-range dependency modeling, a specific attention strategy is employed within the TMLA blocks at different stages. The window-style variant of the TMLA block is used in the first block of stage 1, 2 and 3. The subsequent blocks and stages utilize the full TMLA structure, which concurrently processes information through windowed and two axial attention branches. Besides, the expansion ratio of hidden dimension in the MLP structure in the TMLA block is reduced to 1.0 to form a lightweight design. The complete configuration of FANet is listed in Table 3.

3.3. Penetration state monitoring performance evaluation

The performance of the proposed FANet-based monitoring method is evaluated and compared with several vision algorithms. As shown in Table 4, the selected vision algorithms include classic CNNs (ResNet [27], ConvNeXt [28]), ViTs (Swin-Transformer [29], CSwin-Transformer [30], CAFormer [31]), and lightweight ViTs (EfficientViT [32], FastViT [33]). Hybrid temporal models (CNN-BiLSTM, CNN-TCN), designed for time-series analysis and widely used in weld process monitoring, are also included for comparison. All models are trained from scratch on our custom dataset to ensure a fair comparison. The temporal models are trained with a sequence length of 12 to capture dynamics of the welding process. The results are analyzed in terms of accuracy, per-class capability, and computational efficiency.

Overall Performance. The results in Table 4 demonstrate the superior capability of our FANet. It presents the best performance on all primary evaluation metrics, with an Accuracy of 93.24% and an F1-Score of 91.55%. This result exceeds the next best performed models CAFormer (89.42%) and CSwin (89.30%) by 2.13 and 2.25 points on F1-Score. This highlights the effectiveness of the fusiform-aware design in extracting molten pool information. Additionally, the time-series model CNN-TCN exhibits competitive accuracy by leveraging temporal dependencies. However, FANet still outperforms CNN-TCN by 1.01% in accuracy and 3.01% in F1-Score. This suggests that FANet's spatial geometry modeling is more effective at capturing critical defect features than the temporal dependencies built by convolutions in this scenario.

Efficiency and Application Feasibility Analysis. For an online industrial monitoring system, low latency is as important as high accuracy. Considering a typical laser welding scenario, the monitoring system generally requires a frame rate at 100 FPS (i.e., latency < 10 ms) to provide effective feedback control at a precision of submillimeter. Table 5 reports the efficiency metrics, where "End-to-End Latency" accounts for the entire pipeline from image preprocessing to result output.

FANet achieves an end-to-end latency of 5.41 ms, which is well within the real-time operational margin. In contrast, although CNN-BiLSTM and CNN-TCN perform well in classification, their latencies (15.37 ms and 18.61 ms) are significantly higher due to the overhead of processing image sequences, making them less suitable for online monitoring. Moreover, while generic lightweight models like EfficientViT (0.27G MACs) are theoretically efficient by simplifying attention mechanisms, they lack the specific design to model specific geometric shapes effectively. In the context of laser welding, the penetration state is strongly correlated with the morphology of the molten pool. This performance gap indicates that generic lightweight architectures lack the specialized structure necessary to capture critical welding defects contained in pool shape, keyhole, weld seam, etc. Although FANet has higher MACs than the most aggressive lightweight models, a significant gain in accuracy and robustness is worthwhile.

Per-Class Capability. A deeper look into per-class capability reveals the robustness of the monitoring algorithm across different conditions. FANet achieves the highest or near-highest accuracy in three categories. Notably, FANet achieves a roughly 2 point advantage on Complete Penetration state and the most difficult category Humping defect, compared to the second-best result. Though temporal models outperform on Incomplete Penetration defect, their overall performance is not balanced on other categories because of insufficient analysis on spatial information. Results indicate that the FANet has successfully learned the distinct visual signals for each penetration state, even under difficult defect conditions.

