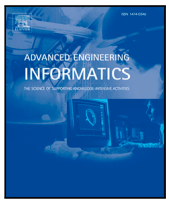
Advanced Engineering Informatics 69 (2026) 103936

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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



Full length article

Automatic measurement system for aircraft rivet flushness on surfaces empowered by multi-modal large-scale models

Kaijun Zhang [a], Zikuan Li [d,e], Xiaojie Zheng [a], Chenghan Pu [c], Tianhao Huang [a], Jun Wang [a,b,*]

[a] College of Mechanical and Electrical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016, Jiangsu, China
[b] State Key Laboratory of Mechanics and Control for Aerospace Structures, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016, Jiangsu, China
[c] School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, 211106, Jiangsu, China
[d] The Sixty-third Research Institute, National University of Defense Technology, Nanjing, 210007, Jiangsu, China
[e] Laboratory for Big Data and Decision, National University of Defense Technology, Changsha, 410073, Hunan, China

ARTICLE INFO

Keywords:
Rivet flushness
Automated measurement system
Fusion network
Deep learning
Rivet segmentation

ABSTRACT

In aircraft manufacturing, rivet flushness is defined as the average vertical height difference between the rivet head surface and the adjacent skin surface. This geometric parameter directly impacts the aerodynamic performance and the stealth characteristics of next-generation fighter aircraft. In this paper, we present a comprehensive automated rivet flushness measurement system enabling rapid, precise quantitative assessment of each rivet's flushness. Firstly, an integrated handheld device is designed to enable automated multimodal data acquisition, real-time evaluation, and output display. Secondly, a SAM-Adapter enhanced multi-modal Fusion Network (SAM-AFNet) is proposed for precise rivet segmentation, enabling zero-shot generalization to unseen rivet types and rapid adaptation to industrial environments with minimal training. Finally, a height map optimized rivet flushness calculation method delivers enhanced accuracy, better aligning with ground truth. Experimental results demonstrate state-of-the-art segmentation metrics and superior measurement accuracy. Notably, the system offers a fully deployable industrial solution, has already validated in production environments. The source code is available at: https://github.com/Kai1jun/SAM-AFNet.

1. Introduction

Aircraft surfaces incorporate numerous assembled rivets, where each rivet may protrude or recess relative to its ideal installation position due to assembly-induced deviations. This deviation, formally defined as **rivet flushness** in aeronautical standards, is quantified as the average vertical height difference between the rivet head surface and the adjacent skin surface. As shown in Fig. 1, excessive positive or negative rivet flushness critically compromises laminar flow stability, thereby increasing drag. Furthermore, these deviations impair surface smoothness, consequently increasing the radar cross-section (RCS) [1]. Collectively, these effects degrade aircraft's aerodynamic performance and compromise the stealth characteristics. Consequently, developing more efficient and high-accuracy measurement methods for rivet flushness aligns with the requirements of next-generation fighter aircraft [2].

Nevertheless, significant challenges persist in rivet flushness measurement. Firstly, the vast quantity (millions) and diverse, continuously renewing and evolving morphology (Fig. 2) complicate precise identification. Secondly, detecting micrometer-level flushness on millimeter-scale rivet heads represents a substantial metrological challenge, especially across large-scale aircraft structures spanning dozens of meters. Thirdly, current methods predominantly rely on manual inspection [3], resulting in notably low efficiency. Consequently, accurately and rapidly segmenting all rivet regions presents the primary challenge.

Current rivet segmentation methodologies primarily fall into two categories: 2D image-based and 3D point cloud-based approaches. Jiang et al. [4] employed vision-based detection on 2D images to localize rivet structures on bridge surfaces, demonstrating significant robustness to varying camera angles. However, this approach suffers substantial degradation under strong specular reflection (high glare). Conversely, Xie et al. [5] focused on 3D point clouds, identifying rivets based on their density-based saliency within the structure for subsequent fitting. Nevertheless, this method encounters challenges with modern aircraft point cloud data, where rivet density variations are often less pronounced, and it exhibits poor robustness against point cloud noise. To summarize, prevailing techniques face significant hurdles: 2D methods are inherently sensitive to illumination variations, leading

* Corresponding author.
E-mail address: wjun@nuaa.edu.cn (J. Wang).

https://doi.org/10.1016/j.aei.2025.103936
Received 1 July 2025; Received in revised form 28 September 2025; Accepted 29 September 2025
Available online 13 October 2025

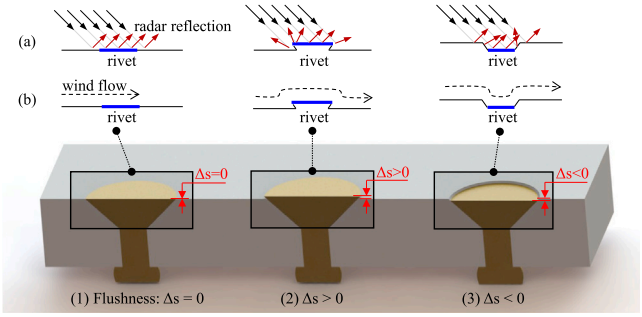


Fig. 1. Schematic illustration of rivet flushness. Case (1): Zero flushness, meeting the standard. Cases (2) and (3): Excessive positive or negative rivet flushness, respectively. These deviations critically compromise aircraft stealth characteristics (a) and degrade aerodynamic performance (b).

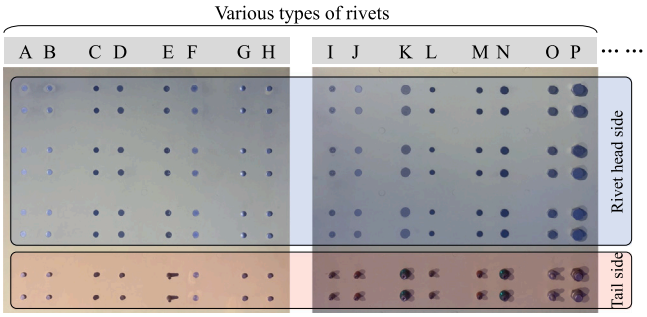


Fig. 2. Diverse rivet structures (partial view). Rivet structures on airplanes are complex and numerous, and evolving with new aircraft models, presents significant challenges for rivet identification and localization.

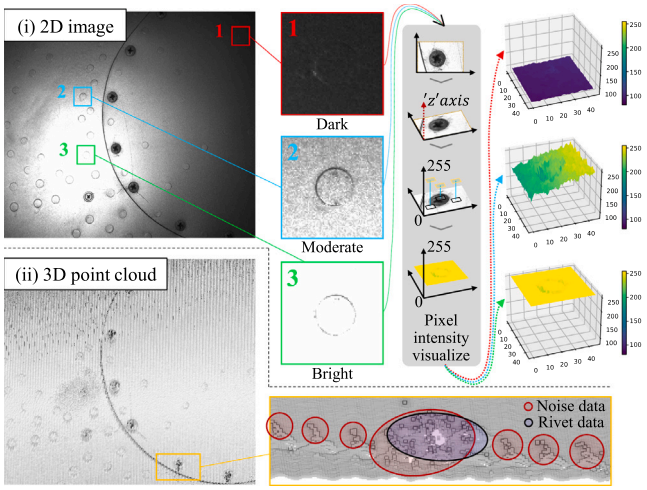


Fig. 3. Limitations of existing methods. (i) For 2D images: Due to high reflectivity, the visualization of pixel-wise height values for rivets across differently illuminated regions exhibits poor feature discriminability and low measurement consistency. (ii) For 3D point clouds: Significant noise arises due to high reflectivity on the aircraft surface, making direct segmentation of the 3D modality challenging.

to severe performance degradation under strong glare (Fig. 3(a)). In contrast, 3D methods are highly susceptible to point cloud noise and often fail when rivet features are weak or indistinct (Fig. 3(b)). These modality-specific flaws underscore a fundamental limitation: relying on a single modality hinders robust rivet recognition and segmentation, thereby motivating the need for multimodal fusion. However, existing fusion methods (e.g., UniSeg [6], LoGoNet [7]), designed for LiDAR datasets at scales of thousands of points, are unsuitable for industrial scenarios requiring high-resolution, massive data processing at millions of points.

In this paper, to address the aforementioned challenge, we present a comprehensive automated rivet inspection system designed for rapid and precise quantitative assessment of each rivet's flushness as the final measurement output. Specifically, (1) an integrated handheld device is proposed, enabling automated multimodal data acquisition, real-time evaluation, and output generation. It acquires multimodal data using structured light and transmits it to an integrated high-performance edge computing module for processing. (2) To achieve accurate rivet segmentation, we propose a multi-modal fusion network specifically designed for rivet region segmentation. This network operates within the edge computing module, leveraging the Segment Anything Model (SAM) [8] for robust zero-shot segmentation capabilities on previously unseen rivet categories. Furthermore, adapter modules are strategically integrated, enabling swift adaptation to new scenes by training only a minimal number of parameters. We term this network the SAM-Adapter Enhanced Multi-modal Fusion Network (SAM-AFNet). (3) To

achieve precise quantitative measurement of rivet flushness, a height map optimized rivet flushness calculation method is developed. Experimental results demonstrate that the proposed method attains leading segmentation performance metrics and high accuracy. The designed system enables successful engineering implementation, with validation in actual production environments confirming its deployability and effectiveness.

Overall, the main contributions in this paper can be summarized as follows:

(1) We present an integrated handheld device enabling automated data acquisition, evaluation, and output, achieving fully integrated rivet flushness measurement.
(2) To achieve high-quality segmentation of rivets, we propose SAM-AFNet, a novel framework that integrates multi-modal data alignment with the SAM module and adapter module. This design not only ensures robust zero-shot capability for unseen rivet categories but also enables rapid convergence with minimal training data.
(3) To ultimately achieve precise quantitative measurement of individual rivet flushness, we develop a novel, enhanced rivet flushness calculation methodology.
(4) We propose a complete solution for rivet flushness measurement in real-world industrial settings. Our method has been validated in actual production environments, offering a deployable and effective methodology.

2. Related works

2.1. High-precision rivet inspection in aviation manufacturing

Rivet inspection plays a critical role in aerospace manufacturing, as the quality of riveted joints directly impacts aircraft structural integrity and aerodynamic performance [9–11]. Conventional non-destructive testing methods [12–15], including ultrasonic testing, radiographic inspection, eddy current detection, and magnetic particle examination, have been widely adopted in industrial practice to identify microscopic cracks and interfacial defects at rivet connections. However, the enhanced stealth performance requirements of next-generation fighter aircraft dictate exceptional aerodynamic smoothness and electromagnetic wave scattering consistency between rivet heads and skin surfaces [16], elevating flushness as a pivotal quality metric especially for advanced riveting processes like friction forge riveting [17]. Existing contact measurement techniques, such as micrometer-based approaches, suffer from inherent limitations including surface coating abrasion, secondary deformation risks, and single-point evaluation errors [11,18], rendering them inadequate for comprehensive characterization of flushness on complex-curvature riveted surfaces where sloping head faults critically degrade joint strength [19,20].

