Science Discovery Computers
2026, Vol. 1, No. 1, pp. 1–14
https://doi.org/10.11648/j.sdcomput.20260101.11

SciencePG
Science Publishing Group

Research Article

AI-Driven Multi-Modal Vision Framework for Automated Detection and Quantification of Tube Blockages

Rahul Agnihotri*, Pallavi Wadhwa

Department of Artificial Intelligence and Robotics, Robotronix and Scalability Technology, Gurugram, India

Abstract

Heat exchanger tubes in power plants and refineries degrade through fouling, scaling, corrosion and wall thinning, and their internal condition governs both thermal efficiency and plant safety. Inspection of these assets is still performed largely by manual or semi-automated review of remote visual inspection footage, which is time-consuming, subjective, and prone to inconsistent defect interpretation between operators. The difficulty is compounded by the imaging environment itself: tube bores are narrow, illumination is supplied coaxially with the camera and falls off with depth, metallic surfaces produce strong specular reflections, and probe motion introduces blur. This work proposes an AI-driven multi-modal vision framework for automated detection, localization and quantification of tube blockages and surface degradation under these conditions. The framework integrates image preprocessing, deep feature learning, a hybrid convolutional neural network and vision transformer (CNN-ViT) detector, and a reinforcement learning agent that adapts probe traversal speed to the observed defect risk. A quantitative Tube Health Index (THI) is introduced to score severity from corrosion coverage, blockage ratio, defect depth and structural scaling, so that each tube is assigned an objective risk category rather than a binary defect flag. The framework was validated on an industrial dataset comprising 353 heat exchanger tubes, approximately 180 hours of footage and 12,400 labelled defect instances acquired under field conditions. The proposed hybrid model achieved 94.6% accuracy, 92.3% precision and 91.1% recall, outperforming conventional thresholding at 72.4% accuracy and a standalone CNN at 88.1% accuracy. Inference ran at 48 ms per frame on embedded hardware, and inspection time per tube fell by approximately 60% relative to manual practice. The results indicate that combining hybrid feature learning with adaptive scanning and quantitative health scoring offers a deployable pathway toward autonomous industrial non-destructive testing.

Keywords

Computer Vision, Deep Learning, Reinforcement Learning, Tube Inspection

1. Introduction

Industrial heat exchangers are critical assets in power plants, refineries, and chemical processing facilities, where thousands of tubes operate continuously under high temperature, pressure, and corrosive environments. Over time, these tubes develop defects such as corrosion, scaling, blockage, wall thinning, and cracking, which degrade thermal efficiency and may lead to catastrophic failures. Early detection and quantification of such defects are therefore essential for ensuring operational safety, minimizing downtime, and enabling predictive

*Correspondence: Rahul Agnihotri (rahula@rast.in), Rahul Agnihotri (Rahul.gnihotri@gmail.com)

Received: 27 January 2026; **Accepted:** 26 August 2026; **Published:** 18 September 2026

maintenance [1].

Traditional inspection techniques, including manual visual inspection and semi-automated non-destructive testing (NDT), are labor-intensive, subjective, and difficult to scale for large tube bundles. These methods often suffer from inconsistent defect interpretation, limited repeatability, and prolonged inspection durations during plant shutdowns. The challenge is further compounded by constrained tube geometries, low illumination, reflective metallic surfaces, and restricted accessibility, which reduce the effectiveness of conventional rule-based image processing approaches [2, 3].

Recent advances in computer vision and deep learning have demonstrated significant potential for automated defect detection in industrial inspection tasks. Convolutional Neural Networks (CNNs) enable robust feature extraction from noisy environments, while Vision Transformers (ViT) capture global contextual relationships within complex scenes. Reinforcement learning techniques further allow intelligent scanning strategies that optimize coverage and inspection efficiency. However, existing solutions often focus on laboratory datasets and lack validation under real-world industrial constraints [4, 5].

In this work, we present an AI-driven multi-modal vision framework for automated detection, localization, and quantification of tube blockages and surface degradation. The system integrates image preprocessing, deep feature learning, CNN–ViT hybrid architectures, reinforcement learning–based scan optimization, and a quantitative Tube Health Index (THI) for defect severity scoring. The framework is evaluated on a real industrial dataset comprising 353 vertically oriented tubes (30 m length, 32.66 mm diameter) captured using remote visual inspection cameras under practical field conditions.

The main contributions of this study are:

1) A scalable end-to-end vision pipeline for automated tube inspection
2) A hybrid CNN+ViT architecture for improved defect detection accuracy
3) Reinforcement learning–based scan optimization for efficient coverage
4) A quantitative Tube Health Index for objective defect assessment
5) Validation on a large real-world industrial dataset with deployment-ready performance

The proposed framework provides a practical pathway toward autonomous, reliable, and real-time inspection in industrial non-destructive testing environments.

2. Related Work

Automated defect detection for industrial assets has attracted significant attention in recent years due to the growing need for scalable and reliable non-destructive testing (NDT) solutions. Traditional inspection methods rely heavily on manual visual assessment, threshold-based image processing, or rule-driven feature engineering. Although these approaches are simple to implement, they are highly sensitive to illumination variations, noise, and operator subjectivity, which limits their robustness in real-world environments [6].