Architectural Advantage Analysis. The success of FANet can be attributed to its hybrid attention mechanism, which integrates the strengths of the Swin-Transformer (windowed) and the CSwin-Transformer (cross-shaped). While Swin-Transformer validates the efficacy of local attention on fine-grained details, CSwin-Transformer, on the other hand, introduces axial attention to model long-range dependencies. FANet integrates the advantages through the TMLA mechanism and further reinforces it with the overlapping partition method and the application of simplified attention variants. The windowed branch consolidates essential local understanding, while the parallel horizontal and vertical strip branch are perfect for the elongated shape of the molten pool. This carefully designed mechanism allows FANet to achieve a comprehensive awareness of the fusiform-shaped molten pool.

In summary, above comparison validates that FANet provides an optimal trade-off between accuracy and speed, outperforming both general-purpose lightweight models and specialized time-series approaches, making it suitable for real-world online application.

3.4. Ablation studies

To understand the contribution of each key component of our algorithm to the overall performance, a series of ablation studies are conducted. These experiments explain the impact of the TMLA mechanism and the strip partition strategy on the final recognition accuracy.

Table 4
Defect monitoring performance comparison between general vision models and our proposed FANet. The best result in each column is in **bold**, and the second-best is underlined. ACC. is accuracy, PRE. is precision, REC. is recall, F1. is F1-score.

Method	Acc.	Pre.	Rec.	F1.	Comp.Pene.	Incomp. Pene.	Humpimg	Burn through
ResNet18[27]	88.69	86.95	86.62	86.78	90.46	87.81	75.46	92.73
ConvNeXt-T [28]	90.50	88.38	88.66	88.52	87.26	90.47	78.48	**98.42**
Swin-T [29]	90.43	88.69	87.85	88.27	90.22	89.36	73.90	97.93
CSwin-T [30]	91.25	89.18	89.43	89.30	89.43	90.91	79.40	97.96
CAFormer-S18[31]	91.29	89.44	89.40	89.42	91.12	90.99	79.12	96.38
EfficientViT-M3[32]	89.76	87.91	88.15	88.03	86.30	90.35	79.12	96.84
FastViT-t12[33]	90.44	88.05	88.70	88.38	88.78	90.83	79.03	96.17
CNN-BiLSTM (seq-len=12)	90.40	85.52	88.22	86.85	82.90	**97.99**	79.38	92.59
CNN-TCN (seq-len=12)	92.23	87.32	89.79	88.54	90.64	96.45	79.38	92.69
FANet (Ours)	**93.24**	**91.68**	**91.41**	**91.55**	**93.05**	93.09	**81.41**	98.10

Table 5
Calculation efficiency performance comparison between general vision models and our proposed FANet. The best result in each column is in **bold**, and the second-best is underlined. "bs" means batch size.

Method	Params (M)	MACs (G)	Throughput (img/s, bs=16)	Throughput (img/s, bs=1)	End-to-End latency (ms)
ResNet18[27]	11.18	1.82	**8318**	775	**4.13**
ConvNeXt-T [28]	27.82	4.47	2212	295	5.77
Swin-T [29]	27.52	4.51	1986	138	9.21
CSwin-T [30]	21.81	4.34	1162	57	14.68
CAFormer-S18[31]	24.30	4.12	1862	137	8.16
EfficientViT-M3[32]	6.58	**0.27**	3300	192	6.72
FastViT-t12[33]	6.53	1.09	3406	268	5.63
CNN-BiLSTM (seq-len=12)	14.33	21.92	2313	**960**	15.37
CNN-TCN (seq-len=12)	24.82	49.60	1228	789	18.61
FANet (Ours)	**4.40**	0.80	4652	348	5.41

Table 6
Impact of attention style on performance. Models using one attention style and FANet-variants are compared with FANet. Acc is accuracy, F1. is F1-score.

Model	Window	Axial	Triplet	Acc.	F1	GMACs
M_W	✔			90.54	88.62	**0.76**
M_A		✔		91.10	89.24	0.84
M_T			✔	92.13	89.91	0.81
M_{NVS}	✔		No VStrip	92.67	91.09	0.79
M_{NHS}	✔		No HStrip	88.42	86.11	0.79
FANet	✔		✔	**93.24**	**91.55**	0.80

Table 7
Analysis on different strip partition strategy on stage 3 and 4 of the FANet. Acc is accuracy, F1. is F1-score.