Furthermore, traditional optical inspection methods [21–24] are significantly compromised under conditions of strong specular reflection (high glare) due to the reflective nature of rivet heads. This results in edge blurring and excessive noise, critically hindering the rapid and precise full-field morphological reconstruction of rivets on aircraft skin. Therefore, the integration of multi-modal data fusion approaches is urgently required in this research domain.

2.2. Recent advances in 2D detection and 3D point cloud processing

Most 2D semantic segmentation models predominantly employ convolutional neural networks (CNNs) [25,26], including architectures such as AlexNet [27], VGG [28], and ResNet [29]. While CNN architectures progressively aggregate contextual information through cascaded convolutions and pooling, the effective utilization rate of this information remains significantly below theoretical potential [30]. The advent of vision Transformers [31], such as ViT [32], DETR [33] and MaskFormer [34], etc, has revolutionized representation learning by inherently modeling long-range dependencies via self-attention, enabling contextually rich feature extraction and seamless fusion of local and global cues. However, effectively translating these capabilities into robust, real-world deployment scenarios presents a distinct challenge. Although efficient models like YOLO [35,36] are widely deployed industrially, they exhibit critical limitations: heavy reliance on large-scale labeled datasets, extended training cycles, and notably poor robustness on out-of-distribution (OOD) [37] samples or challenging scenarios, such as highly reflective surfaces. Current research predominantly focuses on benchmark supremacy or parameter efficiency, leaving the critical challenge of rapidly adapting foundation models to diverse, specialized operational domains largely unexplored.

Significant research focuses on deep learning for 3D point cloud understanding, where PointNet [38] pioneered direct processing of unordered point sets and VoxelNet [39] enabled convolutional operations through volumetric representations. Sparse convolution [40] subsequently alleviated critical memory and computational bottlenecks inherent to dense voxel grids, facilitating practical deep learning on moderate-scale point clouds (e.g., indoor scenes, LiDAR). However, processing truly massive point clouds ($> 10^7$ points) with sparse representations still pushes modern GPU hardware limits (A100/H100), straining memory (batch/grid size) and hindering efficient parallel computation [41], particularly for large/complex models. Balancing computational efficiency against the high spatial precision required for industrial applications—such as millimeter-scale detection or recognition of fine structures—remains challenging, further compounded by point cloud noise characteristics which impede robust industrial deployment.

Table 1
Architectural comparison of SAM-AFNet with prior methods. Unlike conventional approaches requiring full encoder fine-tuning, our SAM-AFNet integrates frozen 2D layers with lightweight adapters, eliminating the need to train the entire encoder. Additionally, it incorporates point cloud data into the prompt encoder, enabling multimodal fusion. These structural innovations make SAM-AFNet particularly effective for rivet detection scenarios.

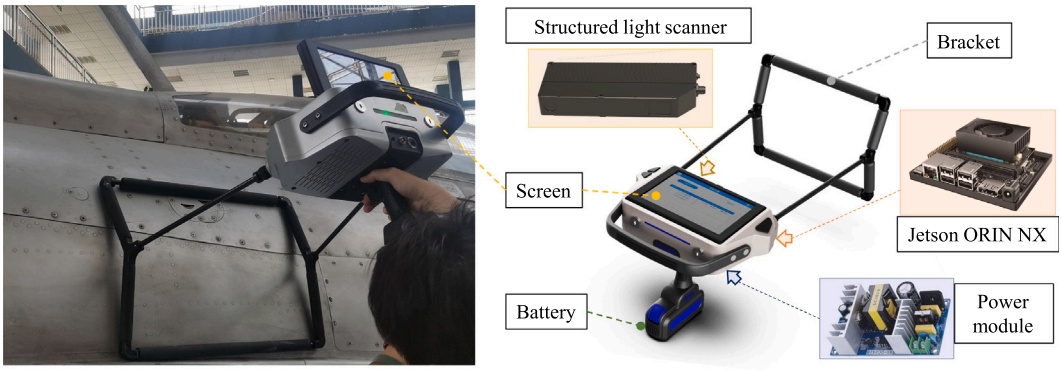
Model	Input Type	Image Encoder	Prompt Encoder	Mask Decoder
SAN	image	2D+adapter	—	2D
SAM3D	point cloud	3D	—	3D
SAM2	video	2D	2D	2D
SAM-Med3D	3D volumetric image	3D	3D	3D
SAM-AFNet (Ours)	image (2D) + point cloud (3D)	2D+adapter	3D (point cloud)	2D & 3D

Frozen module Tunable module

2.3. Directly input-output framework for robust 3D segmentation

While modern machine learning thrives on large-scale annotated data, real-world applications demand recognizing novel concepts without labeled examples. Multimodal foundation models like Contrastive Language-Image Pre-training (CLIP) [42] and Segment Anything Model (SAM) [8] have emerged as pivotal 2D frameworks, exhibit impressive zero-shot ability to new image distributions and tasks. Recently, SAM demonstrates exceptional zero-shot generalization capabilities in 2D vision tasks, sparking extensive research efforts toward its 3D extensions, such as SAM3D [43] and SAM-Med3D [44]. Existing methods suffer from information degradation due to input dimensionality reduction and output projection, failing to meet the sub-millimeter precision requirements for aviation rivet inspection, as Table 1 highlights the gap between these approaches and our work in input type, prompt encoder and mask decoder design.

Our work presents an end-to-end 3D SAM architecture by developing a multimodal adaptation framework that completely resolves projection distortion through direct 3D point cloud and image inputs with 2D/3D mask outputs. This framework significantly enhances



(a) Hardware and measurement scenarios (b) Detailed structural diagram explanation

Fig. 4. Our designed rivet data scanning handheld device. The device is capable of simultaneously capturing both image and point cloud modality data. Additionally, it integrates an NVIDIA chip, enabling edge computing for real-time processing and analysis. The screen can output the measurement results of the flushness of each rivet region.

SAM's robustness for aviation rivet inspection tasks, particularly in addressing two key challenges: (1) cross-category generalization to novel rivet configurations, and (2) effective fusion of multisensor data under real-world manufacturing conditions.

3. System description

To achieve accurate and rapid segmentation of rivet regions for analysis and subsequent rivet flushness measurement, we designed an automated rivet flushness measurement system. First, we developed a handheld device integrating a structured-light camera to support high-precision fused data acquisition. Next, we proposed the SAM-Adapter Enhanced Multi-modal Fusion Network (SAM-AFNet), which simultaneously incorporates 2D and 3D information as input and ultimately outputs the segmented each rivet region. Finally, we designed a high-precision rivet flushness evaluation method. These contributions are detailed in the following three subsections.

3.1. Rivet scanning handheld device design

To facilitate the acquisition of aircraft surface rivet data, we designed a portable handheld device optimized for ease of operation and high-precision measurement. As illustrated in Fig. 4(a), the device is used in a typical data acquisition scenario, while Fig. 4(b) provides a detailed structural diagram explanation. The system is composed of a structured light scanner, a bracket, a Jetson Orin NX computing platform, a touchscreen, a battery, and a power distribution module.

To simultaneously capture both high-resolution images and dense 3D point clouds of rivet surfaces, the scanner employs a high-precision binocular structured light system with adjustable angles. The bracket ensures the scanner operates within its optimal working distance from the rivet surface, thereby maximizing data accuracy and consistency. To enable real-time processing and high-throughput edge computation, the system integrates a Jetson Orin NX, which utilizes its heterogeneous architecture and CUDA acceleration to efficiently perform point cloud preprocessing directly on the device. Additional functional modules include a touchscreen display for interactive control and result visualization, a battery for portable power supply, and a dedicated power management module for regulated distribution across components.

Through the integrated design of above-mentioned components, this system operates independently: structured light acquires high-precision 3D data, while an edge detection module incorporates our specially designed high-precision rivet region segmentation network (3.2 section) and optimized flushness calculation algorithm (3.3 section). The end result is an end-to-end process providing real-time visualization of each rivet's flushness measurement on-screen.

3.2. SAM-adapter enhanced multi-modal fusion network

3.2.1. Overview of network

The limitations observed in current rivet detection methodologies can be summarized as follows: reliance on features derived from a single data modality inherently impedes robust rivet recognition. Specifically, 2D methods are highly sensitive to adverse lighting conditions, while 3D methods struggle with weak feature distinctiveness and significant noise vulnerability, as illustrated in Fig. 3. To address these challenges while accommodating the characteristics of rivets—being numerous and evolving with new aircraft models. In this section, we propose the SAM-Adapter Enhanced Multi-modal Fusion Network (SAM-AFNet). The network robustly integrates visual and geometric features, while maintaining fast adaptability to new scenarios and strong zero-shot capability.

Our network is an end-to-end architecture that directly processes both structured light image data and point cloud data as inputs, and outputs the point cloud corresponding to the rivet region. The key feature of SAM-AFNet lies in its hybrid architecture, which leverages the Segment Anything Model (SAM) to deliver robust zero-shot segmentation capabilities for previously unseen rivet categories. Strategically integrated adapter modules enable rapid adaptation to new scenes by training only a minimal number of parameters, ensuring efficient fine-tuning without extensive computational overhead. The overall architecture of the network is illustrated in Fig. 5. For the image data, we employ a 2D image encoder to extract and encode image features. The architecture is designed with multiple SAM layers, organized hierarchically and interconnected through multi-layer adapters. These adapters are designed to enable rapid adaptation to novel categories by fine-tuning only a small number of parameters, while keeping all SAM layers frozen to preserve their zero-shot segmentation capability. For the point cloud data, we adapt the point cloud to fit the prompt structure of the SAM structure paradigm, which is subsequently encoded by a 3D prompt generator. In the 3D prompt generator, we encode the geometric features of the input point cloud and adapt them to serve as prompts within the traditional SAM framework. These encoded geometric features are integrated into the feature prompt encoder, enabling the system to leverage 3D geometric information for enhanced representation.

Ultimately, our network outputs both 2D image segmentation rivet masks and 3D point cloud rivet regions as the final results. These outputs are designed to support downstream inspection tasks, providing comprehensive and multi-modal representations for further analysis.