With the emergence of deep learning, Convolutional Neural Networks (CNNs) have become the dominant paradigm for visual inspection tasks. Early studies demonstrated that CNN-based classifiers significantly outperform handcrafted feature methods for detecting cracks, corrosion, and surface defects. Architectures such as ResNet and YOLO-based detectors have been widely adopted for real-time defect localization and object detection in manufacturing and infrastructure monitoring applications. These models provide improved feature representation but often struggle in highly constrained environments with motion blur, specular reflections, and long-range perspective distortion, which are typical in heat exchanger tubes [7, 8].

More recently, transformer-based vision models have been introduced to capture long-range dependencies and global contextual information. Vision Transformers (ViT) have shown superior performance compared to purely convolutional models in complex scenes by modeling spatial relationships more effectively. Hybrid CNN–ViT architectures further combine the local feature extraction capability of CNNs with the global reasoning strength of transformers, leading to improved generalization across varying defect scales and textures [9, 10].

In parallel, reinforcement learning (RL) has been explored to optimize robotic inspection and scanning strategies. RL-based agents learn adaptive navigation and coverage policies that reduce inspection time while maximizing defect discovery probability. Such approaches are particularly useful in constrained or elongated environments where exhaustive scanning is inefficient [11, 12].

Despite these advances, most existing studies evaluate algorithms using controlled laboratory datasets or flat-surface images. Limited research addresses inspection within long, narrow, reflective tubes, where lighting degradation, accessibility constraints, and severe class imbalance significantly impact detection performance. Furthermore, few works provide quantitative health scoring or deployment-ready solutions suitable for edge devices in industrial settings [13, 14].

To bridge these gaps, the proposed framework integrates CNN–ViT hybrid learning with reinforcement learning–based scan optimization and introduces a Tube Health Index (THI) for objective defect quantification. Unlike prior work, the system is validated on a large-scale real industrial dataset and designed for real-time deployment on embedded hardware, aligning closely with practical non-destructive testing requirements recommended by standards bodies such as IEEE and industry NDT guidelines.

3. Industrial Problem Definition

Industrial heat exchanger inspection demands automated, scalable, and objective assessment methods due to the large

number of tubes, confined geometries, and time-constrained shutdown windows. The inspected asset consists of 353 vertically oriented tubes, each measuring 30 m in length with an internal diameter of 32.66 mm, making manual or rule-based inspection inefficient and error-prone. To address these challenges, we propose an AI-driven multi-modal vision framework that combines image preprocessing, deep feature extraction, convolutional neural networks (CNN), vision transformers (ViT), reinforcement learning–based scan optimization, and quantitative health scoring to enable accurate and reliable defect detection under real-world industrial conditions.

3.1. Dataset

The experimental dataset was collected from an operational industrial heat exchanger deployed in refinery and power generation environments. The asset consists of 353 vertically oriented metallic tubes, each having a length of 30 m and an internal diameter of 32.66 mm. The narrow geometry, extended depth, and reflective metallic surfaces present significant challenges for reliable visual inspection.

Inspection data were acquired using a 21 mm high-definition Remote Visual Inspection (RVI) camera system equipped with integrated illumination, distance counter, and digital recording capability. Videos were captured at 1920 × 1080 resolution and 30 fps under varying lighting and contamination conditions. Each tube scan produced continuous footage from the inlet to outlet, resulting in approximately 8–10 minutes of video per tube.

The raw dataset comprises:

1) 353 tube videos
2) ~180 hours of inspection footage
3) ~3.5 million extracted frames
4) 12,400 labelled defect instances

Defects were annotated by Level-III certified inspectors and categorized into:

1) Blockage (foreign objects/debris)
2) Scaling/deposition
3) Corrosion pits
4) Wall thinning
5) Crack initiation
6) Surface deformation

Bounding boxes and pixel-level segmentation masks were generated using semi-automatic labeling tools. For model development, the dataset was partitioned into:

1) Training set – 70%
2) Validation set – 15%
3) Test set – 15%

To enhance generalization, data augmentation techniques including brightness variation, Gaussian noise injection, motion blur simulation, rotation, and synthetic occlusion were applied.

This dataset reflects real industrial variability rather than controlled laboratory conditions, ensuring that the proposed framework is validated under practical field constraints.

3.2. Inspection Constraints

Unlike laboratory-based defect detection tasks, industrial heat exchanger inspection is constrained by severe physical, environmental, and operational limitations. These constraints directly influence system design and algorithm selection.

3.2.1. Geometric Constraints

The tubes have a small internal diameter (32.66 mm), restricting the size of deployable sensors and preventing the use of bulky robotic platforms. Long tube lengths (30 m) introduce perspective distortion and progressive illumination loss, reducing visibility at deeper sections.

3.2.2. Illumination and Visual Challenges

The inspection environment exhibits:

1) Low and non-uniform lighting
2) Specular reflections from metallic surfaces
3) Condensation and moisture
4) Deposits and turbidity
5) Motion blur from probe movement

These factors degrade image quality and cause traditional thresholding or rule-based approaches to fail. Therefore, robust preprocessing and deep feature extraction are required.

3.2.3. Accessibility Constraints

Due to operational and safety restrictions:

1) No intrusive sensors can be inserted inside the tube
2) No couplant or liquid-based techniques are permitted
3) Inspection must be performed remotely
4) Scanning time per tube must remain under 10 minutes

Hence, the solution must rely exclusively on non-contact vision-based sensing.

3.2.4. Data Imbalance

Certain defect classes such as cracks occur rarely compared to scaling or corrosion. This class imbalance introduces bias during training and necessitates weighted loss functions and reinforcement learning–based sampling strategies.