Strip partition	Width	Overlap	Acc.	F1	MACs (G)
Overlap	4	Each	**93.24**	**91.55**	0.80
Minimal-overlap	4	Last one	92.75	90.87	0.78
CSwin-style	7	None	92.26	90.48	**0.77**

3.4.1. Effectiveness of the triplet attention mechanism

This study investigates how the design of attention mechanism influences the recognition capabilities. As shown in Table 6, three models with one single attention style, as well as three variants of FANet are compared to the proposed FANet.

The influence of attention mechanism is studied by simplifying the attention blocks to one pure style. The M_W using only local windowed attention achieves a baseline F1-score of 88.62%. By switching to long-range axial attention, M_A improves the score to 89.24%, confirming that global morphological information is crucial for accurate recognition. M_T with proposed triplet attention fuses both local and global information and further boosts the performance, proving that concurrent usage of different attention forms is key to comprehensive understandings of molten pool image. The final FANet architecture, which combines the windowed attention in the early stage or early block with the powerful triplet attention, achieves the highest accuracy. The difference between M_T and FANet indicates that local details serve as the cornerstone for a comprehensive global semantic understanding. Above result validates that the design of physically informed attention is superior for this specific monitoring task.

The necessity of bidirectional axial attention is also studied. The M_{NVS} variant removes the vertical strip-shaped attention from the original TMLA structure and yields a competitive performance with FANet. This performance is caused by the data distribution of our dataset, where most of the molten pool images are nearly horizontal. However, the M_{NHS} variant, which removes the horizontal strip-shaped attention, suffers a significant performance drop due to the misalignment between the shape of attention window and the area of molten pool. The contrast reveals that while single-directional axial attention may suit specific datasets, it is vulnerable in industrial applications. In real-world scenarios, the orientation of the welding path and workpiece varies unpredictable. The design of bidirectional strips is essential for a robust and reliable monitoring algorithm.

3.4.2. Analysis on strip partition strategy

This experiment validates how the geometry of the receptive field impacts recognition accuracy. Three attention window division strategies are tested and the result is summarized in Table 7. The first strategy, overlap, is the one adopted in the proposed model. It employs a strip width of 4 with a stride of 2 in Stage 3 and a stride of 3 in Stage 4, ensuring a uniform division of attention windows. The second strategy, minimal-overlap, maintains the strip width of 4 but set stride equals to width, avoiding overlap except the last strip due to feature map dimensions. The third strategy, CSwin-style, emulates the approach in CSwin-Transformer by setting the strip width to 7 without overlap.

As shown in Table 6, the proposed overlap strategy achieved the best performance, reaching an accuracy of 93.24% and an F1-score of 91.55%. It outperformed the minimal-overlap approach by 0.68% in F1-score, suggesting that the overlapping regions enhances the model's

Table 8
Comparison on performance between the SGA and standard attention.

Attention type	Acc.	F1	Params (M)	MACS (G)	Latency (ms)
SGA	93.24	91.55	4.40	0.80	5.41
Standard	93.45	91.78	5.38	1.05	7.63
Difference	−0.21	−0.23	−18.2%	−23.8%	−29.1%

Table 9
Performance on dataset with different rotation augmentations.

Orientation	Acc.	Pre.	Rec.	F1
Original	93.24	91.68	91.41	91.55
Orthogonal	92.74	91.02	91.05	91.03
Omnidirectional	91.05	88.84	88.72	88.78

ability to capture molten pool features that are split across strip boundaries. This performance gain increases computational cost by only 0.02 GMACs. Furthermore, the overlap method surpassed the CSwin-style strategy by a significant margin of 1.07% in F1-score. This improvement suggests that a strip with moderate width is better suited for a whole molten pool because a wider strip may include excessive background and noise. Results indicate that a carefully designed overlapping strip partition strategy minimizes boundary artifacts and ensures accurate coverage of the molten pool target, which is critical for sensitivity and accuracy in penetration state monitoring tasks.

