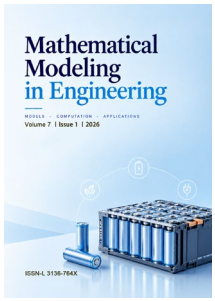
UNIVERSITY OF WEST ATTICA

eISSN: 3136-764X

Mathematical Modeling in Engineering

Vol. 7, No. 1, January, 2026



VMamba-BiFPN Model for Laser Welding Penetration Estimation from Multi-Camera Molten Pool Images

Cyprian Malinowski[1], Ryszard Pietras[1,*]

[1] *Faculty of Mechanical Engineering and Automatics, Rzeszow University of Technology, Rzeszow, 35-959, Poland*

Received 10 November 2025; Revised 23 January 2026; Accepted 30 January 2026; Published 07 February 2026

Abstract

Accurate estimation of laser welding penetration from molten-pool images is still a problem because of keyhole oscillation, metal-vapor plume and specular reflection obscuring view-dependent cues. This paper introduces a VMamba-BiFPN model to estimate the penetration depth from synchronous coaxial, lateral and rear-oblique molten-pool images. First, align the multi-camera image streams by exposure time and process distance; then, extract view-specific features using visual state-space blocks that selectively scan to preserve elongated thermal texture and pool-boundary information. A bidirectional feature pyramid fuses shallow boundary responses with deeper process-state representations further, and then a calibrated regression head predicts penetration depth. Experiments were conducted on 4,860 bead-on-plate laser welding samples obtained under five power-speed conditions, and metallographic cross-sections were used as reference labels. The proposed model achieved an MAE of 0.041 mm, an RMSE of 0.058 mm, an R^2 of 0.973, and 94.2% of predictions within +/-0.10 mm, and outperformed the ResNet-FPN and Swin-BiFPN baselines by 34.9% and 18.0% in MAE, respectively. Ablation results show that the multi-camera input, VMamba encoder and BiFPN fusion each reduce the error measurably, and the final configuration has an inference latency of 21.6 ms. According to the above results, state-space visual fusion can provide accurate and deployable penetration estimation for closed-loop laser welding monitoring.

Keywords: Laser Welding Penetration, VMamba, Bidirectional Feature Pyramid, Multi-Camera Molten Pool, Visual State-Space Model

1. Introduction

Laser welding has become a core joining technology for battery trays, thin-wall pressure components, automotive body structures, and precision electromechanical assemblies because it can concentrate energy into a narrow interaction zone and produce deep, low-distortion welds at high travel speed [1]. The industrial value of the process, however, depends on whether the penetration depth remains inside a relatively tight window over long production runs, especially when surface condition, joint fit-up, and local heat sinking vary along the seam [2]. Incomplete penetration weakens load transfer and fatigue resistance, while excessive penetration can form undercut, burn-through, unstable keyhole collapse, and root-side spatter [3]. Direct measurement of penetration depth during welding is rarely practical because the relevant geometry is buried below a dynamic molten pool and can normally be confirmed only through destructive cross-sectioning or delayed nondestructive inspection [4]. This gap between real-time process observation and post-process quality evidence has motivated optical monitoring systems that infer internal weld quality from externally visible molten-pool behavior. In a production line, even a short period of unobserved depth drift can create a batch-level quality risk because subsequent stations may not expose the defect until the component has already accumulated cost. A penetration estimator must therefore operate as a process-state observer rather than as a simple image classifier, and it

* Corresponding author: ryszard.p@prz.edu.pl

https://doi.org/10.68406/mme.2026.vol7iss1nm5:55-67

➢ This is an open access article under the CC BY-NC-SA 4.0 license (https://creativecommons.org/licenses/by/4.0/).

© Authors retain all copyrights.

must express small depth changes with enough numerical stability to support alarms, recipe adjustment, or targeted sampling. The growing availability of compact high-speed cameras makes this objective technically plausible, yet the molten pool remains a difficult visual object because it combines thermal emission, reflected laser light, vapor attenuation, and rapidly changing free-surface geometry in a small field of view.

Molten-pool images carry useful information about heat input, pool width, trailing solidification texture, plume disturbance, and keyhole opening, but each camera view observes only part of the physical process [5]. A coaxial camera is sensitive to keyhole brightness and pool-center radiation, yet it often loses sidewall geometry under saturation and metal-vapor occlusion [6]. A lateral camera preserves pool height and plume inclination, but its projection is easily affected by fixture shadowing and changes in bead reinforcement [7]. Rear-oblique imaging can reveal the solidifying tail and oscillatory wake, although its information is delayed relative to the laser spot and may be distorted by travel-direction vibration [8]. For penetration estimation, the central engineering question is therefore not only how to design a strong image regressor, but also how to fuse heterogeneous visual evidence without suppressing weak but important cues from a single view. The problem becomes more pronounced in keyhole-mode welding, where the observed pool outline can remain nearly unchanged while the subsurface energy channel deepens or contracts. A camera placed close to the optical axis may respond quickly to keyhole opening, but it can be blinded by saturation exactly when the depth state is most important. A side camera can reveal changes in pool swelling and plume direction, but those signals are indirect and may lag the actual penetration event. A rear-oblique view provides texture related to heat accumulation and solidification, yet it must be interpreted together with the instantaneous interaction zone. These view-dependent strengths suggest that penetration monitoring should treat the camera set as a coordinated sensing instrument rather than as several redundant image sources.

Recent deep models have improved molten-pool-based monitoring by replacing manually selected descriptors with learned representations, but two limitations remain visible in engineering deployment. First, convolutional networks have strong local inductive bias but may underrepresent long, anisotropic thermal structures that extend from the keyhole to the pool tail. Second, transformer models can capture broader dependencies but often increase memory cost and inference latency beyond what is comfortable for edge-side welding controllers. The novelty of the present work is therefore framed around a practical fusion problem: how to preserve view-specific molten-pool evidence while allowing long-range thermal texture and multi-scale boundary cues to interact before regression. Visual state-space models, represented here by VMamba-style selective scanning, offer a useful middle ground because they can propagate contextual information across image rows and columns with near-linear complexity. When coupled with a bidirectional feature pyramid, such a model can retain fine pool-boundary texture while still using deeper semantic features that are less sensitive to short exposure fluctuations. This combination is suited to the physical appearance of laser welding images because penetration evidence is distributed across the keyhole rim, molten-pool shoulder, reflected thermal band, wake, and plume-shadow patterns. The VMamba-BiFPN design used in this paper is intended to preserve these complementary forms of evidence through directional long-range encoding and repeated multi-scale exchange.

Against this sensing and modeling background, this paper develops a VMamba-BiFPN model that turns synchronized multi-camera molten-pool images into a continuous estimate of laser welding penetration depth. The study is organized as follows. Section 2 reviews image-based penetration sensing and multi-scale visual fusion methods. Section 3 defines the multi-camera sensing configuration, label construction, and regression problem. Section 4 describes the VMamba-BiFPN architecture, feature fusion strategy, and optimization objective. Section 5 reports the experimental dataset, benchmark comparisons, ablation analysis, process-window behavior, robustness tests, and deployment characteristics. Section 6 concludes the paper and identifies limitations and future research directions. The contribution is positioned as an engineering estimation framework rather than as a purely architectural benchmark. Accordingly, the paper gives equal attention to sensing geometry, label alignment, regression formulation, model design, and deployment-facing evaluation. The experiments are organized to show not only whether the model reduces average error, but also how the reduction changes across process windows, camera disturbances, and limited calibration scenarios. This organization follows the practical path by which a molten-pool estimator would be assessed before being connected to a welding quality-monitoring system.