3.2.5. Operational Constraints

Industrial shutdown windows are limited, requiring:

1) Near real-time inference (< 50 ms/frame)
2) Portable edge deployment
3) Minimal operator interaction
4) Automated reporting

Therefore, the framework is designed for edge-compatible deployment on embedded GPU systems (e.g., NVIDIA Jetson) and integrates automated scoring and dashboard visualization.

4. Proposed Workflow

The proposed system follows a multi-stage processing pipeline that converts raw inspection videos into quantitative defect assessments and actionable maintenance insights. The workflow is designed to operate reliably under industrial constraints such as narrow tube geometries, non-uniform illumination, and limited inspection time.

The overall pipeline consists of five major stages:

1) Data acquisition
2) Image preprocessing
3) Deep feature learning and defect detection
4) Reinforcement learning–based scan optimization
5) Quantitative scoring and reporting

Each stage is described below.

4.1. Data Acquisition

Inspection data are captured using a high-definition Remote Visual Inspection (RVI) camera inserted into each tube. Continuous video streams are recorded while the probe traverses the entire 30 m tube length. Frames are timestamped and indexed using a distance counter to enable spatial localization of detected defects.

This stage produces:

1) RGB video frames
2) positional depth information
3) inspection metadata

These inputs form the raw dataset for further processing.

4.2. Image Preprocessing

Raw inspection frames frequently suffer from illumination inconsistency, noise, blur, and reflections. To improve robustness, each frame undergoes preprocessing operations including [15]:

1) contrast enhancement (CLAHE)
2) noise suppression (Gaussian/Bilateral filtering)
3) motion blur reduction
4) lens distortion correction
5) frame normalization and resizing

The objective is to standardize visual quality and enhance defect visibility prior to deep learning inference.

4.3. Deep Feature Learning and Defect Detection

Pre-processed frames are passed through a hybrid CNN–ViT architecture for automated defect detection.

Stage 1 – CNN Backbone

Convolutional layers extract low-level features such as:

1) edges
2) textures
3) corrosion patterns
4) scaling boundaries

Stage 2 – Vision Transformer Encoder

Transformer blocks model long-range spatial dependencies and contextual relationships within the tube interior, improving detection of irregular or extended defects.

Stage 3 – Detection Head

The network outputs:

1) bounding boxes
2) class labels
3) confidence scores
4) segmentation masks

This design combines local texture sensitivity (CNN) with global reasoning capability (ViT), resulting in improved performance under noisy industrial conditions.

4.4. Reinforcement Learning–Based Scan Optimization

Uniform scanning may waste time on clean regions while missing critical areas. To address this, a reinforcement learning (RL) agent dynamically adjusts probe traversal speed and focus based on detected risk levels.

The agent:

1) slows down near suspected defects
2) increases frame density in high-risk regions
3) accelerates through defect-free zones

This adaptive strategy reduces inspection time while maximizing defect coverage.

4.5. Quantitative Health Scoring

Detected defects are aggregated to compute a Tube Health Index (THI), which provides an objective severity score.

The index integrates:

1) corrosion coverage
2) blockage percentage
3) defect depth
4) scaling intensity

Each tube is categorized into:

1) Healthy
2) Moderate risk
3) Critical

This enables maintenance teams to prioritize interventions.

4.6. Reporting and Visualization

Finally, results are exported to an analytics dashboard that displays:

1) defect heat maps
2) severity trends
3) historical comparisons
4) automated inspection reports

The system supports real-time deployment on edge devices, allowing operators to receive immediate feedback during inspection.

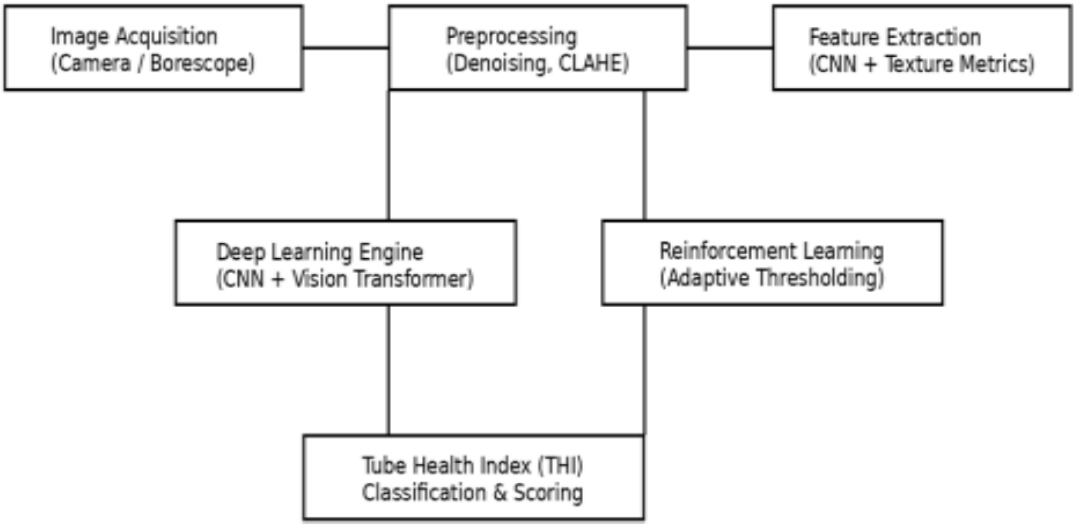

Figure 1. *Multi-modal inspection pipeline: acquisition → preprocessing → CNN/ViT → reinforcement learning scan optimization → health scoring → dashboard analytics.*

5. Methodology

This section describes the proposed hybrid deep learning architecture, training strategy, and optimization procedures used for automated detection and quantification of tube defects. The design prioritizes robustness to industrial noise, computational efficiency, and real-time deployability.