3.2.2. 2D image encoder

As previously mentioned, to fully leverage the advantages of SAM model, such as its ability to segment anything and its zero-shot

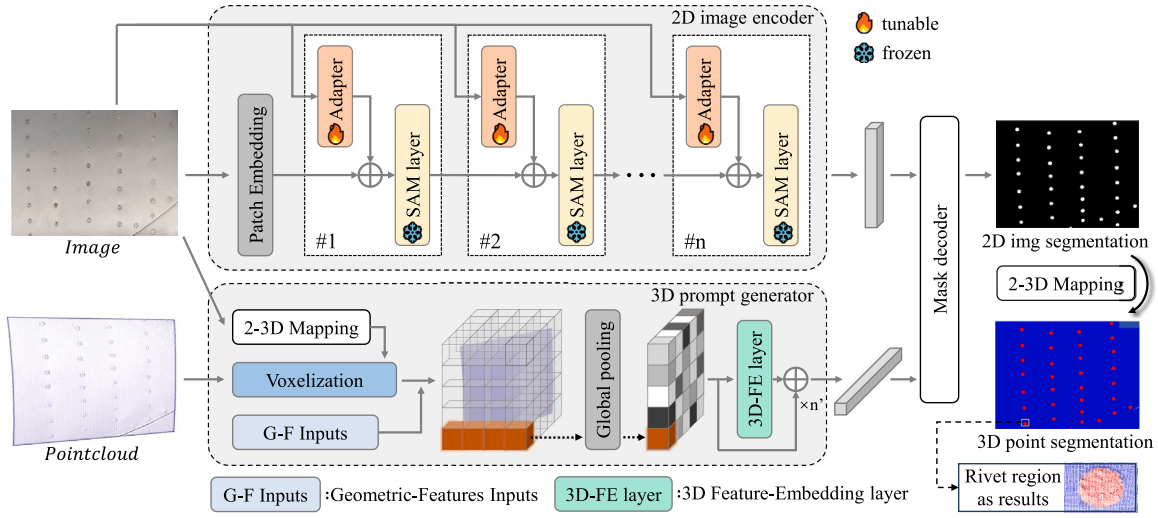


Fig. 5. Overview of our proposed method. The architecture follows the SAM-based paradigm, where the 2D image encoder is responsible for extracting semantic features from the image, while a 3D prompt generator encodes geometric information from the point cloud to guide the segmentation process. Ultimately, the network outputs both 2D image segmentation masks and 3D point cloud rivet regions.

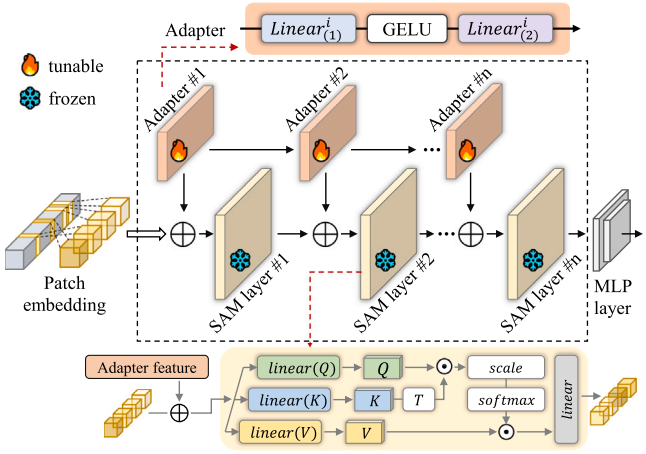


Fig. 6. 2D image encoder structure. We design an adapter structure built upon multiple SAM layers, which are hierarchically organized and interconnected via multi-layer adapters. During training, the SAM layers are kept frozen, and only the adapter structures are fine-tuned.

capabilities, we have designed an adapter within the 2D encoder module. Adapter enables rapid adaptation by fine-tuning only a small number of parameters, while keeping SAM layers frozen to preserve zero-shot segmentation capability.

The 2D encoder module structure with adapter is shown in Fig. 6. Specifically, the 2D image is first input and processed through patch embedding. It is then passed into multiple SAM layers, which are organized hierarchically and interconnected through multi-layer adapters. Specifically, the adapter takes the input f_i and obtains the prompt P_i:

$$P_i = Linear^i_{(2)}(GELU(Linear^i_{(1)}(f_i))), \tag{1}$$

where the $Linear^i_{(1)}(\cdot)$ denotes the tuned linear layers used to generate task-specific prompts for each adapter. $GELU(\cdot)$ is the GELU activation function [45]. $Linear^i_{(2)}(\cdot)$ is a shared upward projection layer across all adapters, which is responsible for adjusting the dimensions of transformer features.

The SAM layer primarily consists of the ViT attention module, which can be represented as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right)V, \tag{2}$$

where d_k denotes the dimensionality of the key vectors. Through the fusion of multi-layer adapter and SAM layer features, the output is fed into the mask encoder via a MLP layer.

3.2.3. 3D prompt generator

Due to its spatial structural characteristics, point cloud data is not affected by lighting conditions or scanner angles. To address the issue of image reflectivity, we adopt the SAM paradigm by replacing the prompt encoder with one adapted for 3D point clouds, incorporating spatial information from the point clouds to enhance performance.

The 3D prompt generator structure, as shown in Fig. 7, begins with a 2D-3D mapping. Given an input point cloud, it is first divided into 3D voxels. To establish the mapping between the voxels and the point cloud, the structured light scanner's projection matrix $\mathcal{M}$ is used to calculate reference points $\mathbf{Img}_i$ in the image plane from the center c_i of each voxel,

$$\mathbf{Img}_i = \mathcal{M} \cdot c_i, \tag{3}$$

where $\mathcal{M}$ is the product of the camera intrinsic matrix and the extrinsic matrix, obtained through camera calibration.

To improve the performance of spatial 3D segmentation and enhance the representation of 3D features, spatial geometric features are

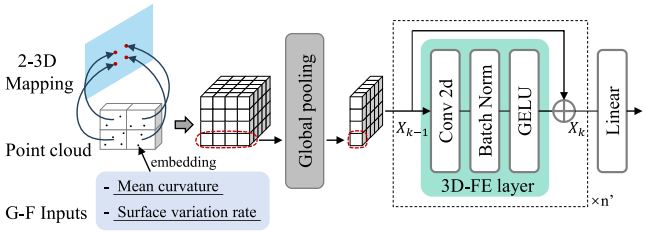


Fig. 7. 3D prompt generator structure. The embedded geometric features are used as prompts in the prompt encoder to guide the segmentation of rivet regions.

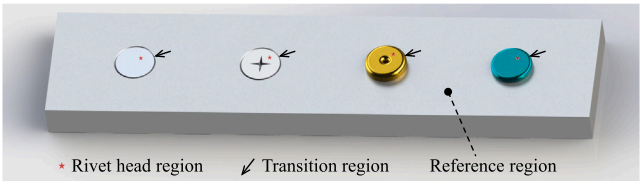


Fig. 8. Schematic diagram of different rivet head regions. It shows that the transition areas between classes may introduce ambiguity and negatively affect segmentation accuracy.

computed and embedded into each point $\mathbf{p}_i$ of the input point cloud. Specifically, this includes the mean curvature and surface variation rate. The mean curvature is a geometric quantity that describes the degree of curvature at a point on a surface or in three-dimensional space. It is calculated as the arithmetic mean of the two principal curvatures:

$$H(\mathbf{p}_i) = \frac{k_1(\mathbf{p}_i) + k_2(\mathbf{p}_i)}{2}, \tag{4}$$

here, $k_1(\mathbf{p}_i)$ and $k_2(\mathbf{p}_i)$ are the two principal curvatures computed by fitting a surface to the neighborhood of the point $\mathbf{p}_i$. The surface variation rate (SVR), a critical metric for quantifying local geometric irregularities in 3D point cloud analysis, is conventionally defined as the spatial variance of points relative to a locally fitted reference surface. Mathematically, it is expressed as:

$$\mathrm{SVR} = \frac{1}{N}\sum_{i=1}^{N} \left\|\mathbf{p}_i - \mathbf{p}_{\mathrm{fit}}\right\|^2, \tag{5}$$

where $\mathbf{p}_i$ represents the coordinates of the ith point within a neighborhood, $\mathbf{p}_{\mathrm{fit}}$ represents the reference point of the fitted plane or surface, $\|\cdot\|$ is the Euclidean norm, and N is the neighborhood point size.

After obtaining the image mapping relationships and the embedded geometric features, the voxel features are then processed. Considering the characteristic of the aircraft surface point cloud, which consists of a single layer, we first apply global pooling, similar to a projection approach, to reduce the dimensionality of the features. Next, the features are passed into the 3D Feature Embedding (3D-FE) layer, a multi-layer feature encoder. For the features at the kth layer, X_k, the computation is as follows:

$$X_k = \mathrm{concat}\left(X_{k-1}, GELU\left(BN\left(\mathrm{Conv2d}\left(X_{k-1}\right)\right)\right)\right),$$
$$\text{for } k = 1, 2, \ldots, n',$$

where concat combining the features from the previous and current transformations. $GELU(\cdot)$ is the GELU activation function. $BN(\cdot)$ indicates batch normalization. Conv2d$(\cdot)$ represents the 2D convolution operation. After passing through the 3D-FE layer, the features are ultimately adapted to match the dimensions of the mask decoder through a linear layer, completing the point cloud feature embedding for prompt encoding.

3.3. Height map optimized rivet flushness calculation

The proposed SAM-AFNet network outputs the point cloud of the rivet region, which can then be used to measure rivet flushness. Rivet flushness is an average estimation metric, and point cloud data allows for more accurate estimation results. However, the rivet region point cloud obtained through segmentation is not directly suitable for use. As shown in Fig. 8, rivet flushness only considers the deviation of the rivet head area relative to the reference region. The presence of the transition region, however, can affect the computation, and it is difficult for the network to distinguish this directly. This is a structural challenge inherent to the problem rather than a limitation of the network's performance.

To address the aforementioned issue, we propose a method that can accurately extract and optimize the rivet head region. First, for each segmented rivet head point cloud region, we estimate an average normal $\vec{N}$ as follows, which best represents the vertical direction of the rivet:

$$\vec{N} = \frac{1}{m}\sum_{i=1}^{m} \vec{n}\left(\mathbf{p}_i\right), \tag{6}$$

where $\vec{n}\left(\mathbf{p}_i\right)$ represents the normal vector corresponding to the point $\mathbf{p}_i$, which is either obtained from the original acquisition device or estimated based on the local neighborhood. m denotes the number of points within the region. In the next step, the height map is computed along the $\vec{N}$ direction by dividing it into small intervals, and the number of points within each interval is then counted. We select the height interval with the most points h_{max}, along with the points from the intervals within $[-\sigma, \sigma]$ of this range, (where σ is set to 2.5–5 times the precision of the scanner, $\sigma = 0.05$ mm in this paper), defined as following:

$$P_{\mathrm{head}} = \left\{\mathbf{p}_i \in P \mid h_{\max} - \sigma \le h_i \le h_{\max} + \sigma\right\}. \tag{7}$$

These extracted points from the small intervals represent the smooth surface of the rivet head, excluding the transition region and noise, as shown in Fig. 9. Finally, the reference region is determined by applying RANSAC to the neighborhood of each rivet head point cloud and fitting a plane $\Pi : ax + by + cz + d = 0$. The rivet flushness of the kth rivet region can be represented as follows:

$$Flu(k) = \frac{1}{\left|P_{\mathrm{head}}\right|}\sum_{p_i \in P_{\mathrm{head}}} D\left(\mathbf{p}_i, \Pi\right). \tag{8}$$

$\left|P_{\mathrm{head}}\right|$ is the number of points in the set P_{head} . $D\left(\mathbf{p}_i, \Pi\right)$ represents the distance between $\mathbf{p}_i$ and the plane Π.