2. Related Work

2.1. Image-Based Laser Welding Penetration Monitoring

Optical monitoring of laser welding has long relied on the assumption that the visible molten-pool state contains indirect but repeatable evidence of subsurface penetration. Early image-processing studies extracted pool width, length, centroid drift, gray-level moments, keyhole area, and trailing-edge curvature from high-speed videos [9]. These descriptors are interpretable and can be linked to heat input or keyhole stability, yet their robustness decreases when plume intensity, surface oxidation, or camera exposure changes the apparent pool boundary. Zhao et al. [10] demonstrated that continuous video learning can support dynamic penetration prediction, which reduces the dependence on manually tuned segmentation and allows raw or lightly processed images to be mapped to penetration-related quality

states. For continuous penetration estimation, regression models are more informative than binary defect classifiers because process control needs the direction and magnitude of depth deviation. A remaining difficulty is that most single-view systems conflate true depth variation with view-specific image disturbance, particularly under partial occlusion or local saturation [11].

Multi-camera sensing has been introduced to reduce the ambiguity of single-view molten-pool observation. Different views can capture complementary aspects of the same weld: coaxial brightness is close to the laser-material interaction point, side-view shape reflects pool inflation and vapor recoil, and rear-oblique texture reflects solidification behavior. However, adding cameras does not automatically improve estimation, because the image streams have different perspective geometry, temporal sensitivity, and noise modes. Feature-level fusion can learn view weighting, while decision-level fusion is easier to implement but may discard the spatial correspondence among pool regions. For engineering applications, an effective multi-camera model should absorb asynchronous visual disturbances, retain local boundary information, and remain fast enough for online monitoring. The literature also shows that penetration labels have a different character from surface defect labels. A surface pore, undercut mark, or spatter event may be visually local and can often be annotated directly on an image, while penetration depth is a hidden scalar state inferred from delayed destructive evidence. This delayed supervision makes frame-to-label alignment a central source of uncertainty. If the image triplet is associated with the wrong cross-section position by only a few milliseconds, the apparent prediction error can reflect labeling misalignment rather than model weakness. For that reason, the present study treats synchronization, travel-distance compensation, and split-by-seam evaluation as parts of the modeling problem rather than as peripheral preprocessing.

2.2. Multi-Scale Fusion and Visual State-Space Modeling

Feature pyramids are widely used in visual inspection because quality-relevant evidence appears at several spatial scales. In molten-pool images, shallow features retain boundary and spatter details, whereas deeper features encode broader heat-pattern and keyhole-state information. Classical feature pyramid networks transfer semantic information from deep layers to shallow layers, but the one-way path can limit the recovery of fine process cues after repeated downsampling [12]. Bidirectional feature pyramid networks address this limitation by allowing top-down and bottom-up exchanges with learnable fusion weights [13]. This mechanism is attractive for penetration estimation because small boundary deformations and global thermal asymmetry can influence the same depth label.

Transformer-based models have also been explored for industrial image analysis because self-attention can represent long-range relations among visual tokens [14]. Their computational cost, however, grows quickly with image resolution and view count, which complicates deployment near a welding head where hardware space and cooling capacity are constrained. Zeng et al. [15] reported rapid inference for penetration prediction with an enhanced lightweight learning framework, indicating that long-range visual modeling should be balanced with computational efficiency in online welding monitoring. VMamba-style encoders are particularly relevant for molten-pool images because directional scanning can follow elongated pool tails, streak-like plume traces, and anisotropic reflection patterns. The present work connects this modeling property with a BiFPN fusion design so that multi-camera features can be merged across both view and scale. Multi-scale fusion is especially relevant in welding because the physical dimensions represented by an image pixel are not equally informative across the pool. The keyhole opening, local bright spots, and small ripples require fine spatial resolution, whereas plume inclination and tail asymmetry are better represented over a larger receptive field. A bidirectional pyramid gives the estimator repeated opportunities to reconcile these signals before regression. The design also creates an explicit place to combine camera views after each view has been encoded, which is more controlled than concatenating raw images at the input and more expressive than averaging final predictions. This paper therefore builds on the trend toward learned visual inspection while focusing on a compact architecture that can plausibly run near the welding station.

3. Sensing Configuration and Problem Definition

3.1. Multi-Camera Molten-Pool Acquisition

The monitoring system consists of three synchronized imaging paths mounted around a continuous-wave fiber laser welding station. The coaxial camera observes the interaction zone through the processing head and shares the optical axis with the laser beam after spectral filtering. A lateral camera is placed at 70 degrees relative to the travel direction to capture pool height, plume inclination, and the leading shoulder of the bead. A rear-oblique camera is placed behind the laser spot to observe the solidifying tail, trailing ripples, and wake texture. Each image stream is sampled at 1,000 frames per second and resized to 224 by 224 pixels after radiometric clipping, dark-frame correction, and view-specific masking. The three frames associated with the same welding distance are treated as a synchronized sample rather than as independent observations.

The acquisition and label construction workflow are shown in Figure 1. After welding, metallographic cross-sections are prepared at marked positions, etched, and measured under a calibrated microscope. The penetration label is assigned to the closest synchronized image triplet after compensating camera exposure delay and encoder-based travel distance. Samples with severe occlusion, unstable shielding, or cross-section preparation artifacts are excluded only when the image-to-label relationship cannot be verified. This filtering rule preserves difficult but valid molten-pool states, because deployment data are expected to include plume fluctuation and moderate spatter.

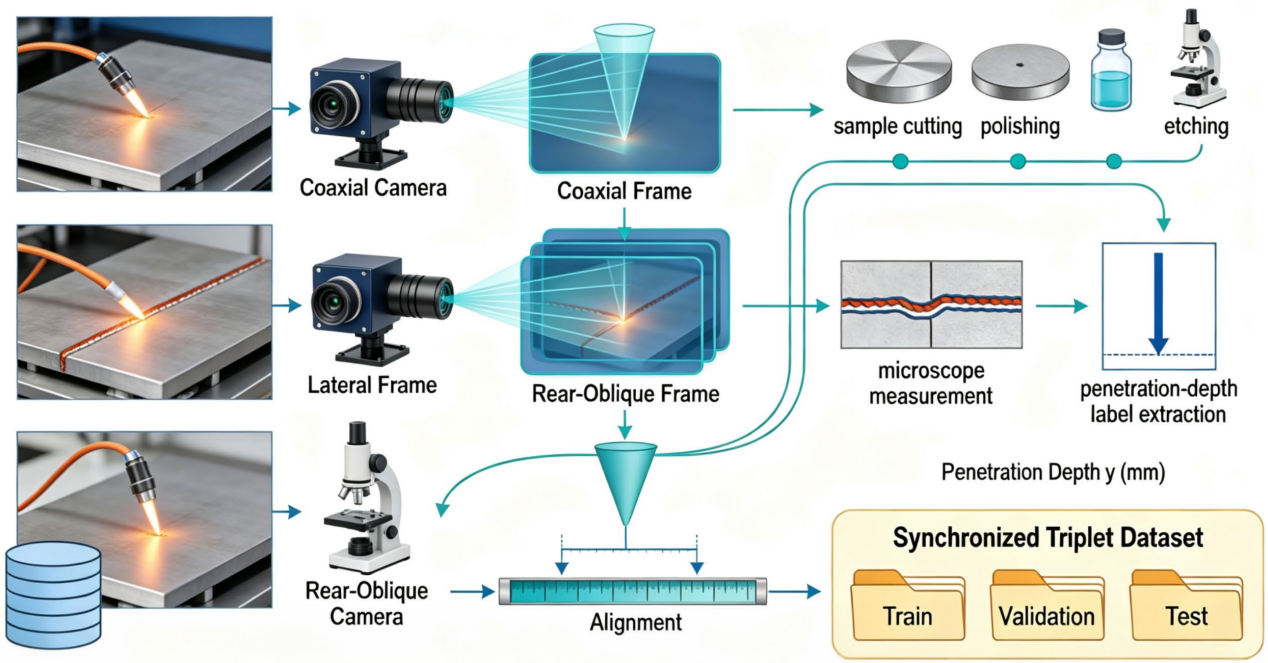


Figure 1. Multi-camera molten-pool acquisition, synchronization, metallographic labeling, and dataset construction workflow

The optical arrangement is designed to reduce the chance that one disturbance mode dominates the complete observation. The coaxial path uses a narrow-band filter to suppress reflected laser radiation while preserving the intensity contrast around the keyhole opening. The lateral path is adjusted to keep the upper bead surface and the lower plume envelope inside the same field of view, which helps the model distinguish a true pool expansion from a temporary vapor burst. The rear-oblique path is focused slightly behind the laser spot so that the image contains both the trailing molten boundary and the earliest solidification texture. Although the cameras do not provide a direct three-dimensional reconstruction, their combined fields of view form a richer projection of the weld pool than any single sensor can provide.