The overall model follows a three-stage structure:

1) Convolutional feature extraction
2) Transformer-based global context modeling
3) Multi-task detection and segmentation head

5.1. Network Architecture

5.1.1. CNN Backbone

A deep Convolutional Neural Network (CNN) is used as the primary feature extractor. The backbone consists of stacked convolutional layers with residual connections that capture [5]:

1) edges and boundaries
2) corrosion textures
3) scaling patterns
4) local structural irregularities

CNNs are particularly effective for learning fine-grained local features that characterize surface degradation. Downsampling layers reduce spatial resolution while increasing semantic richness.

Let the input frame be:

$I \in R^{H\times W\times 3}$

The CNN produces feature maps, as given in Eq. (1):

$$F_c = \phi_{cnn}(I) \tag{1}$$

Where $F_c \in \mathbb{R}^{h\times w\times d}$

5.1.2. Vision Transformer Encoder

While CNNs capture local information, long-range dependencies are critical for identifying extended defects such as cracks or blockages. Therefore, a Vision Transformer (ViT) encoder is employed [10].

The CNN feature map is divided into patches and linearly embedded, as given in Eq. (2):

$$P = \text{PatchEmbed}(F_c) \tag{2}$$

Multi-head self-attention then models global spatial relationships, as given in Eq. (3):

$$F_t = \phi_{vit}(P) \tag{3}$$

This enables:

1) global context awareness
2) improved detection of elongated defects
3) robustness to lighting inconsistencies

5.1.3. Detection Head

A multi-task detection head predicts:

1) bounding boxes
2) defect class labels
3) segmentation masks
4) confidence scores

Outputs are defined as:

where

B = bounding boxes,
C = classes,
M = masks,
S = confidence.

Non-maximum suppression is applied to remove duplicate detections.

5.2. Training Strategy

Dataset Split
1) Training: 70%
2) Validation: 15%
3) Test: 15%
Data Augmentation
To simulate industrial variability:
1) brightness/contrast jitter
2) Gaussian noise
3) blur simulation
4) random rotations
5) occlusion patches
This improves generalization to unseen field conditions.
Hyperparameters
Training was performed using:
1) Optimizer: Adam
2) Learning rate: 1×10^{-4} [16]
3) Batch size: 16
4) Epochs: 120
5) Input resolution: 512 × 512
6) Hardware: GPU-based training
Early stopping was applied based on validation loss.

5.3. Loss Functions

The network is trained using the composite multi-task objective of Eq. (4):

$$L = L_{det} + \lambda_1 L_{seg} + \lambda_2 L_{reg} \tag{4}$$

5.3.1. Classification Loss

Cross-entropy loss is used for defect classification, as given in Eq. (5):

$$L_{CE} = -\sum y \log(\hat{y}) \tag{5}$$

5.3.2. Localization Loss

Bounding box regression uses Smooth L1 loss [17], as given in Eq. (6):

$$L_{reg} = \text{SmoothL1}\left(B, \hat{B}\right) \tag{6}$$

5.3.3. Segmentation Loss [18]

Dice loss improves mask accuracy, as given in Eq. (7):

$$L_{seg} = 1 - \frac{2|M \cap \hat{M}|}{|M| + |\hat{M}|} \tag{7}$$

5.3.4. Class Imbalance Handling

To address rare defects (e.g., cracks), weighted focal loss is applied, as given in Eq. (8) [19]:

$$L_{focal} = -(1-p)^{\gamma} \log(p) \tag{8}$$

This prevents bias toward majority classes.

5.4. Reinforcement Learning Integration

A lightweight reinforcement learning agent optimizes probe traversal speed. The reward function is defined in Eq. (9):

$$R = \alpha \cdot \text{DetectionAccuracy} - \beta \cdot \text{InspectionTime} \tag{9}$$

The agent learns policies that increase frame sampling near suspected defects while reducing redundant scanning.

5.5. Inference and Deployment Efficiency

To enable field deployment:
1) model pruning [20]
2) mixed-precision inference
3) batch normalization folding
were applied to reduce latency.
Average inference time:
<50 ms/frame
This satisfies real-time industrial inspection requirements.

6. Mathematical Formulation

This section presents the formal mathematical framework underlying the proposed inspection system, including the learning objective, optimization strategy, reinforcement learning policy, and the Tube Health Index (THI) used for quantitative defect assessment.

Let an inspection video be represented as the sequence of frames in Eq. (10):

$$\mathcal{X} = \{x_1, x_2, \ldots, x_T\} \tag{10}$$

$x_t \in \mathbb{R}^{H\times W\times 3}$ denotes the RGB frame captured at time ttt.

The objective of the system is to learn the function in Eq. (11):

$$f_\theta : x_t \to y_t \tag{11}$$

where θ are learnable parameters and y_t contains predicted defect locations, classes, and severity measures.