4. Experiments

In this section, we validate the proposed method on both our dataset and a public benchmark dataset. We perform a performance

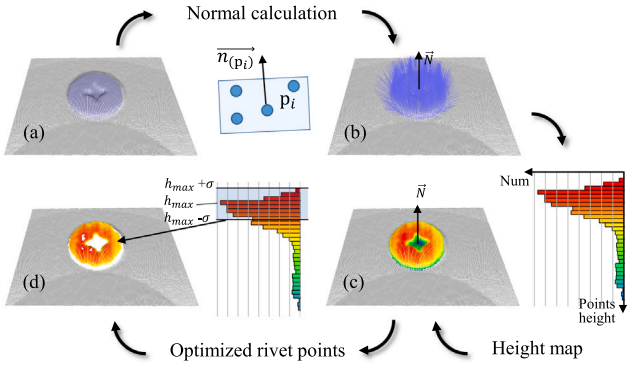


Fig. 9. Schematic diagram of the proposed height map optimization measurement method. (a) A typical irregular rivet with an unevaluable rivet head plane. (b) Visualization of all normal vectors and estimation of a average normal $\vec{N}$ as vertical direction. (c) Generation of height map based on direction $\vec{N}$ (horizontal axis: number of intervals; vertical axis: different heights, with red indicating the highest point, blue indicating the lowest point.) (d) Extraction of points satisfying $h_{\max} - \sigma \le h_i \le h_{\max} + \sigma$ as optimized rivet points.

Algorithm 1 Rivet Flushness Calculation

Require: Segmented rivet head point cloud regions P, neighborhood size σ
Ensure: Rivet flushness $Flu(k)$ for each rivet region
Step 1: Estimate Average Normal
for each region P **do**
  Compute average normal $\vec{N}$:
  $\vec{N} \leftarrow \frac{1}{m}\sum_{i=1}^{m} \vec{n}(\mathbf{p}_i)$ ▷ $\vec{n}(\mathbf{p}_i)$ is the normal of point $\mathbf{p}_i$
end for
Step 2: Compute Height Map
for each region P **do**
  Divide height small intervals along $\vec{N}$
  Count the number of points in each interval
  Find the interval with the most points, $h_{\max}$
  Extract points within $[h_{\max} - \sigma, h_{\max} + \sigma]$:
  $P_{\text{head}} \leftarrow \{\mathbf{p}_i \in P \mid h_{\max} - \sigma \le h_i \le h_{\max} + \sigma\}$
end for
Step 3: Fit Reference Plane
for each region P **do**
  Apply RANSAC to the neighborhood of P_{head}
  Fit a plane $\Pi : ax + by + cz + d = 0$
end for
Step 4: Calculate Rivet Flushness
for each region P **do**
  Compute flushness $Flu(k)$:
  $Flu(k) \leftarrow \frac{1}{|P_{\text{head}}|}\sum_{\mathbf{p}_i \in P_{\text{head}}} D(\mathbf{p}_i, \Pi)$ ▷ $D(\mathbf{p}_i, \Pi)$ is the distance from $\mathbf{p}_i$ to Π
end for
return $Flu(k)$ for all rivet regions

evaluation by comparing our approach with state-of-the-art (SOTA) instance segmentation algorithms. Subsequently, ablation studies are conducted to investigate the impact of various factors on the performance of our method. Finally, we assess the accuracy and efficiency of our approach on a finely processed high-precision standard block.

4.1. Experimental setup

4.1.1. Dataset

AISD dataset. To facilitate the measurement of rivet flushness on aircraft surfaces, we have constructed a dedicated dataset, the Aircraft Rivet Surface Dataset (AISD). This dataset was acquired using our rivet scanning handheld device integrated with a structured light scanner, which, based on binocular imaging principles, allows for the synchronized capture of image and point cloud data within 1 to 2 s. Due to the calibration of the structured light system, the captured images and point cloud data have precise corresponding relationships. The collected data covers various aircraft surface regions and different rivet categories.

The AISD dataset consists of 312 images and corresponding point cloud data. Each image has a resolution of 1236 × 1032 pixels, while each point cloud contains 1236 × 1032 points. The images and point cloud data are stored on the device's hard drive for subsequent processing and analysis. The scenes and sample data collected are shown in Fig. 10(a), illustrating the diversity of surface regions and rivet categories. For model training, the dataset is partitioned into training, validation, and test subsets with a ratio of 7:2:1, see also in Table 2.

CT airplane dataset. To validate our methodology, we also employ the publicly available CT Airplane Dataset [46], which comprises seven high-resolution volumetric sub-scans of a ME 163 aircraft obtained through X-ray computed tomography. Notably, these sub-volumes capture intricate structural details including rivet assemblies, as illustrated in Fig. 10(b).

A key distinction arises from data modalities: conventional images and point clouds are inherently limited to external surfaces, whereas CT leverages X-ray penetration to reconstruct volumetric internal structures. For comparative analysis, we employ volume rendering to generate 2D/3D visualizations from CT data, while surface data are discretized into point clouds via structured sampling. Since the 2D images are rendered from 3D data, this rendering process inherently preserves the correspondence between pixels and point clouds. Due to the limited size of the CT dataset, we directly evaluate each method by testing with weights pre-trained on the AISD dataset.

4.1.2. Evaluation metrics

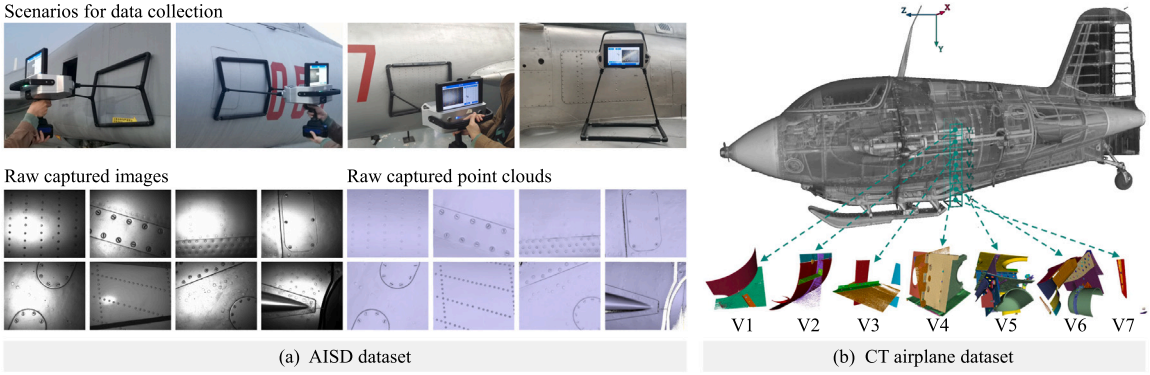
Selecting appropriate evaluation metrics is crucial for assessing model performance. In this study, to effectively evaluate the network's segmentation results and the accuracy of rivet flushness measurement, we have designed two sets of evaluation metrics, referred to as **Metrics #1** and **Metrics #2**.

Metrics #1: To evaluate the performance of the network in segmenting the rivet category in both image and point cloud data, we adopted four commonly used evaluation metrics: Intersection over Union (IoU), Precision (P), Recall (R), and the Matthews Correlation Coefficient (MCC). These metric formulations are presented as follows:

$$IoU = \frac{TP}{TP + FP + FN}, \tag{9}$$

$$Precision(P) = \frac{TP}{TP + FP}, \tag{10}$$

$$Recall(R) = \frac{TP}{TP + FN}, \tag{11}$$



(a) AISD dataset (b) CT airplane dataset

Fig. 10. Overview of our proposed method. (a) The AISD dataset, which is collected using a handheld rivet scanning device, capturing data from rivet structures of various aircrafts. (b) The CT airplane dataset. It is a public dataset of airplane ST scan data, containing numerous rivet structures.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \tag{12}$$

where true positives (TP) are the pixels or points that are present in both the predicted result and the ground truth (GT). False positives (FP) refer to those present in the predicted result but not in the GT. True negatives (TN) are those that are neither included in the predicted result nor in the GT and false negatives (FN) are those not in the predicted result but in the GT.

IoU measures the overlap between the predicted and ground truth regions. High precision indicates the excellent performance of the model. Recall represents the ability to recognize all the true value pixels or points. MCC is a statistical measure used to evaluate the performance of segmentation models, particularly in situations involving imbalanced distributions.

Besides, we also propose a dual-component evaluation metric, that is, $Num(E/M)$, where the error term (E) counts false positive rivets (detected but absent in ground truth) and the error term (M) enumerates localization failures (IoU < 30% with ground truth). In industrial applications, M represents critical detection failures that compromise inspection reliability [46].

Metrics #2: To evaluate the accuracy of rivet flushness calculations, we employed two key evaluation metrics: the Mean Absolute Error (MAE) and the Mean Squared Error (MSE). These metrics serve as robust indicators of measurement precision and are formally defined as follows:

$$\text{MAE} = \frac{1}{k}\sum_{i=1}^{k}|y_i - \hat{y}_i|, \tag{13}$$

$$\text{MSE} = \frac{1}{k}\sum_{i=1}^{k}(y_i - \hat{y}_i)^2, \tag{14}$$

where y_i denotes the measured flushness value of the ith rivet, $\hat{y}_i$ represents the ground truth value of the ith rivet, and k be the total number of rivets. The MAE is the average of the absolute errors between the measured values and the ground truth values. The MSE is the average of the squared errors between them.

4.1.3. Implementation details

The hardware used in this experiment includes an AMD EPYC 7443 CPU and an NVIDIA GeForce RTX A6000 GPU, with the system configured for CUDA 11.8. The network environment is set up with Python 3.8 and PyTorch 2.1. The hyperparameters used during training are provided in Table 2.

4.2. Comparison

In this section, since the structured light sensor simultaneously captures rivet data in both 2D image and 3D point cloud modalities, each can serve as input for rivet segmentation. To comprehensively compare existing methods, we divide the experiments into two parts: Comparative analysis on 2D images and Comparative analysis on 3D point clouds.

Table 2
Training configuration parameters for the experiments.

Configuration parameters	Values
Optimizer	Adam
Initial learning rate	0.001
Voxel size	16×16×16
Batch size	4
Epoch	300
Train: Validation: Test	7:2:1
Adapter & SAM layer (n)	8
3D-FE layer (n')	5

4.2.1. Comparative analysis on 2D images

To validate the effectiveness of our method in 2D image segmentation tasks, we conducted extensive comparative experiments against six historically significant state-of-the-art pixel-level segmentation methods: ANN [47], SETR [48], Mask2Former [49], SAN [50], SAM [8] and SAM2 [51]. For dataset selection, comprehensive evaluations were performed on our self-constructed AISD dataset. To further assess the generalizability of the trained model, we evaluated its zero-shot transfer performance on the publicly available CT airplane dataset without any fine-tuning, thereby rigorously testing its robustness to domain shifts. Since the segmentation task is formulated as a binary classification problem at the pixel level, we adopted Metric #1 as the evaluation metrics. The comparative results are presented quantitatively in Table 3 and qualitatively in Fig. 11.