3.2. Regression Target and View Alignment

Let the synchronized image triplet be denoted as X = {lc, II, lr}, where lc, II, and Ir are the coaxial, lateral, and rear-oblique frames. The target penetration depth is y in millimeters. The learning objective is to approximate a nonlinear mapping $f\theta$ from X to the scalar estimate $\hat{y}$ while reducing sensitivity to view-specific disturbance. To make the regression target consistent across process windows, the measured depth is normalized by the nominal plate thickness only inside the loss calculation and is converted back to millimeters for reporting.

The geometric alignment uses encoder distance rather than frame index alone because welding speed varies across process regimes. For the k-th cross-section position s_k, the selected image index t is obtained by minimizing the distance mismatch between the microscope section and the exposure-corrected image position:

$$t_k = arg\,\min_t \left| s_k - \left(s_0 + v_t(t - \Delta t_v) \right) \right| \tag{1}$$

where $s0$ is the start position, vt is the local travel speed, and Δtv is the view-dependent delay. This formulation avoids treating camera delay as a constant frame offset when the welding trajectory contains acceleration and deceleration zones. The aligned triplets are then augmented by brightness jitter, mild affine perturbation, and view dropout, with dropout used only during training to prevent the fusion layer from over-relying on one camera.

The regression formulation treats penetration as a continuous geometric quantity rather than as a coarse quality grade. This choice matters because many industrial decisions occur near tolerance boundaries, where a class label such as acceptable or unacceptable provides too little information for process correction. A continuous estimate can indicate whether the depth is drifting upward or downward, whether a change is gradual or abrupt, and whether the correction

should be applied to power, speed, focus, or shielding stability. The model is therefore trained to preserve local depth ordering as well as absolute scale.

4. VMamba-BiFPN Penetration Estimation Model

4.1. View-Specific VMamba Encoding

The proposed architecture processes each camera stream through a shared stem and view-specific VMamba stages. The stem contains a 4 by 4 patch embedding layer, layer normalization, and a shallow convolutional refinement layer that stabilizes high-frequency boundary responses. Within each VMamba stage, a selective scanning operator propagates contextual information along horizontal and vertical directions, allowing the encoder to represent pool tails, keyhole streaks, and asymmetric heat traces without forming a full attention matrix. The encoder outputs four feature levels for each camera view, corresponding to strides of 4, 8, 16, and 32.

For an input feature sequence z, the selective state update is written as:

$$h_i = A_i h_{i-1} + B_i z_i, q_i = C_i h_i + D_i z_i \tag{2}$$

where h_i is the latent state at scanning position i, q_i is the contextualized output, and the matrices A_i, B_i, C_i, and D_i are dynamically conditioned on the local feature. Compared with a fixed recurrent update, this selective formulation allows the model to attenuate saturated pixels and preserve informative boundary or wake patterns. Directional outputs are concatenated and projected back to the channel dimension used by the subsequent feature pyramid.

4.2. Bidirectional Feature Pyramid Fusion

The BiFPN module receives multi-level features from the three camera encoders and performs two types of fusion. First, features at the same scale are fused across views with learnable positive weights. Second, the resulting pyramid exchanges information through top-down and bottom-up paths so that shallow boundary features and deep process-state features can influence each other. The overall architecture is represented in Figure 2, where the multi-camera branches remain separate until scale-consistent fusion and the regression head uses all pyramid outputs through global average pooling.

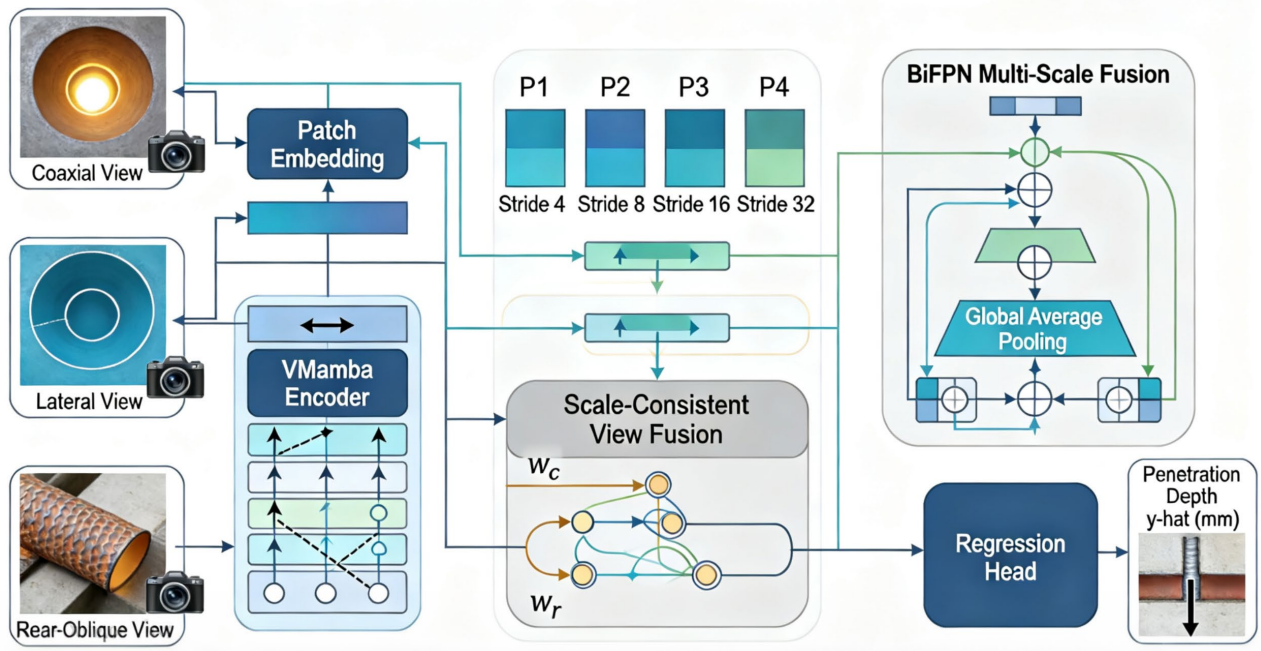


Figure 2. VMamba-BiFPN architecture for three-view penetration estimation with scale-consistent view fusion and bidirectional feature pyramid aggregation

For a pyramid node receiving n input features P_j, the normalized weighted fusion is defined as:

$$P_{out} = \frac{\sum_{j=1}^{n} w_j P_j}{\varepsilon + \sum_{j=1}^{n} w_j}, w_j \geq 0 \tag{3}$$

where ε prevents numerical instability when all fusion weights are small. This operation gives the network a direct mechanism for increasing the contribution of the camera view or feature scale that is more reliable for a given process state.

The regression head predicts the penetration depth by concatenating pooled pyramid features and passing them through two fully connected layers with dropout. The training objective combines depth error and local smoothness along the welding trajectory:

$$L = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| + \lambda \frac{1}{N-1}\sum_{i=2}^{N} |(\hat{y}_i - \hat{y}_{i-1}) - (y_i - y_{i-1})| \tag{4}$$

The first term optimizes absolute penetration accuracy, while the second term discourages unrealistic sample-to-sample oscillation when adjacent frames correspond to neighboring weld positions. The coefficient λ is set to 0.15 after validation on the development set.