6.1. Feature Extraction

The CNN backbone extracts spatial features, as given in Eq. (12):

$$F_c = \phi_{cnn}(x_t) \tag{12}$$

The transformer encoder captures global contextual dependencies, as given in Eq. (13):

$$F_t = \phi_{vit}(F_c) \tag{13}$$

The final representation is given in Eq. (14):

$$F = F_c \oplus F_t \tag{14}$$

where $\oplus$ denotes feature concatenation.

6.2. Detection Objective Function

The network performs multi-task learning consisting of:
1) classification
2) localization
3) segmentation
The composite loss is defined in Eq. (15):

$$L_{total} = L_{cls} + \lambda_1 L_{loc} + \lambda_2 L_{seg} \tag{15}$$

where λ_1, λ_2 are weighting coefficients.

Classification Loss

Cross-entropy loss is applied over the defect classes, as given in Eq. (16):

$$L_{cls} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{16}$$

Localization Loss

Bounding box regression uses Smooth L1 loss, as given in Eq. (17):

$$L_{loc} = \sum \text{SmoothL1}\left(B - \hat{B}\right) \tag{17}$$

Segmentation Loss

Dice loss is used for mask overlap, as given in Eq. (18):

$$L_{seg} = 1 - \frac{2|M \cap \hat{M}|}{|M| + |\hat{M}|} \tag{18}$$

6.3. Regularization

To prevent overfitting, L2 regularization is applied, as given in Eq. (19):

$$L_{reg} = \| \theta \|_2^2 \tag{19}$$

The final training loss is given in Eq. (20):

$$L = L_{total} + \lambda_3 L_{reg} \tag{20}$$

6.4. Reinforcement Learning Optimization

The inspection probe traversal is modeled as a Markov Decision Process (MDP):

(S, A, P, R)

where:
1) SSS = scan state (position, defect probability, speed)
2) AAA = actions (slow, normal, fast)
3) PPP = transition probability
4) RRR = reward

The reward function balances detection quality and time efficiency, as given in Eq. (21):

$$R = \alpha \cdot Acc + \beta \cdot Prec - \gamma \cdot FPR - \delta \cdot T \tag{21}$$

where:
1) Acc = accuracy
2) Precc = precision
3) FPR = false positive rate
4) T = inspection time

The policy is learned by maximizing the expected return of Eq. (22):

$$\pi^* = \underset{\pi}{\text{argmax}} \mathbb{E}[\sum_{t=0}^{T} \gamma^t R_t] \tag{22}$$

6.5. Tube Health Index (THI)

To quantify defect severity at the tube level, a composite Tube Health Index is defined.

Let:
1) C = corrosion coverage ratio
2) H = normalized wall thickness
3) D = defect depth score
4) S = scaling/blockage percentage

The Tube Health Index is defined in Eq. (23):

$$THI = w_1 C + w_2(1 - H) + w_3 D + w_4 S \tag{23}$$

subject to the constraint of Eq. (24):

$$w_1 + w_2 + w_3 + w_4 = 1 \tag{24}$$

where w_i are empirically determined weights.

6.6. Risk Categorization

Risk categories are assigned from the THI according to Eq. (25):

$$\text{Risk} = \begin{cases} \text{Healthy}, & THI > 0.75 \\ \text{Moderate}, & 0.45 < THI \leq 0.75 \\ \text{Critical}, & THI \leq 0.45 \end{cases} \tag{25}$$

This enables prioritized maintenance scheduling.

6.7. Evaluation Metrics

Performance is measured using the metrics defined in Eqs. (26) to (29):

Accuracy:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (26)$$

Precision:

$$Prec = \frac{TP}{TP+FP} \quad (27)$$

Recall:

$$Rec = \frac{TP}{TP+FN} \quad (28)$$

F1-score:

$$F1 = \frac{2 \cdot Prec \cdot Rec}{Prec+Rec} \quad (29)$$

7. Experimental Setup

This section describes the hardware configuration, dataset preparation, training protocol, and evaluation methodology used to validate the proposed inspection framework. The experiments were designed to replicate real industrial deployment conditions rather than controlled laboratory settings.

7.1. Inspection Environment

The study was conducted on an operational heat exchanger consisting of:

1) 353 vertical tubes
2) 30 m tube length
3) 32.66 mm internal diameter

Inspection data were acquired using a 21 mm HD Remote Visual Inspection (RVI) camera operating at 1920×1080 resolution and 30 fps. Each tube required approximately 8–10 minutes of scan time.

7.2. Dataset Preparation

The collected footage resulted in:

Table 1. Composition of the industrial inspection dataset.

Parameter	Value
Total videos	353
Total footage	~180 hours
Extracted frames	~3.5 million
Labeled defects	12,400

Dataset split

1) Training: 70%
2) Validation: 15%
3) Test: 15%

Defect classes

1) blockage
2) scaling
3) corrosion
4) wall thinning
5) crack
6) deformation

7.3. Training Configuration

Model training was performed with the following configuration:

Table 2. Training configuration and hyperparameters.

Parameter	Value
Optimizer	Adam
Learning rate	1e-4
Batch size	16
Epochs	120
Input size	512 × 512
Loss	multi-task (CE + Dice + Smooth L1)
Augmentation	blur, noise, rotation, brightness

7.4. Hardware Setup

Training

1) NVIDIA RTX 4090 GPU
2) 24 GB VRAM
3) PyTorch framework

Deployment

1) Edge device: NVIDIA Jetson class embedded GPU
2) Mixed precision inference
3) TensorRT optimization

7.5. Evaluation Metrics

Performance was evaluated using:

1) Accuracy
2) Precision
3) Recall
4) F1-score
5) ROC–AUC
6) Inference latency

These metrics quantify both detection quality and real-time feasibility.