3.4.3. Analysis on SGA mechanism

To evaluate the contribution of the Scalar Gated Attention (SGA) mechanism towards calculation efficiency, we conduct a comparative analysis by replacing the SGA module in FANet with the standard Scaled Dot-Product Attention while keeping other hyperparameters unchanged. The results are summarized in Table 8.

As observed, the standard attention mechanism yields a minor improvement in performance, with Accuracy and F1 scores increasing by 0.21% and 0.23%, respectively. However, this slight gain comes at an increase in computational cost and latency. The proposed SGA mechanism reduces the parameter by 18.2% and MACs by 23.8%. Moreover, SGA reduces the inference latency from 7.63 ms to 5.41 ms, representing a 1.4× improvement in FPS. The trade-off between 0.21% drop in accuracy and a 29.1% saving in processing time is favorable for online laser welding monitoring scenarios where low latency is required.

3.4.4. Robustness evaluation on welding orientation

The primary dataset consists of horizontally aligned molten pools due to the specific camera mounting in our industrial setup. To validate that the proposed bidirectional axial attention mechanism ensures robustness against orientation variations, we conducted two additional experiments using geometric data augmentation. The result is listed in Table 9.

Orthogonal rotation: We retrain and test the FANet with a dataset where images are randomly rotated by 90° or 270° with a 50% probability. This setup simulates a scenario where the welding path includes both horizontal and vertical directions. The results showed a minor accuracy drop of approximately 0.5% compared to the baseline. This suggests that the V-Strip attention branch effectively activates when the fusiform molten pool aligns vertically.

Omnidirectional rotation: To further assess the model's performance under irregular welding paths, we applied omnidirectional random rotation to the dataset. The performance decrease is observed to be around 2%–3%. This slightly higher drop is attributed to the orthogonal design of the strip attention, which is optimized for axis-aligned features. However, the model still maintains a high performance, demonstrating that the triplet attention mechanism retains its discriminative power under non-orthogonal conditions.

In summary, although the rotation is not a perfect substitution to real welding data, these simulation results confirm that FANet is capable of handling multi-directional welding tasks. Future work will focus on validating the model's performance on real-world welding paths with appropriately collected experimental data to further substantiate its industrial robustness.

3.5. Visualization of the TMLA mechanism

To provide intuitive insight into the working principle of the TMLA algorithm, visualization on the final activation and attention on three distinct branches is conducted as shown in Fig. 8. The visualization results serve as a qualitative evidence to reveal how the algorithm selectively focuses on the most informative areas of the input visual signal to make decisions about actual penetration states. Each row picks an example of a specific category, while the columns deconstruct the attention from different components.

Firstly, the second column shows the final activation before the classification head. The activation is correctly concentrated on the molten pool area, suggesting that FANet successfully identifies the most critical area for the penetration state classification task. The subsequent columns then reveal the contributions of each attention branch.

The third column illustrates the behavior of the windowed attention branch. The heatmaps are relatively diffuse and show activations on localized features, including the distribution of spatters and the plasma plume. These dynamic features are indicators of process stability and keyhole status. For instance, the attention map shows that the model effectively captures the excessive spatter and chaotic plume patterns typical in Incomplete Penetration states, while the suppressed spatter and weak plume characteristics are often observed in Humping states.

In contrast, the horizontal strip column (column 4) validates our core hypothesis regarding geometric alignment. The horizontal strip-shaped attention generates an accurate activation band aligns with the molten pool. This branch focuses on the overall morphology of the molten pool, extracting critical geometric parameters such as pool length, width, boundary edges, keyhole, and tail shape. Since the geometric dimensions of the molten pool are direct responses to heat input and thermal diffusion, this branch is effective in distinguishing welding states with distinctive molten pool geometric features, e.g., the incomplete state usually has a large molten pool.