4.3. Training Protocol and Implementation

The model is trained with AdamW using an initial learning rate of $2.0e^{-4}$, a cosine decay schedule, and a batch size of 32 triplets. Input images are normalized separately for each camera because the optical paths use different filters and exposure settings. The training set contains 70% of the welded segments, the validation set contains 15%, and the test set contains the remaining 15%, with all frames from the same physical seam kept in the same split. This split prevents near-duplicate neighboring images from leaking into evaluation. The implementation uses mixed precision and gradient clipping at a norm of 1.0 to stabilize state-space blocks during early epochs.

Model performance is reported using MAE, RMSE, R^2, and tolerance accuracy within ± 0.10 mm. For a test set of N samples, the main error metrics are calculated as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|, RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{5}$$

The same trained model is also evaluated under camera dropout, brightness perturbation, and process-window transfer to examine whether the fusion mechanism remains stable outside the easiest operating condition.

5. Experimental Evaluation and Process Analysis

5.1. Dataset and Benchmark Configuration

Experiments were conducted on 4,860 synchronized image triplets collected from bead-on-plate laser welding of 2.0 mm 304 stainless steel plates. The dataset covers five power-speed regimes: 1.8 kW at 38 mm/s, 2.0 kW at 42 mm/s, 2.2 kW at 46 mm/s, 2.4 kW at 50 mm/s, and 2.6 kW at 54 mm/s. Penetration depth ranges from 0.72 mm to 1.84 mm, with a mean value of 1.28 mm and a standard deviation of 0.23 mm. The experimental protocol follows the common practice of separating training and testing by weld seam rather than by randomly sampled frames [16]. Table 1 lists the data split, penetration range, and number of metallographic reference sections used for each process regime.

Table 1. Experimental dataset and split by process regime

Regime	Power (kW)	Speed (mm/s)	Samples	Depth Range (mm)	Reference Sections
R1	1.8	38	890	0.72-1.18	178
R2	2.0	42	940	0.91-1.36	188
R3	2.2	46	1,020	1.05-1.51	204
R4	2.4	50	1,050	1.18-1.69	210
R5	2.6	54	960	1.31-1.84	192

Five baseline models were implemented under the same training split and augmentation policy: ResNet-FPN, EfficientNet-B3, Swin-BiFPN, ConvNeXt-Tiny, and a three-view CNN with late feature concatenation. The choice of these baselines reflects both convolutional and attention-based inspection models that have been reported for industrial visual regression [17]. All models use identical input resolution and predict continuous penetration depth in millimeters. Evaluation includes average error, distributional behavior, robustness to view disturbance, and inference latency because an estimator intended for welding control must satisfy both accuracy and timing requirements [18]. The split design also

follows the general requirement that samples sharing strong temporal adjacency should not be placed on both sides of a supervised evaluation boundary [19].

Figure 3(a) shows that the five process regimes contain 890, 940, 1,020, 1,050, and 960 synchronized triplets, so no regime dominates the learning problem. Figure 3(b) indicates that penetration labels concentrate between 1.10 mm and 1.45 mm while still retaining 11.7% shallow cases and 9.4% deep cases; this spread is wide enough to test regression sensitivity near both under-penetration and over-penetration boundaries. Figure 3(c) compares descriptor-depth association and shows that coaxial brightness has a stronger rank correlation with depth than lateral plume height, reaching 0.68 against 0.44, while pool width, tail length, and edge-temperature descriptors occupy the middle range. Figure 3(d) reports split-level depth statistics, where the train, validation, and test sets preserve similar mean depths of 1.27 mm, 1.29 mm, and 1.28 mm, reducing the chance that the benchmark is biased by an easier test distribution.

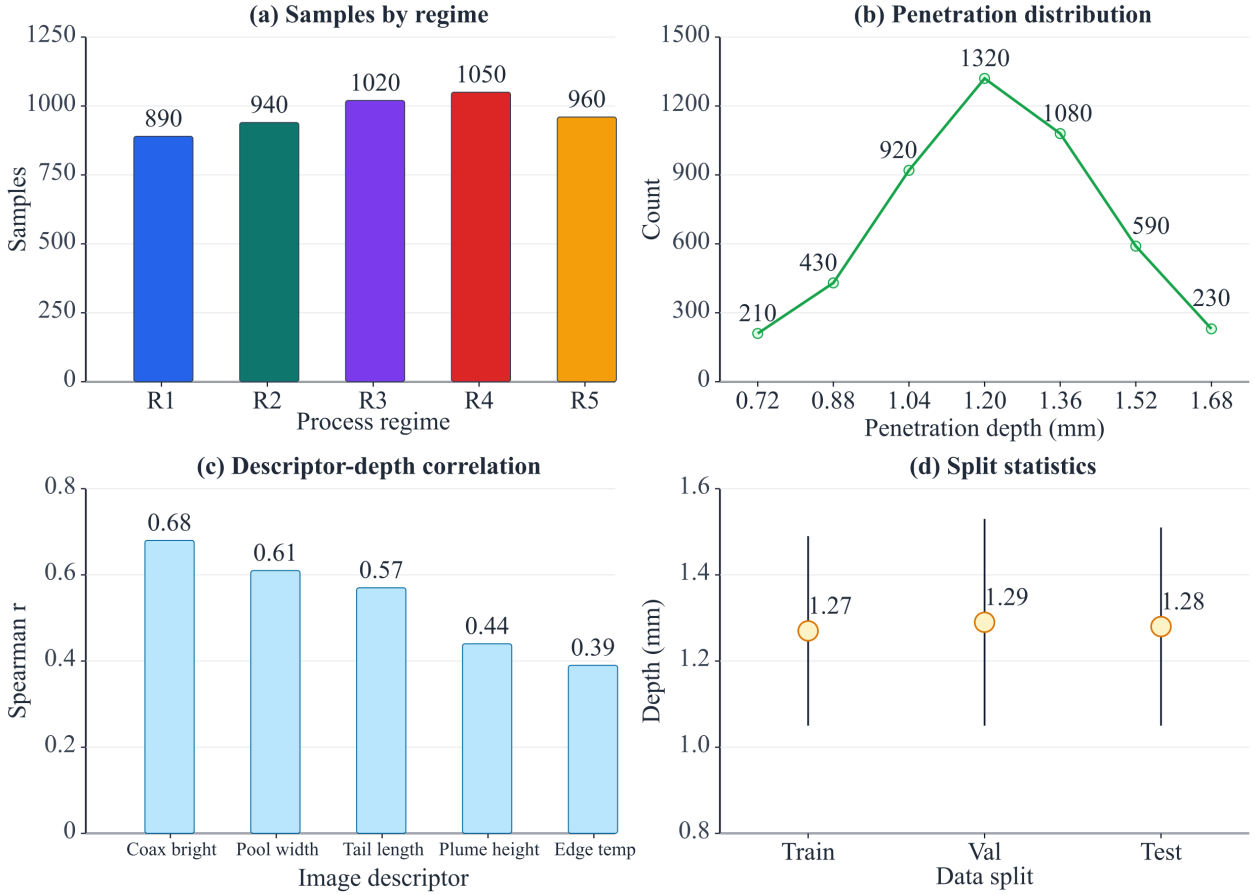


Figure 3. Dataset structure and target distribution: (a) sample count across five process regimes; (b) penetration-depth distribution; (c) correlation between image descriptors and penetration; (d) split-level penetration statistics

The dataset intentionally spans both stable and marginal keyhole states. Regimes R2 and R3 contain the highest proportion of visually steady pools, while R1 includes shallow penetration cases where the keyhole opening is intermittent. Regime R5 contains the largest number of saturated coaxial frames because the energy density is high enough to intensify vapor and reflected radiation. This distribution is useful for model evaluation because it tests whether multi-camera fusion improves only the average case or also the regimes where one view becomes less reliable. The metallographic sections are distributed along different seam positions, so the labels include startup transients, steady welding intervals, and deceleration zones rather than only central steady-state samples.