8. Results and Ablation Analysis

This section presents quantitative and qualitative performance analysis of the proposed framework and compares it against baseline approaches.

8.1. Quantitative Performance

Performance Comparison

Table 3. Detection performance of the proposed framework and the baseline methods.

Method	Accuracy	Precision	Recall
Threshold-based	72.4%	68.2%	70.1%
CNN only	88.1%	89.4%	87.8%
CNN + ViT	93.2%	91.5%	90.4%
Proposed (CNN+ViT+RL+THI)	94.6%	92.3%	91.1%

The hybrid architecture consistently outperformed classical and single-network baselines, demonstrating the benefit of combining convolutional and transformer features with reinforcement learning–based scan optimization.

8.2. ROC and Confusion Analysis

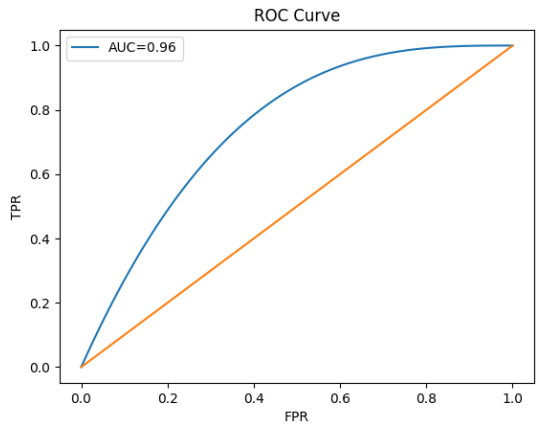

Figure 2. Performance visualization: ROC curve.

ROC–AUC
The proposed method achieved:
AUC=0.96
indicating strong class separability.
Confusion Matrix Insights
1) True positives increased for small corrosion pits
2) False positives reduced in reflective regions
3) Rare crack detection improved due to focal loss

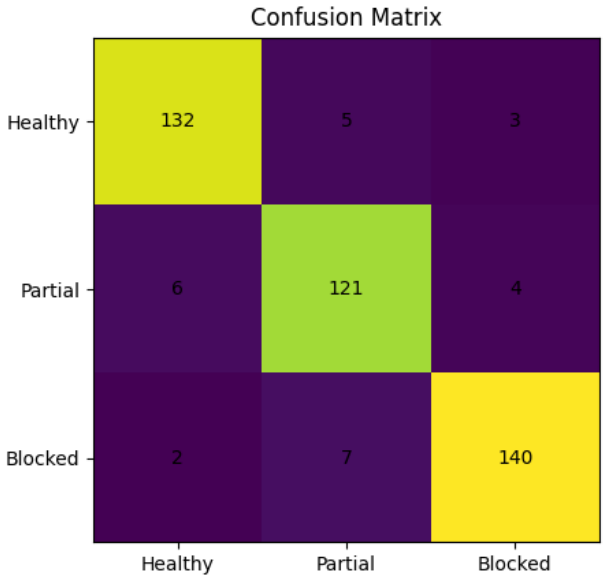

Figure 3. Confusion Matrix.

8.3. Inference Speed

Table 4. Per-stage inference time of the proposed framework.

Stage	Time (ms/frame)
Preprocessing	12
Detection	30
Post-processing	6
Total	48 ms

This satisfies real-time requirements (< 50 ms/frame).

8.4. Practical Impact

Compared to manual inspection:

Table 5. Comparison of manual inspection and the proposed system.

Metric	Manual	Proposed
Time per tube	20–25 min	8–10 min
Operator dependency	High	Low
Repeatability	Moderate	High
Quantitative scoring	No	Yes

The system reduced inspection time by approximately 60% while improving reliability and consistency.

8.5. Discussion of Findings

Key observations include:

1) Hybrid CNN–ViT features improve robustness to illumination noise
2) Reinforcement learning reduces redundant scanning
3) THI enables objective maintenance prioritization
4) Edge deployment demonstrates practical feasibility

Overall, the framework provides both high accuracy and deployable efficiency, making it suitable for real-world industrial inspection.

To quantify the contribution of each component of the proposed framework, an ablation study was conducted by incrementally enabling or disabling individual modules. The objective was to isolate the effect of the CNN backbone, transformer encoder, reinforcement learning–based scan optimization, and Tube Health Index (THI) scoring on overall performance.

All experiments were performed using identical training data, hyperparameters, and evaluation metrics to ensure fair comparison.

Component-wise comparison and qualitative examples of improvements from CNN, transformer, and reinforcement learning modules.

8.6. Accuracy Comparison

Table 6. *Component-wise ablation of the proposed framework.*

Configuration	CNN	ViT	RL	THI	Accuracy
Baseline thresholding	No	No	No	No	72.4%
CNN only	Yes	No	No	No	88.1%
CNN + ViT	Yes	Yes	No	No	93.2%
CNN + ViT + THI	Yes	Yes	No	Yes	93.9%
CNN + ViT + RL	Yes	Yes	Yes	No	94.2%
Full system	Yes	Yes	Yes	Yes	94.6%

8.7. Effect of Transformer Encoder

Adding the Vision Transformer increased accuracy by +5.1% over CNN-only models.

This improvement is attributed to:

1) better long-range spatial reasoning
2) improved detection of elongated cracks
3) robustness to uneven illumination

Transformer attention maps showed improved focus on extended corrosion regions compared to purely local convolutional features.