The vertical strip column (column 5) shows the contribution of the last branch. Although the activation is less precise because of the misalignment with the pool's main axis, it provides complementary information. As observed in the visualization, this branch tends to observe the weld pool tail and the solidified weld seam right after the tail. Capturing the formed weld seam helps verify the final state of the molten area, acting as a secondary validation to the molten pool features.

In summary, the algorithm learns to integrate the shape-aware focus of the stripe branches with the detailed understanding of the windowed branch. This effective fusion allows FANet to form a comprehensive and robust representation of the molten pool, which is fundamental for its superior performance on penetration state monitoring task.

4. Conclusion

In this paper, a novel vision-based algorithm is designed for online monitoring of penetration state in laser welding. The core innovation lies in its physically informed visual signal processing algorithm, the Fusiform-Aware Network (FANet), which is specifically designed to align with the geometric characteristics of the molten pool. By introducing the Triplet Multi-head Linear Attention (TMLA) mechanism, FANet effectively integrates long-range, shape-aware feature extraction with localized detail analysis. The scalar-gated attention design ensures that the system achieves high accuracy without compromising the high throughput required for real-time industrial applications.

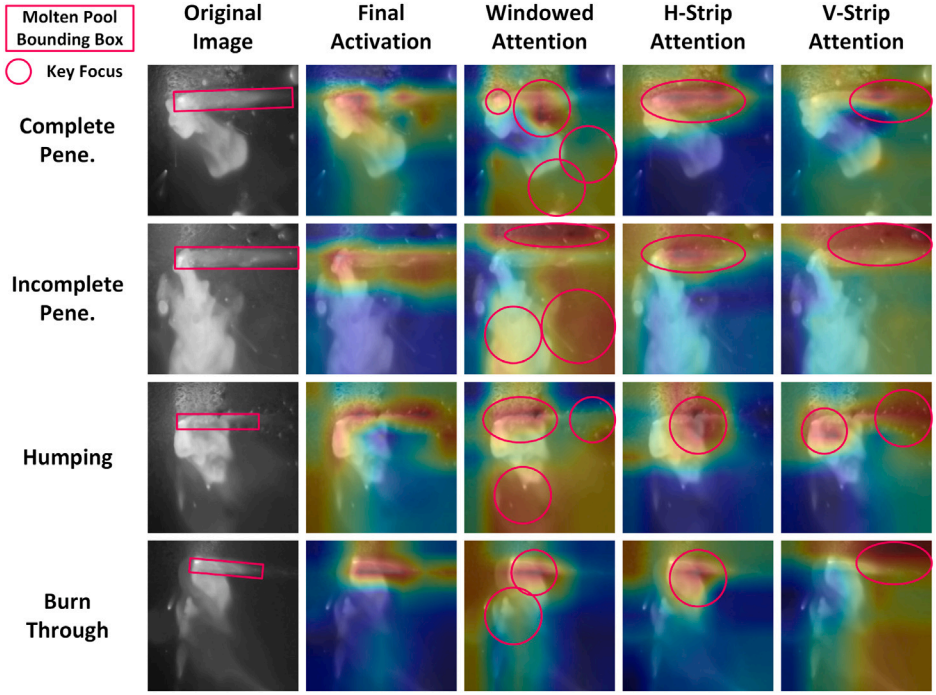

Fig. 8. Visualization on activations of the final activation and attention on different branches on molten pool images. Pixels with higher red saturation indicate a higher sensitivity of the model, while pixels with a more bluish color denote reduced sensitivity. The key focus of three branches are circled in the heatmap. The Window branch (Col 3) highlights localized features like molten pool, spatters and plumes. The H-Strip branch (Col 4) captures the primary morphology and length of the molten pool. The V-Strip branch (Col 5) provides complementary attention to the pool edges and the solidified weld seam.