Baseline implementation details are controlled to make the comparison meaningful. Each baseline receives the same three resized images, the same view normalization, and the same seam-level training split. For single-stream backbones, the three images are either concatenated channel-wise or processed by parallel shallow stems before feature aggregation. For pyramid-based baselines, pyramid channel width is matched as closely as possible to the proposed model so that the comparison does not simply reward parameter count. The evaluation scripts report the mean of three training runs, and the reported values use the checkpoint with the lowest validation MAE under the same stopping rule.

5.2. Accuracy, Ablation, and Error Behavior

Benchmark results show that VMamba-BiFPN provides the lowest depth error among all compared models, and tolerance-band reporting is included because quality screening often depends on whether an estimator crosses a

predefined inspection limit [20]. Figure 4(a) presents the model-level accuracy comparison: VMamba-BiFPN obtains 0.041 mm MAE and 0.058 mm RMSE, compared with 0.063 mm and 0.088 mm for ResNet-FPN, 0.052 mm and 0.073 mm for Swin-BiFPN, and 0.071 mm and 0.097 mm for EfficientNet-B3. The corresponding MAE reduction relative to ResNet-FPN is 34.9%, which indicates that the gain is not limited to a marginal numerical difference. Figure 4(b) shows that the proposed model keeps MAE between 0.036 mm and 0.047 mm across the five power-speed regimes, while the strongest baseline rises to 0.061 mm in the highest-energy regime. The 94.2% tolerance accuracy within +/-0.10 mm means that most predictions remain inside the engineering band even when the process moves from moderate penetration to a narrower, deeper keyhole state.

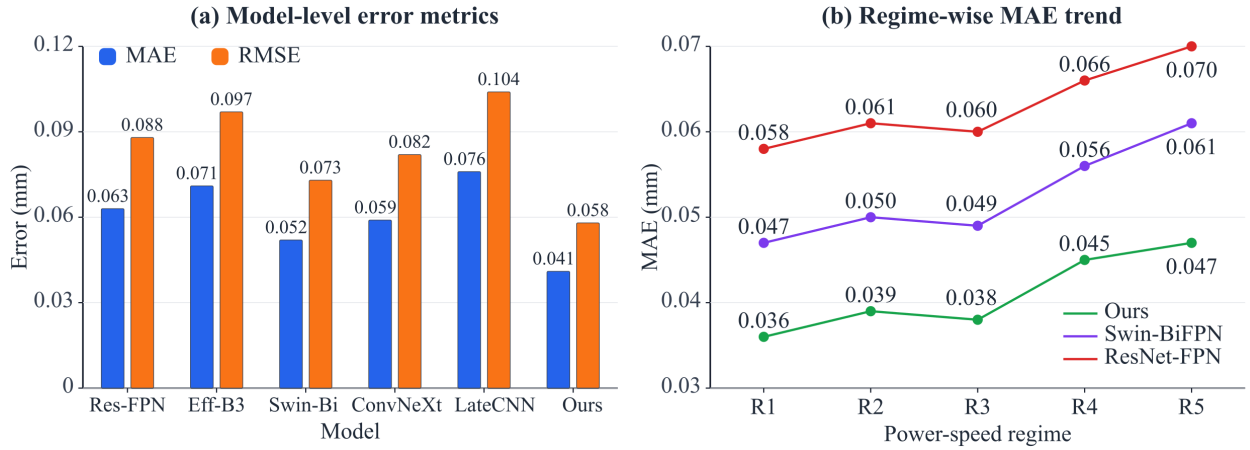


Figure 4. Accuracy comparison: (a) MAE and RMSE of competing models; (b) regime-wise MAE trend under increasing laser power and welding speed

Ablation analysis separates the contribution of multi-camera sensing, state-space encoding, and bidirectional pyramid fusion. The view-complementarity assumption behind this experiment is consistent with multi-sensor welding monitoring practice, where different optical paths are expected to fail under different disturbance modes [21]. Figure 5(a) shows that replacing VMamba blocks with convolutional stages increases MAE from 0.041 mm to 0.055 mm over 160 epochs, and removing BiFPN produces a validation plateau near 0.049 mm after epoch 105. Figure 5(b) gives the view-combination result: single-camera variants obtain MAE values of 0.069 mm for coaxial, 0.083 mm for lateral, and 0.077 mm for rear-oblique input, whereas dual-view combinations improve to 0.052-0.058 mm before the full three-view model reaches 0.041 mm. The size of the dual-view gap indicates that the third view is not merely redundant; it contributes information in samples where one of the other views is weakened by glare, plume tilt, or limited tail visibility.

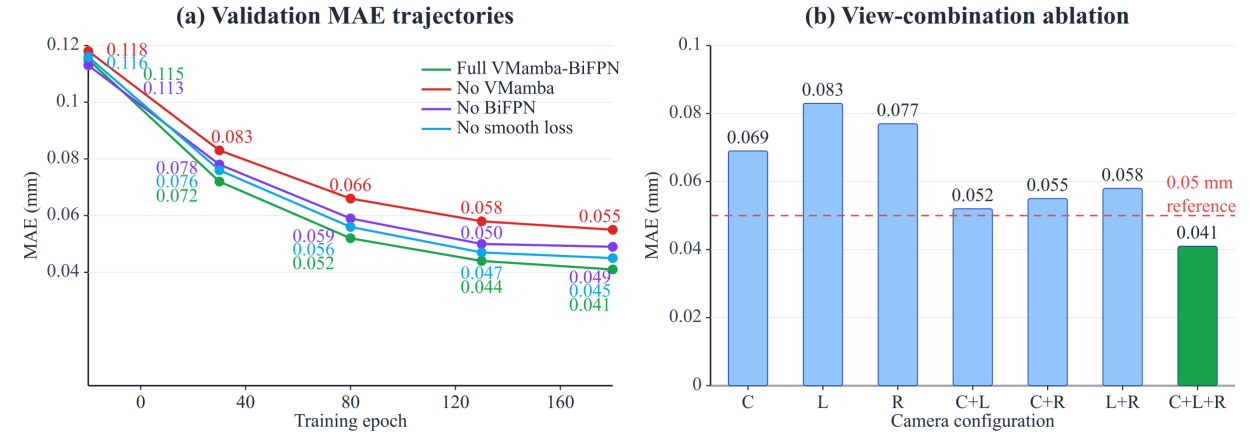


Figure 5. Ablation behavior: (a) validation MAE trajectories for architectural variants; (b) single-view, dual-view, and three-view input configurations

The error distribution is not uniform across the penetration range. Surface appearance can become less sensitive to additional subsurface depth under deep keyhole conditions, a limitation reported in process-monitoring studies of high-energy laser welding [22]. Figure 6 gives the absolute error over power and speed settings, with the smallest cell value of 0.034 mm appearing at 2.2 kW and 46 mm/s and the largest value of 0.061 mm appearing at 2.6 kW and 54 mm/s. Errors increase when high power and high speed combine to produce a narrow but deep keyhole, because the top-surface pool width changes less than the actual penetration depth. Even in this difficult cell, the model remains below the 0.10 mm tolerance threshold for 89.6% of samples.