The comparative result analysis is as follows. the ANN-based method introduces innovations primarily aimed at improving computational efficiency, yet its performance lags significantly behind other approaches in both quantitative metrics and qualitative visualizations. In contrast, transformer-driven architectures like SETR and Mask2Former demonstrate distinct improvements: SETR enhances generalization through expanded receptive fields, while Mask2Former employs cross-attention mechanisms to extract discriminative local features. Experimental results confirm that Mask2Former's design philosophy aligns more closely with our method's requirements. However, these three approaches remain constrained by their foundation in traditional fully convolutional network (FCN) architectures trained on fixed datasets. The emergence of large-scale models, exemplified by SAN's integration of CLIP [52] and culminating in SAM, marks a paradigm shift. Notably, such large models achieve substantial superiority in cross-domain generalization metrics (e.g. MCC) on the CT airplane

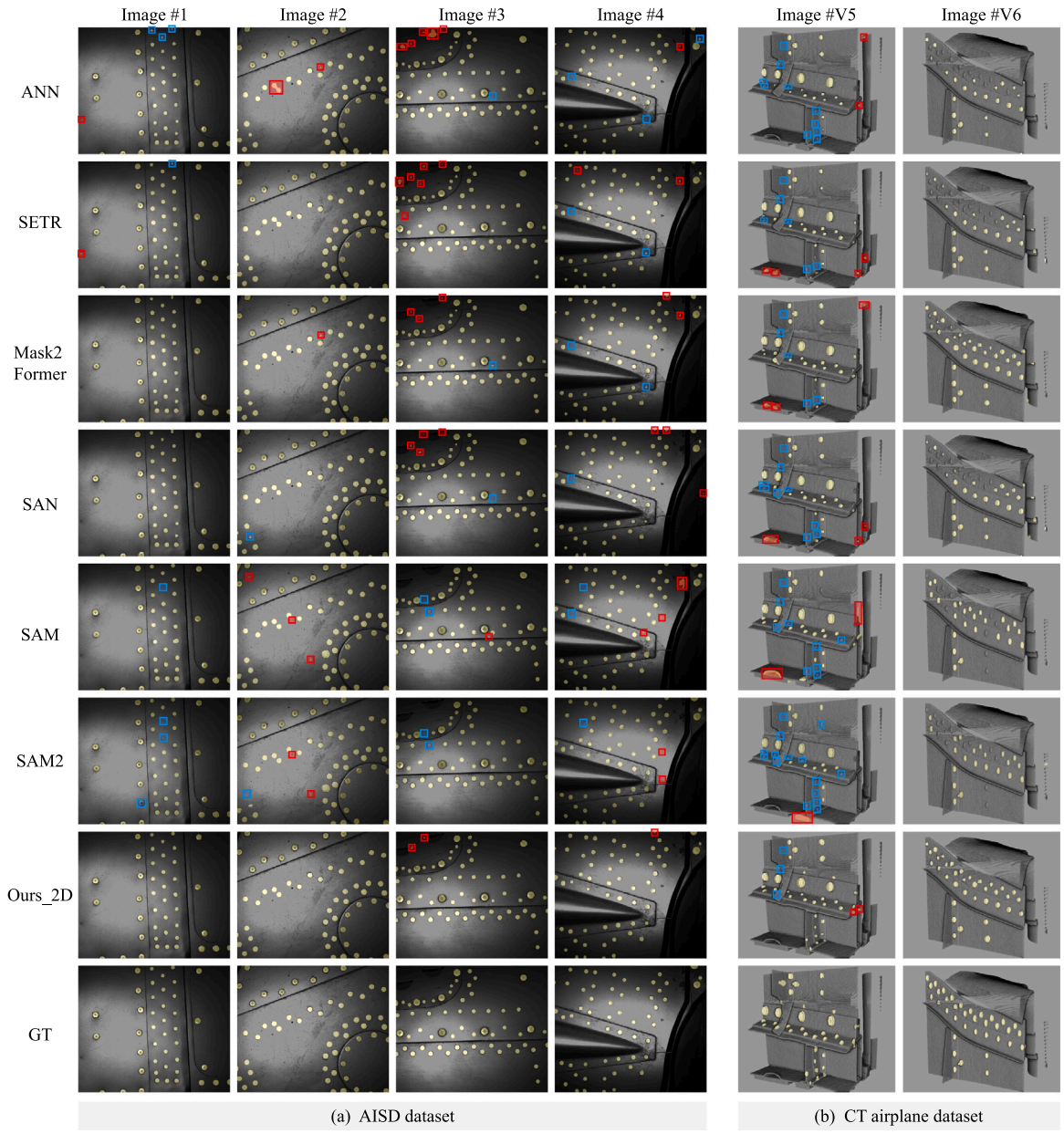


Fig. 11. Comparison of 2D image segmentation results on the AISD dataset (a) and the CT airplane dataset (b). The yellow regions indicate the predicted rivet masks. For visualization purposes, the display has been enhanced. Regions marked with " ▫ " represent the error term rivets (not present in the ground truth), while " ▫ " indicates missed rivet regions where the predicted mask has an IoU less than 30% with any ground truth annotation. This corresponds to the Num (E/M) metric used in our evaluation metrics.

Table 3
Quantitative comparison of 2D image segmentation results using both the AISD dataset and the CT airplane dataset.

Methods	Backbone	AISD dataset					CT Airplane dataset				
		Num (E/M)	IoU (%)	P (%)	R (%)	MCC (%)	Num (E/M)	IoU (%)	P (%)	R (%)	MCC (%)
ANN	ResNet-101	2.3/1.9	74.75	82.50	88.94	84.91	2.5/12.0	33.71	**78.93**	37.03	53.14
SETR	ViT-L/16	2.5/0.8	78.03	83.76	91.96	87.14	4.0/10.0	34.86	71.96	40.28	52.75
Mask2Former	ResNet-101	1.4/0.7	77.46	83.41	91.55	86.75	3.0/3.5	37.00	67.89	44.68	53.95
SAN	ViT-B/16	1.7/0.7	76.33	81.56	92.21	86.05	3.5/8.0	46.69	69.78	58.66	62.89
SAM	ViT-B/16	1.9/1.1	80.13	81.99	**97.16**	88.70	2.5/7.5	48.04	74.01	58.34	**64.51**
SAM2	ViT-L/16	1.3/1.8	80.27	83.50	95.48	88.73	**2.0**/11.5	44.40	77.58	51.04	61.96
Ours_2D	ViT-B/16	**0.3/0.6**	**81.58**	**87.03**	92.58	**89.27**	**2.0/1.5**	**48.25**	72.14	**59.41**	64.45

dataset, outperforming conventional methods by significant margins, particularly in handling heterogeneous data distributions.

The introduction of SAM2 represents an advancement over the original SAM framework. While SAM2 retains the core architecture of its predecessor, its primary innovation lies in the incorporation of temporal dimensions, enabling robust cross-frame object tracking in video sequences through contextual continuity. However, comparative analysis on the SA-1B benchmark reveals that SAM2 does not exhibit

Fig. 12. Comparison of 3D point cloud segmentation results on the AISD dataset (a) and the CT airplane dataset (b). The red regions represent rivet areas segmented by different methods.

Table 4
Quantitative comparison of 2D image segmentation results using both the AISD dataset and the CT airplane dataset.

Methods	AISD dataset				CT Airplane dataset			
	IoU (%)	P (%)	R (%)	MCC (%)	IoU (%)	P (%)	R (%)	MCC (%)
PointNet++	34.34	39.31	73.01	50.04	20.52	31.18	37.48	31.71
PAConv	24.06	41.94	34.65	34.72	14.70	65.79	15.78	30.21
RRCNet	38.59	43.43	77.27	54.08	35.96	**85.09**	38.17	56.07
SAM3D	80.13	81.99	**97.16**	88.70	48.04	74.01	58.34	**64.51**
Ours_3D	**81.58**	**87.03**	92.58	**89.27**	**48.25**	72.14	**59.41**	64.45

statistically significant improvements over SAM in single-frame image segmentation tasks, as demonstrated in [51]. Nevertheless, owing to its training on an expanded dataset compared to SAM, SAM2 achieves marginal performance gains in single-frame inference metrics, with higher IoU and MCC metrics on the AISD dataset.

In contrast, the proposed method adopts SAM as its baseline. Although our approach achieves consistently superior performance across a range of evaluation metrics, it records a slightly lower recall rate compared to SAM. In view of this, qualitative analyses reveal that SAM generates segmentation masks with marginally more accurate and well-defined edge delineations. Nevertheless, this minor difference in edge precision has negligible impact on the accuracy of individual rivet segmentation. Importantly, our method demonstrates a marked improvement in the Num (E/M) metric—a key indicator of detection effectiveness in industrial scenarios. This metric measures the ratio of correctly identified elements to missed detections, serving as a direct determinant of the algorithm's practical utility in real-world manufacturing applications.

4.2.2. *Comparative analysis on 3D point clouds*

The effective integration of deep learning with 3D point cloud data has enabled direct detection tasks using raw point clouds as input. To evaluate the advantages of our method over pure 3D approaches in rivet detection, we conducted comparative experiments with four representative models: PointNet++ [53], PAConv [54], RRCNet [55], and SAM3D [43]. This 3D point cloud binary classification task mirrors our 2D comparison experiments, employing identical datasets and evaluation metrics, with results detailed in Table 4 and Fig. 12.

Experimental analysis reveals distinct performance characteristics. PointNet++, as the most classical deep learning approach for point clouds, achieves satisfactory recall rates according to quantitative metrics. However, qualitative visualization demonstrates significant noise in its predictions. While PAConv generates cleaner outputs with reduced noise, its performance on rivet segmentation remains suboptimal due to insufficient feature representation of rivets. RRCNet, specifically designed for rivet segmentation, incorporates feature map projection and CNN receptive fields to localize rivet regions. Nevertheless, it

Table 5
Statistical comparison of 2D image segmentation performance on the AISD and CT Airplane datasets, including mean IoU, 95% confidence intervals (CI), and significance tests against baselines. Significance codes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

AISD Dataset (IoU %)		
Method	Value	95% CI
SAM	80.13	[79.21, 81.05]
SAM2	80.27	[79.35, 81.19]
Ours_2D	**81.58**	[80.72, 82.44]
p-value:	0.015* (vs SAM), 0.022* (vs SAM2)	
CT Airplane Dataset (IoU %)		
Method	Value	95% CI
SAM	48.04	[46.78, 49.30]
SAM2	44.40	[43.10, 45.70]
Ours_2D	**48.25**	[47.02, 49.48]
p-value:	0.152 (vs SAM), 0.004** (vs SAM2)	

shows limited generalization to novel rivet types and tends to misclassify other high-contrast features. All three methods exhibit notable noise sensitivity, primarily attributable to Z-axis fluctuations in scanned data caused by high reflectivity. In contrast, SAM3D adopts SAM as its backbone, projecting 2D segmentation results into 3D space through mapping relationships. However, being fundamentally 2D-based with single-frame projection, its core remains rooted in 2D segmentation paradigms.

Overall, in the domain of rivet segmentation, both qualitative and quantitative results demonstrate the clear advantages of 2D-based methods. Compared to their 3D counterparts, 2D approaches offer superior performance, with sharper boundaries, lower noise levels, and more favorable evaluation metrics. These findings highlight the rationality and effectiveness of employing 2D methods for rivet prediction. Nevertheless, 3D methods are not without merit. For instance, RRCNet achieves partial predictions for each rivet region, indicating the potential value of 3D features. Given the inherent sensitivity of 3D representations to geometric characteristics of rivets, this study proposes a novel approach that embeds 3D geometric information to guide 2D predictions through a prompt mechanism integrated into SAM. Experimental results show a substantial reduction in both false positives and false negatives, validating the effectiveness of the proposed 3D prompt encoder in enhancing 2D detection accuracy.