In laser welding visual monitoring systems, the primary sources of uncertainty include the random obscuration from plasma plume, strong light emissions, reflections of the workpieces and imaging noise. Conventional visual algorithms exhibit significant performance fluctuations when confronted with these disturbances. Our proposed FANet algorithm, particularly its overlapping partitioning strategy and adaptive focus mechanism on the molten pool morphology, is designed to enhance robustness against these disruptive factors.

Comprehensive experiments conducted on laser welding platform confirm the superior performance of our proposed FANet, which achieved an accuracy of 93.24% and surpassed various general-purpose vision architectures. Visualization of the attention mechanism provided intuitive evidence of the principle of TMLA mechanism, demonstrating its ability to focus precisely on the fusiform-shaped molten pool. This work underscores the significant potential of aligning deep learning architectures with the physical priors of the target, and provides a reliable solution for quality control in industrial laser welding manufacturing.

CRediT authorship contribution statement

Songlin Li: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation. **Ting Yuan:** Writing – review & editing, Supervision, Formal analysis. **Jiawei Fan:** Visualization, Data curation. **Haonan Zhang:** Validation, Data curation. **Zhuguo Li:** Project administration, Funding acquisition. **Uwe D. Hanebeck:** Writing – review & editing, Validation. **Jiuchao Qian:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

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

Acknowledgments

This work is funded by Shanghai central guided local science and technology development fund, No. YDZX20233100004030004.

Data availability

Data will be made available on request.

References

[1] Angel-Iván García-Moreno, A fast method for monitoring molten pool in infrared image streams using gravitational superpixels, J. Intell. Manuf. 33 (6) (2022) 1779–1794, http://dx.doi.org/10.1007/s10845-021-01761-8.

[2] Yi-Wei Huang, Xiang-Dong Gao, Perry P. Gao, Bo Ma, Yan-Xi Zhang, Laser welding monitoring techniques based on optical diagnosis and artificial intelligence: A review, Adv. Manuf. 12 (1) (2024) 1–25, http://dx.doi.org/10.1007/s40436-024-00493-1.

[3] Tao Sun, Zhengjie Fan, Xiaomao Sun, Yichun Ji, Wanqin Zhao, Jianlei Cui, Xuesong Mei, Femtosecond laser drilling of film cooling holes: Quantitative analysis and real-time monitoring, J. Manuf. Process. 101 (2023) 990–998, http://dx.doi.org/10.1016/j.jmapro.2023.06.059.

[4] Wenqi Cui, Kechen Song, Yanyan Wang, Guotong Lv, Yunhui Yan, Han Yu, Xingjie Li, A rapid screening method for suspected defects in steel pipe welds by combining correspondence mechanism and normalizing flow, IEEE Trans. Ind. Inf. 20 (9) (2024) 11171–11180, http://dx.doi.org/10.1109/TII.2024.3399934.

[5] Zhenying Xu, Rong Wang, Rong Zuo, Prediction of weld penetration status based on sparse representation in fiber laser welding, Proc. 9th Int. Symp. Precis. Mech. Meas. 11343 (2019) 412–417, http://dx.doi.org/10.1117/12.2548813.

[6] Wei Wei, Yang Liu, Jindou Wu, Zhilin Wei, Zhukun Zhou, Yu Long, In-situ monitoring method of femtosecond laser welding between glass and copper with acoustic emission, Measurement 240 (2025) 115568, http://dx.doi.org/10.1016/j.measurement.2024.115568.

[7] Jing Huang, Zhifen Zhang, Rui Qin, Yanlong Yu, Guangrui Strategy Wen, Wei Cheng, Xuefeng Chen, Lightweight neural network architecture for pipeline weld crack leakage monitoring using acoustic emission, IEEE Trans. Instrum. Meas. 72 (2023) 1–10, http://dx.doi.org/10.1109/TIM.2023.3298393.