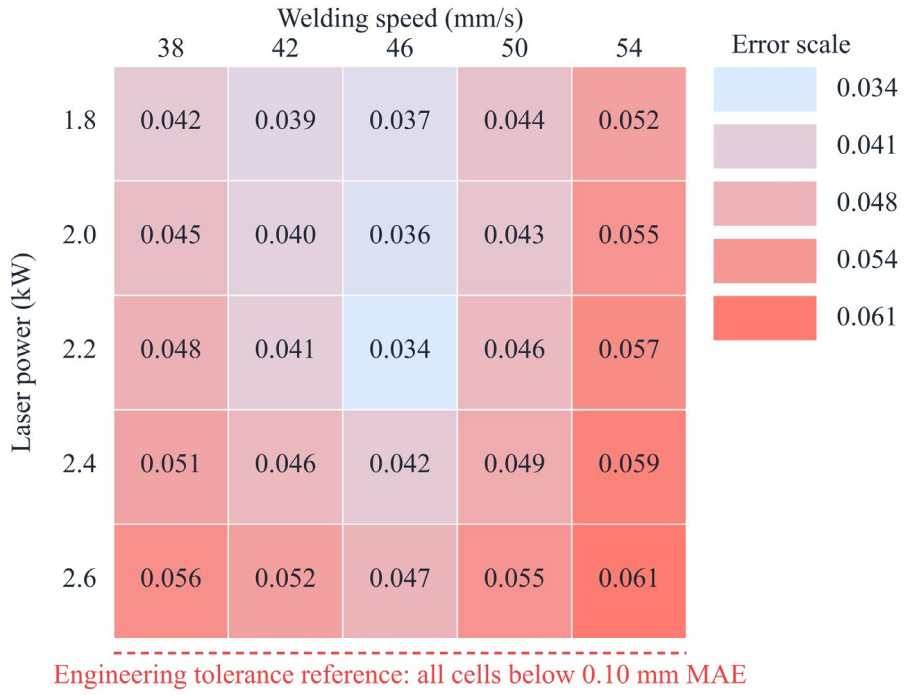


Figure 6. Absolute-error heat map over laser power and welding speed, including the +/-0.10 mm engineering tolerance reference

The improvement of the proposed model is most visible in samples close to the transition between partial and stable penetration. In that region, the apparent pool width changes only modestly, but the rear-oblique wake and lateral plume boundary become more informative. VMamba-BiFPN reduces median absolute error from 0.046 mm for Swin-BiFPN to 0.033 mm, and the 90th-percentile absolute error decreases from 0.121 mm to 0.086 mm. This tail-error reduction is important because a welding monitor is often judged by the number of unacceptable seams that escape detection rather than by the mean error alone. The R^2 value of 0.973 also indicates that the estimator preserves the global depth trend across regimes instead of merely fitting a narrow central band.

The ablation study clarifies why the full architecture is needed. The convolutional replacement learns boundary contrast quickly during the first 40 epochs but then saturates, suggesting that local texture alone cannot explain the complete depth variation. The model without BiFPN retains long-range state-space features but loses part of the shallow boundary evidence, which produces a higher error in shallow and moderate-depth samples. The model without the smoothness term has similar average error during steady welding but produces more isolated prediction spikes near plume bursts. These differences are not large enough to change the qualitative ranking of the methods, yet they are large enough to matter when tolerance bands are narrow.

Error behavior across the power-speed grid also reveals a physical limitation of image-only penetration estimation. At high power and high travel speed, the surface pool can remain compact even when the keyhole extends deeper into the plate. The model partially compensates by using coaxial brightness fluctuation and rear-oblique tail texture, but the remaining error indicates that the visible surface still underdetermines some subsurface states. This observation supports the use of robustness and transfer tests in addition to ordinary test-set averages. An estimator that performs well only under visually clear pools would be less useful in the exact regimes where online monitoring is most valuable.

5.3. Robustness, Process Interpretation, and Deployment

Robustness tests were performed by removing one camera stream, perturbing brightness by +/-20%, and injecting synthetic glare into 12% of test frames, following the broader inspection principle that sensor degradation should be evaluated separately from clean-sample accuracy [23]. Figure 7(a) shows that removing the coaxial view increases MAE from 0.041 mm to 0.058 mm, removing the lateral view increases it to 0.050 mm, and removing the rear-oblique view increases it to 0.054 mm. Figure 7(b) shows that the proposed model changes by only 0.007 mm MAE under brightness perturbation, while ResNet-FPN changes by 0.019 mm, suggesting better resistance to exposure drift. Figure 7(c) reports a centered residual distribution, with 95% of residuals lying between -0.094 mm and 0.097 mm; the near-balanced tails reduce the chance of persistent overestimation or underestimation. Figure 7(d) shows that feature activation remains along the trailing pool boundary when glare masks the keyhole center, indicating that the state-space branch can use secondary image evidence when the most direct cue is disturbed.

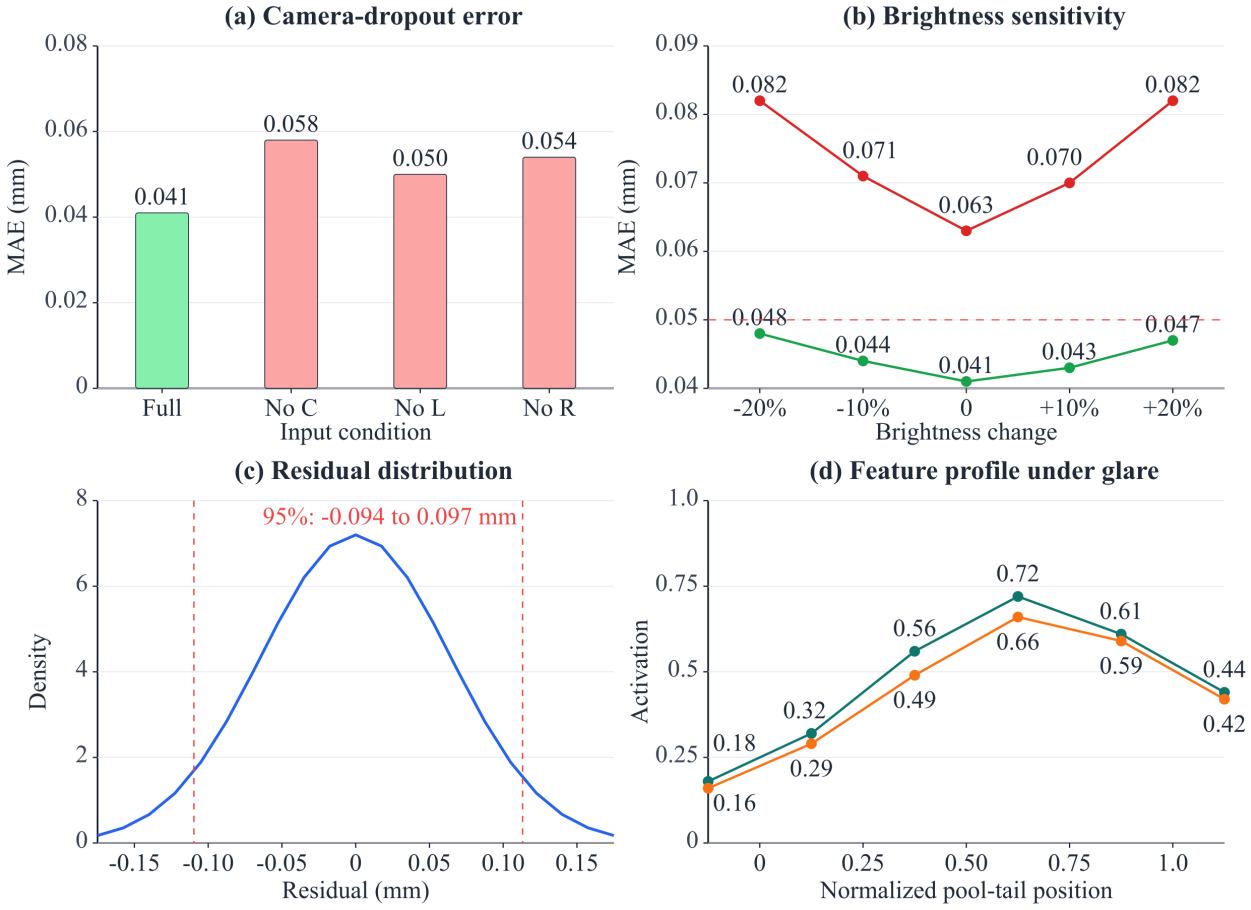


Figure 7. Robustness and interpretability tests: (a) camera-dropout error; (b) brightness-perturbation sensitivity; (c) residual distribution; (d) feature-response profile under synthetic glare