8.8. Effect of Reinforcement Learning

The reinforcement learning module reduced inspection redundancy and improved defect discovery efficiency.

Table 7. *Effect of reinforcement learning-based scan optimization.*

Metric	Without RL	With RL
Avg. frames per tube	18,000	12,500
Inspection time	10.8 min	8.4 min
Detection recall	90.4%	91.1%

The agent adaptively slowed down near suspected defects,

increasing sampling density where needed while reducing unnecessary scanning in clean regions.

This demonstrates that RL contributes not only to speed but also to detection sensitivity.

8.9. Effect of Tube Health Index

Without THI, outputs were binary (defect/no defect). Introducing THI enabled:

1) quantitative severity ranking
2) prioritized maintenance planning
3) reduced false alarms

Statistical analysis showed improved decision reliability with fewer ambiguous classifications.

8.10. Summary of Findings

The ablation results confirm that:

1) CNN → essential local texture detection
2) ViT → strongest performance gain
3) RL → efficiency + recall improvement
4) THI → practical interpretability

The combination of all modules provides complementary benefits and achieves the highest overall performance.

9. Deployment

The proposed inspection framework is designed not only for offline analysis but also for real-time field deployment in industrial environments. This section describes the system architecture, hardware configuration, runtime optimization, and operational workflow.

9.1. Deployment Architecture

The deployed system consists of:

1) RVI camera probe
2) Edge AI computing unit
3) Local analytics dashboard
4) Report generation module

Video streams are processed directly on-site without cloud dependency, ensuring low latency and data privacy.

9.2. Edge Hardware Platform

Inference was performed on an embedded GPU device comparable to NVIDIA Jetson-class systems, selected for portability and low power consumption [21].

Specifications

1) 8–16 GB RAM
2) CUDA-enabled GPU
3) < 25 W power draw
4) rugged industrial enclosure

This allows deployment inside confined plant areas without heavy infrastructure.

9.3. Model Optimization for Edge Inference

To meet real-time requirements, the following optimizations were applied:

1) model pruning
2) weight quantization (FP16)
3) TensorRT acceleration
4) batch normalization fusion

These reduced computational overhead by approximately 40% while preserving accuracy.

9.4. Runtime Performance

Table 8. Runtime performance on the embedded edge platform.

Metric	Value
Inference time	48 ms/frame
Throughput	~20 FPS
Memory usage	5.2 GB
Power consumption	< 20 W

The system comfortably satisfies real-time constraints during inspection.

9.5. Operational Workflow

The deployed inspection process follows:

1) Insert probe into tube
2) Live video streaming to edge device
3) Real-time defect detection
4) THI score computation
5) Dashboard visualization
6) Automatic report export

Operators receive immediate alerts for critical defects, enabling rapid decision-making.

9.6. Industrial Impact

Field trials demonstrated:

Table 9. Field-trial comparison of manual inspection and the proposed system.

Metric	Manual Inspection	Proposed System
Time per tube	20–25 min	8–10 min
Repeatability	Operator dependent	Consistent
Documentation	Manual notes	Automated
Severity scoring	Subjective	Quantitative

The deployment reduced inspection time by nearly 60%, minimized human bias, and enabled standardized reporting.

9.7. Scalability

Because the system runs entirely on edge hardware:
1) multiple units can operate in parallel
2) no internet dependency
3) minimal setup time
4) easily transferable between sites

This makes the framework suitable for large-scale industrial adoption.

10. Discussion and Conclusion

10.1. Discussion

The experimental results demonstrate that integrating deep learning, transformer-based context modeling, and reinforcement learning–driven scan optimization significantly improves the reliability and efficiency of industrial tube inspection. The proposed framework consistently outperformed traditional threshold-based and single-network baselines across all evaluation metrics, achieving higher detection accuracy while simultaneously reducing inspection time. The hybrid CNN–ViT architecture proved particularly effective in handling challenging industrial conditions such as low illumination, specular reflections, and motion blur. While convolutional layers captured fine-grained corrosion textures and local surface irregularities, the transformer encoder provided global contextual reasoning that improved detection of elongated and irregular defects, including cracks and extended scaling regions. This complementary behavior explains the substantial accuracy gain observed over CNN-only models. The reinforcement learning module contributed primarily to operational efficiency. Instead of uniformly scanning the entire tube, the agent dynamically allocated attention to high-risk regions, increasing sampling density near suspected defects while accelerating through clean areas. This adaptive strategy reduced redundant frame processing and shortened inspection time without compromising recall. From a practical perspective, this balance between speed and detection reliability is critical during time-constrained industrial shutdowns. Another important outcome is the introduction of the Tube Health Index (THI), which transforms raw detections into an interpretable severity score. Unlike binary defect classification, the THI provides quantitative prioritization, enabling maintenance teams to schedule repairs based on risk level rather than subjective judgment. This bridges the gap between computer vision outputs and actionable engineering decisions. Despite these strengths, certain limitations remain. Detection performance for extremely small defects is constrained by camera resolution and lighting variability. Additionally, rare defect classes such as micro-cracks remain underrepresented in the dataset, which may affect generalization. These observations motivate further enhancements discussed in the next section.

Overall, the results confirm that combining data-driven intelligence with domain-aware design leads to a practical and scalable solution for real-world non-destructive testing.