[8] M.F.M. Yusof, M. Ishak, M.F. Ghazali, Weld depth estimation during pulse mode laser welding process by the analysis of the acquired sound using feature extraction analysis and artificial neural network, J. Manuf. Process. 63 (2021) 163–178, http://dx.doi.org/10.1016/j.jmapro.2020.04.004.

[9] Jie Li, Yi Zhang, Wen Liu, Bin Li, Xuni Yin, Cong Chen, Prediction of penetration based on plasma plume and spectrum characteristics in laser welding, J. Manuf. Process. 75 (2022) 593–604, http://dx.doi.org/10.1016/j.jmapro.2022.01.032.

[10] Shixuan Li, Ping Jiang, Yu Gao, Minjie Song, Leshi Shu, A penetration depth monitoring method for Al-Cu laser lap welding based on spectral signals, J. Mater. Process. Technol. 317 (2023) 117972, http://dx.doi.org/10.1016/j.jmatprotec.2023.117972.

[11] Da Zeng, Di Wu, Hongxing Huang, Biao Peng, Yutong Wei, Hui Du, Peilei Zhang, Haichuan Shi, Qinghua Lu, Xiaoyu Cai, Online identification of laser welding penetration through multi-photoelectric decomposition-reconstruction and shifted-windows-based transformer deep learning framework, Measurement 247 (2025) 116872, http://dx.doi.org/10.1016/j.measurement.2025.116872.

[12] Chenpeng Jia, Yiming Huang, Shengbin Zhao, Changxing Li, Jiong Yuan, Feng Zhang, Lijun Yang, Statistical analysis of plasma thermodynamic behavior based on laser deep penetration welding of TC4 titanium alloy, Measurement 225 (2024) 114004, http://dx.doi.org/10.1016/j.measurement.2023.114004.

[13] Wei Meng, Xiaohui Yin, Junfei Fang, Lijie Guo, Qunshuang Ma, Zhuguo Li, Dynamic features of plasma plume and molten pool in laser lap welding based on image monitoring and processing techniques, Opt. Laser Technol. 109 (2019) 168–177, http://dx.doi.org/10.1016/j.optlastec.2018.07.073.

[14] Yinrui Gao, Ping Zhong, Xin Tang, Haowei Hu, Peng Xu, Feature extraction of laser welding pool image and application in welding quality identification, IEEE Access 9 (2021) 120193–120202, http://dx.doi.org/10.1109/ACCESS.2021.3108462.

[15] ZhenZhou Wang, The active visual sensing methods for robotic welding: Review, tutorial, and prospect, IEEE Trans. Instrum. Meas. 73 (2024) 1–19, http://dx.doi.org/10.1109/tim.2024.3485460.

[16] Shaojie Wu, Weichen Kong, Yingchao Feng, Peng Chen, Fangjie Cheng, Penetration prediction of narrow-gap laser welding based on coaxial high dynamic range observation and machine learning, J. Manuf. Process. 110 (2024) 91–100, http://dx.doi.org/10.1016/j.jmapro.2023.12.017.

[17] Thomas Will, Jannis Kohl, Silvana Burger, Claudio Hölbling, Lars Müller, Michael Schmidt, Weld classification with feature extraction by FRESH algorithm based on surface topographical optical coherence tomography data for laser welding of copper, IEEE Access 10 (2022) 109795–109802, http://dx.doi.org/10.1109/access.2022.3208877.

[18] Shuai Ma, Jiewu Leng, Zhuyun Chen, Yixian Du, Xiaoji Zhang, Qiang Liu, Intrinsically and post-hoc interpretable Kolmogorov–Arnold Network and genetic algorithm for laser deep penetration welding parameters optimization, IEEE Trans. Instrum. Meas. 74 (2025) 1–14, http://dx.doi.org/10.1109/tim.2025.3551494.

[19] Wang Cai, Ping Jiang, LeShi Shu, ShaoNing Geng, Qi Zhou, Real-time laser keyhole welding penetration state monitoring based on adaptive fusion images using convolutional neural networks, J. Intell. Manuf. 34 (3) (2023) 1259–1273, http://dx.doi.org/10.1007/s10845-021-01848-2.