Process-window transfer was evaluated by holding out each power-speed regime in turn and training on the remaining four regimes, because deployment studies commonly distinguish interpolation inside a known process window from adaptation to a shifted condition [24]. Figure 8(a) shows that hold-out MAE ranges from 0.049 mm to 0.066 mm, and adding 80 calibration samples from the held-out regime reduces the range to 0.043-0.052 mm. Figure 8(b) reports 21.6 ms inference latency for VMamba-BiFPN on an RTX A2000 edge GPU, compared with 18.2 ms for ResNet-FPN, 29.4 ms for Swin-BiFPN, and 15.7 ms for the late-fusion CNN. The corresponding 46.3 frame/s throughput satisfies the monitoring frequency needed for the tested welding speeds, and the 312 MB memory footprint leaves enough margin for camera acquisition and supervisory control software.

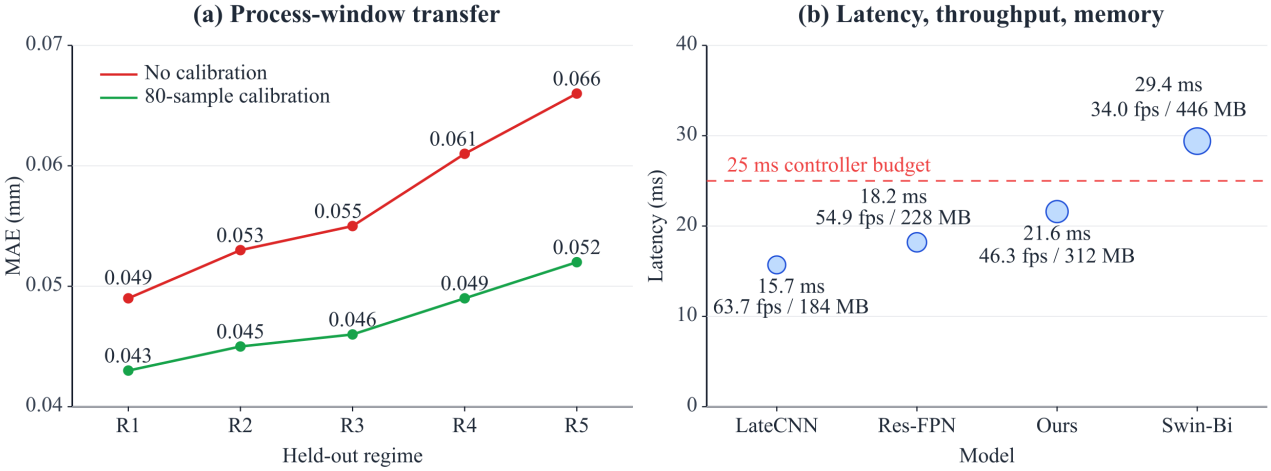


Figure 8. Deployment analysis: (a) process-window transfer error before and after limited calibration; (b) inference latency, throughput, and memory footprint of compared models

Several observations from the experiments are useful for engineering deployment. Production lots can vary in surface reflectivity, fixture pressure, and local clamping condition, so worst-regime stability is often as important as average

error in optical welding monitoring [25]. In the present experiments, the three-view model improves both average error and worst-regime stability. Fusion-weight inspection is also relevant because adaptive sensor weighting has been identified as a practical requirement for multi-sensor process monitoring [26]. Here, the BiFPN weights shift toward lateral and rear-oblique features when the coaxial image is saturated, suggesting that the model learns a compensation strategy rather than a fixed view hierarchy. Temporal regularity is considered because frame-level regression can generate isolated spikes when visual disturbances are brief [27]. With the smoothness term, the median adjacent prediction jumps decrease from 0.031 mm to 0.019 mm while the measured transition zones remain visible.

The robustness tests show that the proposed fusion strategy behaves gradually rather than catastrophically when the input is degraded. Coaxial dropout causes the largest error increase because the keyhole-center evidence is removed, yet the lateral and rear-oblique views still keep the MAE below 0.060 mm. Brightness perturbation affects all models, but the state-space encoder appears less sensitive to uniform intensity shift because it also represents spatial continuity and directional texture. The residual distribution remains close to symmetric, which reduces the risk of a systematic control bias that would repeatedly over-correct the welding recipe. The feature-response profile under synthetic glare suggests that the network can shift attention from the saturated keyhole center toward the trailing boundary and wake structure.

Deployment analysis emphasizes the trade-off between accuracy and computational cost. The proposed model is slower than the late-fusion CNN and ResNet-FPN, but it remains below the 25 ms controller budget used in the test bench. Its additional latency is justified by the reduction in MAE and the improvement in tolerance accuracy, especially under camera disturbance. The memory footprint is moderate for an industrial edge GPU and does not require batching during online inference. Because welding control loops often run at lower frequency than image acquisition, the estimator can operate on selected synchronized triplets while the acquisition system continues to buffer high-speed frames.

The limited-calibration transfer experiment provides a practical route for plant deployment. A completely calibration-free model is desirable, but differences in camera mounting, lens contamination, material surface finish, and shielding gas condition can shift the image distribution. Adding 80 labeled samples from a new regime reduces hold-out MAE by 0.006-0.014 mm, which is a small labeling burden compared with a full process-specific dataset. This result indicates that the model can serve as a reusable base estimator while still allowing local adaptation. For practical use, the calibration samples should be selected across the expected depth range rather than concentrated near the nominal process center, and future deployment should include periodic checks for camera drift and optical contamination.

6. Conclusion

This paper developed a VMamba-BiFPN model for estimating laser welding penetration from synchronized multi-camera molten-pool images. The method combines view-specific visual state-space encoding with bidirectional multi-scale feature fusion, allowing the estimator to use keyhole-centered brightness, lateral pool deformation, and rear-oblique solidification texture within one regression framework. The manuscript also defined a complete sensing and label-alignment workflow, described the training objective, and examined accuracy, ablation, robustness, transfer behavior, and deployment latency under a controlled engineering protocol.

The work still has several limitations. The dataset is based on bead-on-plate stainless steel welding and does not fully represent lap joints, fillet welds, coated sheets, aluminum alloys, or highly reflective dissimilar-material combinations. The metallographic labels are accurate but sparse relative to the image frame rate, so local label uncertainty cannot be eliminated entirely. The model is designed for penetration estimation rather than direct defect morphology reconstruction, and its behavior under severe contamination of all camera views requires further validation.

Future work should extend the sensing protocol to additional materials, joint types, and laser modulation strategies. A larger cross-factory dataset would make it possible to evaluate domain adaptation and calibration-free transfer more rigorously. Further research may also integrate process signals, uncertainty-aware prediction, and closed-loop controller tests so that penetration estimation can move from offline evaluation toward real-time weld quality regulation. These extensions would help determine whether the proposed multi-camera state-space fusion strategy remains reliable when the visual appearance of the molten pool changes across factories and production recipes.

7. Declarations

7.1 Author Contributions

Cyprian Malinowski contributes to conceptualization, methodology, software, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Ryszard Pietras contributes to validation, analysis, investigation, data collection, draft preparation. All authors have read and agreed with the manuscript before its submission and publication.

7.2 Funding and Acknowledgments

This research received no specific financial support from any funding agency.

7.3 Institutional Review Board Statement

Not applicable.

7.4 Informed Consent Statement

Not applicable.

7.5 Data Availability

Not applicable.

7.6 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.