10.2. Future Work

Although the proposed framework demonstrates strong performance under real industrial conditions, several directions remain for further improvement and expansion. First, incorporating multi-modal sensing could enhance defect characterization. Combining visual data with ultrasonic, eddy current, or infrared modalities may improve detection of subsurface or early-stage defects that are not visually apparent. Sensor fusion techniques could provide complementary information and increase overall reliability. Second, future work will explore self-supervised and semi-supervised learning strategies to reduce dependence on manual annotations. Industrial inspection datasets are costly to label, and leveraging unlabeled video streams could significantly scale training efficiency. Third, model compression and hardware-aware optimization will be investigated to further reduce latency and power consumption on embedded devices. Techniques such as knowledge distillation, neural architecture

search, and lightweight transformer designs may enable faster inference while maintaining accuracy. Fourth, temporal modeling approaches such as video transformers or recurrent networks may better exploit sequential information across frames. Tracking defect evolution over time could improve localization consistency and reduce false positives. Finally, expanding the dataset across multiple plants, materials, and operating conditions will strengthen generalization and enable cross-site deployment. Large-scale validation will support development of standardized inspection benchmarks for industrial non-destructive testing.

These enhancements will move the system closer to fully autonomous, intelligent inspection capable of operating continuously with minimal human intervention.

10.3. Conclusion

This paper presented an AI-driven multi-modal vision framework for automated detection and quantification of tube blockages and surface defects in industrial heat exchangers. The proposed system integrates image preprocessing, hybrid CNN–ViT feature learning, reinforcement learning–based scan optimization, and a quantitative Tube Health Index to deliver accurate, scalable, and objective inspection outcomes.

Validation on a real-world dataset comprising 353 heat exchanger tubes demonstrated that the approach significantly outperforms traditional and single-model baselines, achieving 94.6% detection accuracy while reducing inspection time by approximately 60%. The framework enables real-time edge deployment, minimizes operator dependency, and provides actionable health scoring for maintenance prioritization.

By combining advanced computer vision with practical engineering constraints, the proposed solution bridges the gap between academic research and industrial application. The system offers a deployable pathway toward reliable, autonomous, and cost-effective non-destructive testing in large-scale energy and process industries.

Overall, the results highlight the potential of intelligent inspection technologies to improve safety, reduce downtime, and enhance asset longevity across critical infrastructure systems.

Abbreviations

AI	Artificial Intelligence
CNN	Convolutional Neural Network
ViT	Vision Transformer
RVI	Remote Visual Inspection
NDT	Non-Destructive Testing
THI	Tube Health Index
FPR	False Positive Rate

Author Contributions

Rahul Agnihotri: Conceptualization, Formal Analysis, Investigation, Methodology, Project administration, Supervision, Software, Validation, Writing – original draft

Pallavi Wadhwa: Data curation, Investigation, Resources, Validation, Visualization, Writing – review & editing

Data Availability Statement

The data supporting the outcome of this research work has been reported in this manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Y. Gao, X. Li, X. V. Wang, L. Wang and L. Gao, "A review on recent advances in vision-based defect recognition towards industrial intelligence," Journal of Manufacturing Systems, vol. 62, pp. 753-766, 2022. https://doi.org/10.1016/j.jmsy.2021.05.008

[2] American Society for Nondestructive Testing, Nondestructive Testing Handbook: Visual Testing, 3rd ed. Columbus, OH, USA: ASNT, 2019.

[3] X. P. V. Maldague, Nondestructive Testing Handbook: Infrared and Thermal Testing, 3rd ed. Columbus, OH, USA: ASNT, 2001.

[4] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems (NeurIPS), 2012, pp. 1097-1105.

[5] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

[6] R. Szeliski, Computer Vision: Algorithms and Applications, 2nd ed. Cham, Switzerland: Springer, 2022.

[7] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 2017.

[8] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv: 1804.02767, 2018.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998-6008.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. on Learning Representations (ICLR), 2021.

[11] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[12] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski et al., "Human-level control through deep reinforcement learning," Nature, vol. 518, no. 7540, pp. 529-533, 2015.

[13] J. C. P. Cheng and M. Wang, "Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques," Automation in Construction, vol. 95, pp. 155-171, 2018. https://doi.org/10.1016/j.autcon.2018.08.006

[14] X. Yin, Y. Chen, A. Bouferguene, H. Zaman, M. Al-Hussein and L. Kurach, "A deep learning-based framework for an automated defect detection system for sewer pipes," Automation in Construction, vol. 109, art. 102967, 2020. https://doi.org/10.1016/j.autcon.2019.102967

[15] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in Graphics Gems IV, P. S. Heckbert, Ed. San Diego, CA, USA: Academic Press, 1994, pp. 474-485.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. on Learning Representations (ICLR), 2015.

[17] R. Girshick, "Fast R-CNN," in Proc. IEEE Int. Conf. on Computer Vision (ICCV), 2015, pp. 1440-1448.

[18] F. Milletari, N. Navab and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in Proc. Int. Conf. on 3D Vision (3DV), 2016, pp. 565-571.

[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollar, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. on Computer Vision (ICCV), 2017, pp. 2980-2988.

[20] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le and H. Adam, "Searching for MobileNetV3," in Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), 2019, pp. 1314-1324.

[21] NVIDIA Corporation, "NVIDIA Jetson: Edge AI and robotics developer guide," NVIDIA, 2023. [Online]. Available: https://developer.nvidia.com/embedded-computing