[20] Wang Cai, Ping Jiang, Leshi Shu, Shaoning Geng, Qi Zhou, Real-time identification of molten pool and keyhole using a deep learning-based semantic segmentation approach in penetration status monitoring, J. Manuf. Process. 76 (2022) 695–707, http://dx.doi.org/10.1016/j.jmapro.2022.02.058.

[21] Huaping Li, Hang Ren, Zhenhui Liu, Fule Huang, Guangjie Xia, Yu Long, In-situ monitoring system for weld geometry of laser welding based on multi-task convolutional neural network model, Measurement 204 (2022) 112138, http://dx.doi.org/10.1016/j.measurement.2022.112138.

[22] Sikai Liu, Di Wu, Zhongyi Luo, Peilei Zhang, Xin Ye, Zhishui Yu, Measurement of pulsed laser welding penetration based on keyhole dynamics and deep learning approach, Measurement 199 (2022) 111579, http://dx.doi.org/10.1016/j.measurement.2022.111579.

[23] Fuqin Deng, Yongshen Huang, Guangwen Yao, Hufei Zhu, Bing Luo, Shufen Liang, Ningbo Yi, Improving convolution neural networks with window-based transformer blocks for laser welding process monitoring, in: Proc. 2022 IEEE Int. Conf. Ind. Technol., ICIT, 2022, pp. 1–8, http://dx.doi.org/10.1109/icit48603.2022.10002719.

[24] Shenghong Yan, Bo Chen, Han Gao, Caiwang Tan, Xiaoguo Song, Guodong Wang, Cross-attention time-series multi-feature fusion vision transformer for joint formation monitoring in laser scanning welding, Mech. Syst. Signal Process. 229 (2025) 112531, http://dx.doi.org/10.1016/j.ymssp.2025.112531.

[25] Rui Yu, Joseph Kershaw, Peng Wang, YuMing Zhang, How to accurately monitor the weld penetration from dynamic weld pool serial images using CNN-LSTM deep learning model? IEEE Robot. Autom. Lett. 7 (3) (2022) 6519–6525, http://dx.doi.org/10.1109/LRA.2022.3173659.

[26] Francis Ogoke, Peter Pak, Alexander Myers, Guadalupe Quirarte, Jack Beuth, Jonathan Malen, Amir Barati Farimani, Deep learning for melt pool depth contour prediction from surface thermal images via vision transformers, Addit. Manuf. Lett. 11 (2024) 100243, http://dx.doi.org/10.1016/j.addlet.2024.100243.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.

[28] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, Saining Xie, A ConvNet for the 2020s, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2022, pp. 11966–11976, http://dx.doi.org/10.1109/CVPR52688.2022.01167.

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., ICCV, 2021, pp. 9992–10002, http://dx.doi.org/10.1109/ICCV48922.2021.00986.

[30] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, Baining Guo, CSWin transformer: A general vision transformer backbone with cross-shaped windows, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2022, pp. 12124–12134, http://dx.doi.org/10.1109/cvpr52688.2022.01181.

[31] Weihao Yu, Chenyang Si, Pan Zhou, Mi Luo, Yichen Zhou, Jiashi Feng, Shuicheng Yan, Xinchao Wang, MetaFormer baselines for vision, IEEE Trans. Pattern Anal. Mach. Intell. 46 (2) (2024) 896–912, http://dx.doi.org/10.1109/TPAMI.2023.3329173.

[32] Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, Yixuan Yuan, EfficientViT: Memory efficient vision transformer with cascaded group attention, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR, 2023, pp. 14153–14163, http://dx.doi.org/10.1109/CVPR54522.2023.01359.

[33] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, Anurag Ranjan, FastViT: A fast hybrid vision transformer using structural reparameterization, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., ICCV, 2023, pp. 5768–5778, http://dx.doi.org/10.1109/ICCV56702.2023.00532.