4.2.3. Statistical comparison with competitive baselines

To rigorously assess statistical reliability, we compare our method with the two most competitive approaches, SAM and SAM2, and report the mean IoU along with the corresponding 95% confidence intervals (CI) for all methods. Pairwise significance tests are further conducted against these baselines. As shown in Table 5, our Ours_2D approach achieves the highest IoU on both AISD and CT Airplane datasets, with statistically significant improvements over SAM ($p = 0.015$) and SAM2 ($p = 0.022$) on AISD, and over SAM2 ($p = 0.004$) on CT Airplane. These results demonstrate that the performance gains are consistent and unlikely to be due to random variation.

4.3. Ablation study

4.3.1. Ablation analysis of network modules

In this subsection, we conduct ablation experiments to systematically evaluate the effectiveness of the proposed modules. Specifically, the study investigates the individual and combined contributions of the 2D image encoder, the 3D prompt generator, and the adapter module, providing a comprehensive analysis of each component's impact on the overall performance of the network. The metrics are illustrated in Table 6.

Table 6
Ablation analysis of different modules in our network.

Modules			Performance metrics			
2D img	3D prompt	Adapter	IoU(%)	P(%)	R(%)	MCC(%)
✓			80.13	81.99	97.16	88.70
✓	✓		80.73	83.57	95.38	88.89
✓		✓	78.77	85.58	90.79	87.71
✓	✓	✓	81.58	87.03	92.58	89.27

Table 7
Parameters of different size ViT backbone.

Backbone	Embedding dims	Blocks	Attention head	Params
ViT-Base	768	12	12	83 M
ViT-Large	1024	24	16	307 M
ViT-Huge	1280	32	16	632 M

Table 8
Ablation analysis of different size of ViT backbone.

Backbone	IoU (%)	P (%)	R (%)	MCC (%)
ViT-Base	81.58	87.03	92.58	89.27
ViT-Large	81.88	87.29	92.65	89.45
ViT-Huge	81.98	87.44	92.65	89.53

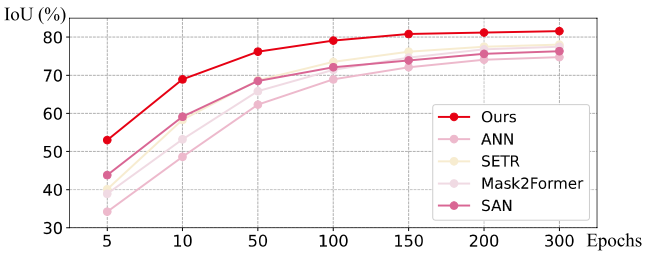
We begin with a baseline configuration that utilizes a pure 2D image encoder composed of a single-layer SAM encoder based on the original Vision Transformer (ViT) architecture [32], along with the standard SAM decoder. On this basis, we evaluate the effectiveness of the proposed 3D prompt generator and the cross-modal adapter module. To assess the impact of the 3D prompt generator, geometric features are encoded into learned embeddings and used as prompts to guide the segmentation process. While this enhances overall accuracy, it still leads to occasional false positives in regions with strong geometric features but no actual rivets. The adapter module, trained while keeping the backbone frozen, is designed to fine-tune the network specifically for rivet segmentation. This improves segmentation accuracy for rivet categories even without incorporating point cloud data; however, due to strong surface reflections, the recall rate drops significantly.

In summary, the proposed method leverages the adapter to enable fine-tuning for diverse rivet types while the 3D prompt generator effectively mitigates the recall degradation caused by high reflectivity, resulting in a more robust and accurate rivet segmentation framework.

4.3.2. Ablation analysis on different ViT backbone

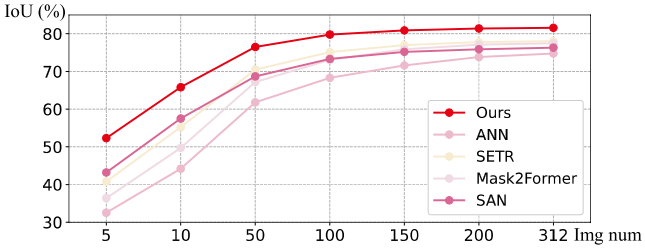
In our method, the 2D image encoder adopts the SAM layer structure, which is based on the Vision Transformer (ViT) architecture [32]. ViT offers scalable performance through three well-established model variants: ViT-Base, ViT-Large, and ViT-Huge, enabling flexible adaptation to various computational budgets and performance requirements. The detailed parameters of different ViT architectures are displayed in Table 7.

Table 8 presents the performance results for Metrics #1 across different backbone sizes. Overall, deeper network architectures tend to yield better performance. However, the performance differences remain marginal, with all variations being less than 0.5%. This can be attributed to the fact that the reasoning process of the ViT-based backbone is jointly guided by both the adapter modules and the 3D prompt generator, which mitigates the dependency on backbone depth alone. Therefore, in practical deployment scenarios where the pursuit of extreme performance metrics is not prioritized, the ViT-Base backbone offers a favorable trade-off between accuracy and computational efficiency.



ANN	34.24	48.63	62.31	68.97	72.18	74.15	74.75
SETR	40.16	58.32	68.75	73.59	76.24	77.53	78.03
Mask2..	38.92	53.27	65.84	71.46	74.63	76.82	77.46
SAN	43.85	59.13	68.52	72.17	73.96	75.64	76.33
Ours	53.07	68.94	76.23	79.15	80.82	81.27	81.58

Fig. 13. Ablation experiments of different methods across various training epochs.



ANN	32.51	44.28	61.85	68.39	71.64	73.87	74.75
SETR	40.82	55.23	70.56	75.19	77.04	77.80	78.03
Mask2..	36.47	49.73	67.29	73.16	75.98	77.12	77.46
SAN	43.25	57.53	68.76	73.37	75.28	75.95	76.33
Ours	52.34	65.87	76.52	79.83	80.96	81.49	81.58

Fig. 14. Ablation experiments of different methods on limited training sample number.

4.3.3. Ablation on limited training epochs and samples

In industrial inspection scenarios, deploying a network often incurs substantial costs in terms of data requirements and training time, which not only prolong the overall inspection cycle but also hinder rapid adaptation to specialized use cases. In this subsection, we design two sets of ablation experiments to evaluate the network's robustness under extreme conditions that may arise in industrial applications. Specifically, we assess performance under reduced training epochs and limited training samples, aiming to test the model's adaptability and efficiency in data-scarce and time-constrained settings. The experimental results are respectively illustrated in Fig. 13 and Fig. 14.

Fig. 13 presents the IoU scores across different training epochs (5, 10, 50, 100, 150, 200, and 300) on the AISD dataset, with the vertical axis representing the IoU. The results show that our method consistently achieves IoU scores above 75% at 300 epochs. Notably, even at only 5 epochs, the proposed approach demonstrates competitive performance due to its design, which involves fine-tuning a pre-trained ViT backbone and leveraging geometric features as prompts. This enables rapid convergence and strong generalization with minimal training. By approximately 50 epochs, our method achieves performance comparable to other approaches trained for 300 epochs. In contrast, other methods, particularly those based on SAN, rely heavily on convergence through extensive training and exhibit suboptimal performance. Furthermore, models using ResNet backbones generally underperform compared to those built upon ViT backbones, highlighting the effectiveness of our architecture choice.

Similar to Fig. 8, Fig. 9 illustrates the IoU scores on the AISD dataset with respect to different training set sizes under a fixed 300-epoch training schedule, where the horizontal axis indicates the number of training samples and the vertical axis represents the IoU. The results demonstrate that our method consistently outperforms other approaches even with a significantly reduced number of training samples. This superior performance is attributed to the large model's inherent capability to enable rapid adaptation through few-shot fine-tuning while maintaining high accuracy. Under extreme conditions, our approach achieves an IoU of 76.52% using only 50 training samples, which matches or exceeds the average performance of other methods trained on the full dataset of 312 samples, thereby enabling efficient and scalable deployment in practical applications.

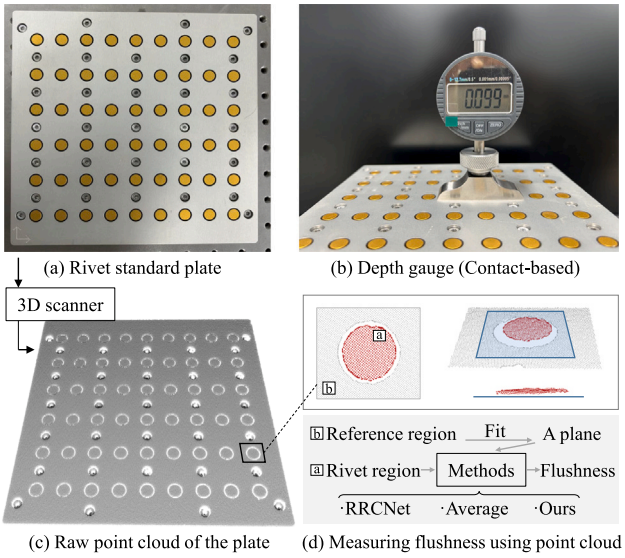


Fig. 15. Rivet standard plate and measurement. (a) Real specimen of the rivet standard plate. (b) Depth gauge (Contact-based) measurement. (c) Acquired raw point cloud of the rivet standard plate. (d) Evaluation of rivet flushness between reference surface and rivet head using different measurement methods.

4.4. Results of standard measuring plate

The previous experiments primarily focused on network performance research. Given the requirements of practical industrial applications, it is essential to ensure the accuracy and reliability of measurement results. To further verify the measurement precision of rivet protrusion/depression, we specially customized a standard measurement plate for rivet protrusion evaluation, whose design and structure are shown in Fig. 15(a). Manufactured using a high-precision CNC machining center combined with precision grinding processes, this

Table 9
Quantitative differences in results among various rivet flushness evaluation methods with significance test.