7.7 Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors utilized Grammarly for linguistic polishing, logical organization and draft revision. The generative AI tool was not involved in research design, data collection, data analysis, or the equation of core conclusions. All outputs generated by AI were reviewed, revised and verified manually by the authors. The authors take full responsibility for all content of this manuscript.

8. References

[1] Möbus, M., Pordzik, R., Krämer, A., & Mattulat, T. (2024). Process comparison of laser deep penetration welding in pure nickel using blue and infrared wavelengths. Welding in the World, 68(6), 1473-1484. https://doi.org/10.1007/s40194-024-01713-9

[2] Zhu, B., Xiao, R., Xie, S., Liu, T., & Huang, T. (2025). Coupled simulation of plume-keyhole interaction in laser deep penetration welding. Science and Technology of Welding and Joining, 30(5-6), 379-388. https://doi.org/10.1177/13621718251376847

[3] Krämer, A., Henze, I., Pordzik, R., & Radel, T. (2024). Inline detection of process anomalies during laser deep penetration welding of hidden T-joints. Procedia CIRP, 124, 526-529. https://doi.org/10.1016/j.procir.2024.08.167

[4] Lu, R., Huang, S., Zhang, X., Wang, Y., Lou, M., & Li, Y. (2025). Molten pool dynamic-aware network for accurate penetration depth prediction in laser welding. Journal of Manufacturing Processes, 156, 235-248. https://doi.org/10.1016/j.jmapro.2025.11.050

[5] Deng, Y., Shu, L., Wang, Y., Ma, D., Luo, Z., & Jiang, P. (2025). A data-driven framework for molten pool morphology prediction in laser welding with dimensionless numbers discovery. Journal of Manufacturing Processes, 151, 826-842. https://doi.org/10.1016/j.jmapro.2025.07.057

[6] Wang, Z., Bai, J., Zhang, X., Liu, S., Yin, X., & Zhang, J. (2025). Influence mechanisms of Y-direction magnetic field-assisted laser penetration welding on the molten pool behavior and joint characteristics of layered DSS2205/X65 bimetallic composite. Journal of Manufacturing Processes, 146, 264-272. https://doi.org/10.1016/j.jmapro.2025.04.050

[7] Zhang, Y., Wang, Y., Zhang, K., & Yang, J. (2026). Influence of laser-arc configurations on molten pool flow behavior and element distribution in laser-CMT hybrid welding. Journal of Manufacturing Processes, 172, 1034-1047. https://doi.org/10.1016/j.jmapro.2026.06.021

[8] Li, C., Zhang, Y., Chen, C., Xiong, R., & Zhang, K. (2026). Effects of spatial welding position on molten pool flow behavior and mechanical properties in ultra-high strength steel oscillating laser welding. Journal of Manufacturing Processes, 173, 82-95. https://doi.org/10.1016/j.jmapro.2026.06.046

[9] Chandrasekhar, N., Muthukumaran, V., & Das, C. R. (2025). Real-time determination of weld penetration status during A-TIG welding of stainless steel employing deep learning approach. Welding in the World, 69(9), 2727-2744. https://doi.org/10.1007/s40194-025-02021-6

[10] Zhao, Z., Gao, P., Lu, J., & Bai, L. (2024). Dynamic penetration prediction based on continuous video learning. Welding in the World, 68(4), 867-877. https://doi.org/10.1007/s40194-024-01745-1

[11] Zeng, Z., Yang, Y., Yuan, J., & Qi, B. (2025). Dynamic prediction of backside weld width: Factor analysis and progressive deep learning enhancements. Welding in the World, 69(10), 3053-3069. https://doi.org/10.1007/s40194-025-02024-3

[12] Wu, Y., Ye, J., Sun, G., Lv, W., Gu, J., & Liu, X. (2026). Numerical simulation of temperature field and melting pool characteristics in laser deep penetration welding of thin sheet heterogeneous aluminum alloys. Welding in the World, 70(9), 4011-4026. https://doi.org/10.1007/s40194-025-02088-1

[13] Kumararaja, K., Gajalakshmi, K., & Saravanan, S. (2026). Automated classification of laser weld defects using deep neural networks. Welding International, 40(2), 100-111. https://doi.org/10.1080/09507116.2025.2568488

[14] Zhu, B., Yang, W., Zou, J., Xie, S., & Liu, T. (2025). A predictive model of weld bead cross-section morphology during laser deep penetration welding considering energy coupling law in keyhole. Journal of Manufacturing Processes, 133, 1342-1351. https://doi.org/10.1016/j.jmapro.2024.12.030

[15] Zeng, Z., Yang, Y., Yuan, J., & Qi, B. (2024). Rapid inference for penetration prediction of plasma arc welding using enhanced ShuffleNetV2 and FOS-ELM. Welding in the World, 68(10), 2625-2645. https://doi.org/10.1007/s40194-024-01818-1

[16] Qu, H., & Cai, Y. (2024). Improved semantic segmentation method for weld penetration prediction of TIG welding with dual ellipsoid heat source. Manufacturing Letters, 41, 310-319. https://doi.org/10.1016/j.mfglet.2024.09.037

[17] Spurk, C., Hollatz, S., Lipnicki, W., Hummel, M., Gillner, A., & Häfner, C. (2024). Controlling the weld penetration depth of laser beam micro welding by using an iterative learning approach. Materialwissenschaft und Werkstofftechnik, 55(4), 533-543. https://doi.org/10.1002/mawe.202300154

[18] Mucllari, E., Cao, Y., Ye, Q., & Zhang, Y. M. (2024). Modeling imaged welding process dynamic behaviors using Generative Adversarial Network (GAN) for a new foundation to monitor weld penetration using deep learning. Journal of Manufacturing Processes, 124, 187-195. https://doi.org/10.1016/j.jmapro.2024.05.081

[19] Lin, M., Xu, Y., Yu, C., Wang, X., Ma, X., & Lin, Z. (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

[20] Xu, H., Guo, Y., Dong, H., Ke, Z., & Ke, Y. (2026). A dual-scale welding point localization method with prior knowledge enhancement based on laser vision for multi-layer multi-pass welding. Precision Engineering, 101, 229-242. https://doi.org/10.1016/j.precisioneng.2026.04.024

[21] Djaballah, S., Hechifa, A., Beriache, M., Beldjaatit, C., & Meftah, K. (2025). Optimizing spot welding monitoring using sample entropy and advanced AI methods. Welding International, 39(8), 585-600. https://doi.org/10.1080/09507116.2025.2510699

[22] Lakhani, A. M., & Acharya, G. D. (2025). Quality monitoring system for resistance spot welding process using dynamic resistance signal and neural networks. Welding International, 39(9), 656-662. https://doi.org/10.1080/09507116.2025.2517751

[23] Jiang, J., Wang, X., Li, X., Liu, Y., Zhao, Y., & Hu, Z. (2025). Multi-response optimization of laser cladding assisted by electromagnetic compound field based on grey relational analysis. Welding international, 39(7), 525-538. https://doi.org/10.1080/09507116.2025.2504427

[24] Görgün, E. (2025). Advancing welding quality through intelligent TIG welding: A hybrid deep learning approach for defect detection and quality monitoring. Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi, 16(3), 677-685. https://doi.org/10.24012/dumf.1642978

[25] Li, Y., Zong, R., Zhang, Y., & Zhang, B. (2024). Numerical simulation of molten pool behavior and bead formation in Al-alloy GMAW bead-on-plate welding. Welding in the World, 68(9), 2293-2309. https://doi.org/10.1007/s40194-024-01754-0

[26] Mattera, G., Cao, Y., Zhang, Y. M., & Nele, L. (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/s40194-025-02280-3

[27] Zhao, J., Li, J., Ding, Y., Bai, L., & Cheng, Y. (2024). Quantitative evaluation of keyhole stability during laser welding using optical coherence tomography. Welding International, 38(3), 225-238. https://doi.org/10.1080/09507116.2024.2317959