Rivet No.	GT	Depth gauge		RRCNet		Average		Ours	
		Value	Error	Value	Error	Value	Error	Value	Error
#1	−0.105	−0.002	0.103	−0.096	0.009	−0.099	0.006	−0.101	0.004
#2	0.054	−0.025	0.079	0.041	0.013	0.045	0.009	0.048	0.006
#3	−0.132	−0.086	0.046	−0.080	0.052	−0.098	0.014	−0.106	0.024
…	…	…	…	…	…	…	…	…	…
#53	−0.076	0.007	0.083	−0.037	0.039	−0.050	0.036	−0.056	0.018
#54	−0.047	−0.023	0.024	−0.024	0.023	−0.032	0.015	−0.035	0.012
MAE	–	–	0.0677	–	0.0376	–	0.0251	–	**0.0188**
MSE	–	–	0.0062	–	0.0021	–	0.0009	–	**0.0005**
p-value (vs ours)		< 0.001***		0.003**		0.013*		–	

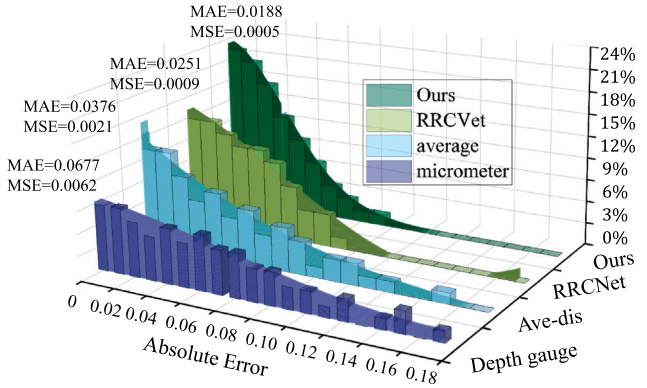


Fig. 16. Different measurement results. The rivet flushness evaluation method proposed in our study demonstrates optimal comprehensive performance.

standard plate achieves a surface roughness of Ra0.2μm, ensuring ultra-high precision in rivet head and surrounding areas. For each rivet head region and adjacent reference area, a high-precision coordinate measuring machine (CMM) was employed to measure 5 × 5 grid sampling points within a 2 mm diameter range at the rivet head center. Weighted averaging calculations were applied to eliminate local machining errors, controlling the comprehensive error of surface processing and measurement accuracy within 0.001 mm, which can be regarded as the ground truth.

To evaluate the accuracy of our method, we conducted comprehensive comparisons with four approaches: (1) Traditional manual measurement using a depth gauge Fig. 15(b); (2) RRCNet - a specialized method for rivet flushness measurement based on maximum-minimum-mean calculations from point clouds Fig. 15(c); (3) Point cloud averaging method, which calculates the mean Euclidean distance between segmented rivet head regions and reference areas; (4) Our proposed method.

Measurement results and absolute errors relative to ground truth (GT) for 18 rivet samples are shown in Table 9. Contact measurement (depth gauge with 0.001 mm precision) exhibited the highest MAE and MSE, indicating significant measurement fluctuations caused by random probe contact positions when rivet heads were deformed. Although specifically designed for rivet segmentation, RRCNet's MAE remained 48.9% higher than our method due to point cloud noise severely affecting the stability of maximum/minimum value extraction (e.g. 0.052 mm abnormal error in rivet #3 caused by noise interference). The point cloud averaging method reduced MAE to 0.0251 mm through global averaging but failed to filter tilted rivet point clouds and density variations at edge regions, introducing systematic errors (e.g. 0.036 mm error in rivet #17). Our method maintained errors $\leq$ 0.024 mm even for complex samples like #3 and #17, achieving the lowest MAE and MSE among all methods.

The data distribution characteristics of the four methods are illustrated in Fig. 16, visually demonstrating our method's significant advantages in stability and precision. Depth gauge measurements showed the widest distribution range (0.08–0.18 mm), with 22% of data points in the high-error zone (> 0.10 mm). Both RRCNet and averaging methods still had numerous data points clustered in the 0.04–0.06 mm error band, with measurement fluctuations caused by point cloud noise and systematic errors. Our method concentrated 84% of measurements within the ±0.02mm error band (vs. 56% for averaging method), with only 4% of data in high-risk zones (> 0.04 mm, 11% reduction compared to RRCNet), verifying the algorithm's robustness against complex rivet deformations and its capability to meet aviation riveting assembly's high-precision and stability requirements.

4.5. Visualization of geometric features in 3D prompt generator

To investigate the contribution of geometric features in the 3D prompt generator, we visualized these features, as illustrated in Fig. 17. The figure demonstrates that the intensity of the geometric features in the rivet head region is significantly higher than that in the surrounding flat areas, which makes the rivet head more prominent and easily distinguishable.

The significance of this design lies in the fact that geometric features can be considered as a form of weighting, directing the network's attention toward important feature locations. This is similar to the Attention mechanism, which reflects the degree of focus the model places on specific regions during decision-making. However, unlike Attention, which is a learned feature matrix, the geometric features in this study are derived from the geometric input of the point cloud itself.

These geometric features are encoded and integrated with the prompt encoder in the SAM paradigm, guiding the image to achieve better recognition in regions of interest. This approach is not only structurally sound but has also been experimentally proven to enhance the effectiveness of the 3D prompt generator. Additionally, the voxel-based prompt design eliminates the need for point-by-point analysis, significantly improving detection speed and reducing memory overhead.

4.6. Deployment performance analysis

In Sections 4.2 and 4.3, the experiments were conducted to evaluate performance, all deployed on a local NVIDIA A6000 (48 GB) platform. However, for the handheld device, the Jetson Orin NX (16 GB) was selected as the edge computing module. Due to its inherent limitations in power consumption, computational capacity, and memory, it is necessary to examine whether the investigated methods can be efficiently deployed on Jetson. In this section, we present an analysis of the deployability on the edge computing module, aiming to verify whether the algorithms can be effectively migrated to the handheld device.

To assess the practical deployability of the proposed system, we conducted a comprehensive evaluation of its resource consumption and inference efficiency across heterogeneous hardware platforms. The

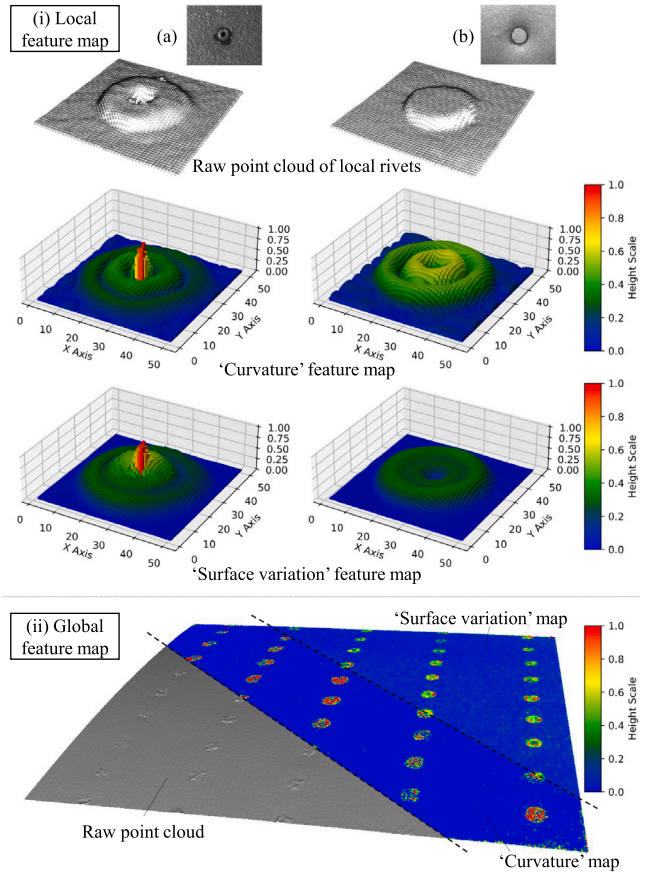


Fig. 17. Visualization of geometric features embedded into the 3D prompt generator. We compute and visualize two geometric features—curvature and surface variation. (i) Local rivet and corresponding geometric feature map. (ii) Global feature map. Both features exhibit significantly stronger responses in the rivet head regions compared to the surrounding flat surfaces. This distinction helps guide the network's focus toward critical areas, enhancing the segmentation of rivet structures.

Table 10
Quantitative comparison of deployment metrics across different methods. Throughput is measured in frames per second (FPS). Power consumption represents the average during inference. The rightmost column shows whether the method is deployable on Jetson.

Method	Input	Platform	Mem. (GB)	FPS	Power (W)	✓ ×
Mask2Former	2D	Jetson	2.24	0.76	62	✓
SAM2	2D	Jetson	2.29	11.7	65	✓
PointNet	3D	A6000	35.21	0.0017	104	×
PAConv	3D	A6000	42.60	0.0025	119	×
Ours	2D & 3D	Jetson	13.39	0.23	97	✓

results, summarized in Table 10, provide a quantitative comparison of memory usage, throughput, and average power consumption for different state-of-the-art methods. Since this evaluation primarily concerns deployability, we used the Jetson as the target testing device whenever feasible. However, for methods that could not be executed on Jetson due to memory overflow, we report their performance metrics obtained on the A6000 platform. In particular, the 3D methods (PointNet and PAConv) exceeded the available memory of Jetson, and thus their results are derived from the A6000 results.

The rightmost column of Table 10 indicates the deployability on Jetson. While our method requires relatively higher memory (13.39 GB) and power consumption (97 W) compared with purely 2D baselines,

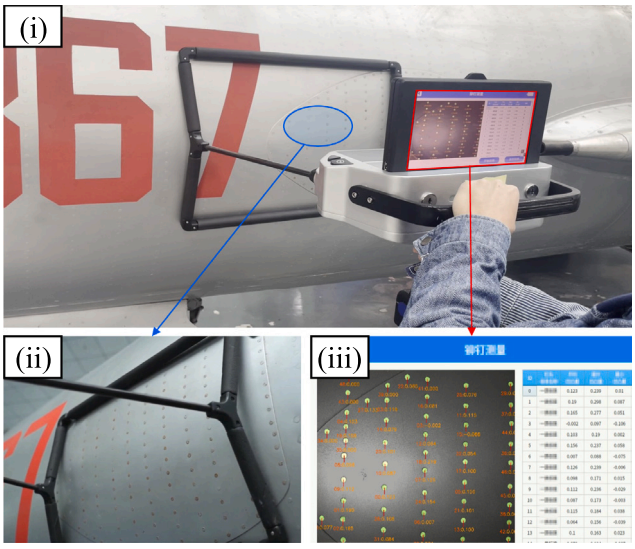


Fig. 18. Our automated rivet flushness measurement system has been successfully deployed in real-world aviation production environments. (i) Actual aircraft rivet measurement scenario. (ii) Rivets to be measured. (iii) The measured rivet flushness results were displayed on-screen, with quantitative flushness values annotated near each rivet.

it is the only approach that supports direct ingestion of 3D point cloud inputs while remaining deployable on Jetson, unlike PointNet and PAConv. The key enabler is our lightweight 3D prompt generator, which bypasses detailed local geometry modeling and instead encodes point clouds into compact geometric prompts, thereby complementing SAM with minimal additional information overhead.

5. Application

In this chapter, we illustrate the effectiveness of our proposed automated rivet flushness measurement system through real-world engineering applications. The system operates autonomously, achieving an all-in-one approach that encompasses data acquisition, real-time detection, and ultimately visualized measurement results for each rivet's flushness.

As shown in Fig. 18, we demonstrate our system's performance through a real-measurement case study on a specific fighter aircraft model. Our system enables comprehensive, full-coverage inspection across the entire field of view, with measured rivet flushness results displayed on-screen for immediate identification of non-conforming rivets. Each measurement cycle, requiring approximately 3 s, can process dozens of rivets simultaneously. This efficiency stems from our innovative 3D prompt-form encoding of point cloud data, which delivers high precision while significantly boosting throughput. The solution meets rigorous process standards for production-line deployment, providing a complete measurement workflow. Practical deployment has substantiated its substantial engineering value in real-world manufacturing environments.

6. Conclusion

In the aviation domain, rapid and precise quantitative assessment of each rivet's flushness aligns with the development requirements of new-generation fighter aircraft. This paper presents a comprehensive automated rivet flushness measurement system. Firstly, an integrated handheld device is designed to enable automated multimodal data acquisition, real-time evaluation, and output display. Secondly, to address the limitations observed in current rivet detection methodologies, which can be summarized as follows: reliance on features

derived from a single data modality inherently impedes robust rivet recognition, while accommodating the characteristics of rivets—being numerous and evolving with new aircraft models— we designed the SAM-Adapter Enhanced Multi-modal Fusion Network (SAM-AFNet). Finally, we propose a height map optimized rivet flushness calculation method. Validated through comprehensive experiments, our method achieves state-of-the-art segmentation accuracy and operational efficiency, outperforming existing approaches in both precision and generalization. The deployed handheld scanning system, combined with SAM-AFNet, bridges the gap between deep learning innovation and industrial practicality, offering a scalable and future-proof solution for next-generation aerospace manufacturing. This work underscores the viability of adaptive vision systems in addressing dynamic real-world engineering demands.

CRediT authorship contribution statement

Kaijun Zhang: Writing – original draft, Visualization, Methodology, Investigation. **Zikuan Li:** Writing – review & editing, Conceptualization. **Xiaojie Zheng:** Visualization, Validation, Data curation. **Chenghan Pu:** Writing – review & editing, Methodology. **Tianhao Huang:** Writing – review & editing. **Jun Wang:** Writing – review & editing.

Declaration of competing interest

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

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 92367301, No. 92267201, No. 52275493, No. 92160301).

Data availability

Data will be made available on request.

References

[1] W. Gao, Y. Wu, C. Hong, R.-J. Wai, C.-T. Fan, Rcvnet: A bird damage identification network for power towers based on fusion of RF images and visual images, Adv. Eng. Informat. 57 (2023) 102104.
[2] R. Chen, J. Yang, R. Xiao, Y. Hui, A. Xu, Q. He, Z. Xue, P. Guo, Scanned point cloud registration for localization of aircraft access panel and complementary frame with weak features, Adv. Eng. Informat. 65 (2025) 103209.
[3] F.-C. Hsu, C.-N. Chen, M.-D. Shieh, Using stepwise backward elimination to specify terms related to tactile sense for product design, Adv. Eng. Informat. 46 (2020) 101193.
[4] T. Jiang, G.T. Frøseth, A. Rønnquist, A robust bridge rivet identification method using deep learning and computer vision, Eng. Struct. 283 (2023) 115809.
[5] Q. Xie, D. Lu, K. Du, J. Xu, J. Dai, H. Chen, J. Wang, Aircraft skin rivet detection based on 3D point cloud via multiple structures fitting, Computer-Aided Des. 120 (2020) 102805.
[6] Y. Liu, R. Chen, X. Li, L. Kong, Y. Yang, Z. Xia, Y. Bai, X. Zhu, Y. Ma, Y. Li, et al., Uniseg: A unified multi-modal lidar segmentation network and the openpcseg codebase, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 21662–21673.
[7] X. Li, T. Ma, Y. Hou, B. Shi, Y. Yang, Y. Liu, X. Wu, Q. Chen, Y. Li, Y. Qiao, et al., Logonet: Towards accurate 3d object detection with local-to-global cross-modal fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 17524–17534.
[8] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, et al., Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026.
[9] B. Kamsu-Foguem, Knowledge-based support in non-destructive testing for health monitoring of aircraft structures, Adv. Eng. Informatics 26 (4) (2012) 859–869.
[10] M. Janovec, M. Smetana, M. Bugaj, Eddy current array inspection of zlin 142 fuselage riveted joints, Transp. Res. Procedia 40 (2019) 279–286.
[11] H.A. Khan, S. Uddin, S. Salik, A. Javaid, T.A. Khan, Z.U. Islam, A lock-in thermography based post-processing scheme for the detection of sub-surface rivet-related defects in aircraft structures, Mech. Syst. Signal Process. 228 (2025) 112423.
[12] T. Vrtač, M. Kodrič, M. Pogačar, G. Čepon, Dynamic substructuring-based identification of the rivet-squeezing force, Mech. Syst. Signal Process. 229 (2025) 112487.
[13] F.H. Stolze, K. Worden, G. Manson, W.J. Staszewski, Phase/frequency analysis of diffuse lamb-wave field for fatigue-crack detection in an aluminium multi-riveted strap joint aircraft panel, Measurement 224 (2024) 113884.
[14] P. Johnson, 7 - quality control and non-destructive testing of self-piercing riveted joints in aerospace and other applications, in: M. Chaturvedi (Ed.), Welding and Joining of Aerospace Materials, in: Woodhead Publishing Series in Welding and Other Joining Technologies, Woodhead Publishing, 2012, pp. 215–234.
[15] A. Katunin, K. Dragan, M. Dziendzikowski, Damage identification in aircraft composite structures: A case study using various non-destructive testing techniques, Compos. Struct. 127 (2015) 1–9.
[16] K. Schulte, S. O'Keefe, M. Rybachuk, S. Stegen, Assessing the structural design of fixed-wing airframes for next-generation electric aircraft, Aerosp. Sci. Technol. 163 (2025) 110224.
[17] M. Soylak, V. Erturun, Friction forge riveting of AA7075-T6 sheets with large diameter 2117-t3 rivets, Aircr. Eng. Aerosp. Technol. 95 (10) (2023) 1651–1658.
[18] F. Luo, Y. Zuo, Riveting damage behavior and mechanical performance investigation of CFRP/CFRP thin-walled single-lap blind riveted joints, J. Manuf. Process. 131 (2024) 129–140.
[19] Z. Silvayeh, M. Brillinger, J. Domitner, Deformation behavior of aluminum alloy rivets for aerospace applications, J. Mater. Res. Technol. 33 (2024) 3482–3491.
[20] M. Soylak, V. Erturun, Investigation of the effect of sloping head fault in solid riveting on the strength of the joint, Aircr. Eng. Aerosp. Technol. 94 (10) (2022) 1892–1897.
[21] T. Jiang, X. Cheng, H. Cui, C. Shi, Y. Li, Dual-camera-based method for identification and location of scattered self-plugging rivets for robot grasping, Measurement 134 (2019) 688–697.
[22] J. Newman, R. Ramakrishnan, Fatigue and crack-growth analyses of riveted lap-joints in a retired aircraft, Int. J. Fatigue 82 (2016) 342–349, 10th Fatigue Damage of Structural Materials Conference.
[23] J. Liu, G. Xie, J. Wang, S. Li, C. Wang, F. Zheng, Y. Jin, Deep industrial image anomaly detection: A survey, Mach. Intell. Res. 21 (1) (2024) 104–135.
[24] X. Zhang, M. Xu, X. Zhou, Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16699–16708.
[25] X. Ma, X. Zhang, M.-O. Pun, M. Liu, A multilevel multimodal fusion transformer for remote sensing semantic segmentation, IEEE Trans. Geosci. Remote Sens. (2024).
[26] J. Mu, S. Zhou, X. Sun, Ppmamba: Enhancing semantic segmentation in remote sensing imagery by SS2d, IEEE Geosci. Remote. Sens. Lett. 22 (2025) 1–5.
[27] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.
[28] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2015, pp. 730–734, arXiv:1409.1556.
[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778.
[30] A. Wang, H. Chen, Z. Lin, J. Han, G. Ding, Rep vit: Revisiting mobile CNN from ViT perspective, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2024, pp. 15909–15920.
[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[33] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, Springer, 2020, pp. 213–229.
[34] B. Cheng, A. Schwing, A. Kirillov, Per-pixel classification is not all you need for semantic segmentation, Adv. Neural Inf. Process. Syst. 34 (2021) 17864–17875.
[35] P. Jiang, D. Ergu, F. Liu, Y. Cai, B. Ma, A review of yolo algorithm developments, Procedia Comput. Sci. 199 (2022) 1066–1073.
[36] J. Terven, D.-M. Córdova-Esparza, J.-A. Romero-González, A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas, Mach. Learn. Knowl. Extr. 5 (4) (2023) 1680–1716.
[37] J. Yang, K. Zhou, Y. Li, Z. Liu, Generalized out-of-distribution detection: A survey, Int. J. Comput. Vis. 132 (12) (2024) 5635–5662.
[38] C.R. Qi, H. Su, K. Mo, L.J. Guibas, Pointnet: Deep learning on point sets for 3d classification and segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 652–660.
[39] Y. Zhou, O. Tuzel, Voxelnet: End-to-end learning for point cloud based 3d object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4490–4499.

[40] Y. Yan, Y. Mao, B. Li, Second: Sparsely embedded convolutional detection, Sensors 18 (10) (2018) 3337.
[41] Y. Tan, K. Han, K. Zhao, X. Yu, Z. Du, Y. Chen, Y. Wang, J. Yao, Accelerating sparse convolution with column vector-wise sparsity, Adv. Neural Inf. Process. Syst. 35 (2022) 30307–30317.
[42] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, J. Yan, Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, 2021, arXiv preprint arXiv:2110.05208.
[43] Y. Yang, X. Wu, T. He, H. Zhao, X. Liu, Sam3d: Segment anything in 3d scenes, 2023, arXiv preprint arXiv:2306.03908.
[44] H. Wang, S. Guo, J. Ye, Z. Deng, J. Cheng, T. Li, J. Chen, Y. Su, Z. Huang, Y. Shen, et al., Sam-med3d: towards general-purpose segmentation models for volumetric medical images, in: European Conference on Computer Vision, Springer, 2024, pp. 51–67.
[45] D. Hendrycks, K. Gimpel, Gaussian error linear units (gelus), 2016, arXiv preprint arXiv:1606.08415.
[46] L. Wang, G. Zhang, W. Wang, J. Chen, X. Jiang, H. Yuan, Z. Huang, A defect detection method for industrial aluminum sheet surface based on improved YOLOv8 algorithm, Front. Phys. 12 (2024) 1419998.
[47] Z. Zhu, M. Xu, S. Bai, T. Huang, X. Bai, Asymmetric non-local neural networks for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 593–602.
[48] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P.H. Torr, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6881–6890.
[49] B. Cheng, I. Misra, A.G. Schwing, A. Kirillov, R. Girdhar, Masked-attention mask transformer for universal image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1290–1299.
[50] M. Xu, Z. Zhang, F. Wei, H. Hu, X. Bai, Side adapter network for open-vocabulary semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2945–2954.
[51] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al., Sam 2: Segment anything in images and videos, 2024, arXiv preprint arXiv:2408.00714.
[52] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PmLR, 2021, pp. 8748–8763.
[53] C.R. Qi, L. Yi, H. Su, L.J. Guibas, Pointnet++: Deep hierarchical feature learning on point sets in a metric space, Adv. Neural Inf. Process. Syst. 30 (2017).
[54] M. Xu, R. Ding, H. Zhao, X. Qi, Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3173–3182.
[55] Q. Xie, D. Lu, A. Huang, J. Yang, D. Li, Y. Zhang, J. Wang, Rrcnet: Rivet region classification network for rivet flush measurement based on 3-D point cloud, IEEE Trans. Instrum. Meas. 70 (2020) 1–12.