GJR Publication
OPEN ACCESS

Global Journal of Research in Engineering & Computer Sciences
ISSN: 2583-2727 (Online)
Volume 06 | Issue 04 | July-August | 2026
Journal homepage: https://gjrpublication.com/gjrecs/

Research Article

Cross-Modal Feature Alignment for Robust Defect Detection in Industrial Inspection: A Comprehensive Review of RGB-Thermal and Multimodal Fusion Techniques

*[1]Hauwa Aminu Kofa [2]Bello Abubakar Imam [3]Shehu Hassan Ayagi and [4]Sadiq Bello Abubakar

[1,2,3,4] Department of Computer Science Kano State Polytechnic, Kano State of Nigeria.

DOI: 10.5281/zenodo.21283386 **Submission Date: 25 May 2026 | Published Date: 09 July 2026**

***Corresponding author:** **Hauwa Aminu Kofa**
Department of Computer Science Kano State Polytechnic, Kano State of Nigeria.

Abstract

Industrial quality inspection has entered a transformative era where single-modality vision systems – predominantly RGB cameras – are proving insufficient for detecting subtle defects involving geometric anomalies, material inconsistencies, and thermal irregularities. The fusion of multimodal sensor data, particularly RGB paired with thermal or depth modalities, offers a pathway to comprehensive defect detection but introduces significant technical challenges: sensor misalignment, computational overhead, and semantic integration across heterogeneous data types. This review provides a systematic examination of cross-modal feature alignment techniques for robust defect detection in industrial inspection. We analyze the fundamental challenges of multimodal fusion, present a comprehensive taxonomy categorizing fusion approaches into data-level, feature-level, and decision-level methods, and critically evaluate state-of-the-art architectures including transformer-based cross-modal alignment, state-space models exemplified by the MambaAlign framework, and reconstruction-based approaches. Quantitative comparisons demonstrate that modern alignment-aware methods achieve 4.8–6.5% improvements in detection accuracy while sustaining real-time performance (~30 FPS) suitable for production lines. We synthesize recent advances from MambaAlign (Tan et al., 2026), cross-modal feature mapping (Nam, 2025), and lightweight optimization strategies, identifying that optimal inspection systems increasingly shift toward alignment-aware architectures with linear computational complexity. Finally, we outline open challenges – dataset scarcity, domain adaptation, explainability – and propose future research directions including self-supervised alignment, physics-informed fusion, and edge-cloud collaborative deployment.

Keywords: *Multimodal fusion, industrial defect detection, RGB-thermal imaging, cross-modal alignment, sensor fusion, anomaly detection, real-time inspection.*

1. Introduction

Industrial quality inspection constitutes the final and often most critical barrier against defective products reaching consumers, with implications spanning automotive safety, aerospace reliability, electronics functionality, and energy system integrity (Zhang et al., 2025). Traditional vision-based inspection systems, predominantly relying on Red-Green-Blue (RGB) cameras, have become ubiquitous due to their speed, affordability, and well-understood computer vision algorithms. These systems excel at detecting surface-level color anomalies and macroscopic defects under controlled lighting conditions (Tan et al., 2026).

However, the limitations of RGB-only inspection are increasingly apparent in modern manufacturing environments. As noted by recent research, RGB cameras "often miss defects related to geometry (scratches or dents), material structure, or heat dissipation" (Tan et al., 2026). A scratch on a metallic surface, a subsurface delamination in composite materials, or an overheating electronic component may present no visible color change detectable by conventional cameras, yet these defects critically compromise product functionality and safety.

@ 2026 | PUBLISHED BY GJR PUBLICATION, INDIA

The integration of complementary sensing modalities offers a compelling solution. Thermal infrared cameras capture temperature distributions and heat dissipation patterns, revealing anomalies invisible to RGB sensors (Brenner et al., 2023). Depth sensors provide three-dimensional geometric information, detecting dents, warpage, or misalignments that appear as subtle shading variations in RGB images (Brenner et al., 2023). Each modality brings distinct advantages: RGB provides high-resolution texture and color information; thermal operates independently of illumination and reveals thermal anomalies; depth captures geometric structure regardless of surface appearance (Brenner et al., 2023; Nam, 2025).

Yet the promise of multimodal fusion confronts a fundamental challenge: how to effectively combine information from heterogeneous sensors into a unified representation that preserves the strengths of each modality while compensating for their individual limitations? This challenge manifests in several interconnected problems:

Sensor Misalignment: Thermal and depth cameras cannot occupy the identical optical path as RGB cameras, resulting in parallax, scale differences, and spatial offsets. Even after calibration, modest misregistration remains common in factory settings, causing fused features to reference incorrect spatial locations (Tan et al., 2026; National Technical University of Athens, 2025).

Semantic Heterogeneity: RGB intensities, thermal radiation values, and depth measurements represent fundamentally different physical phenomena. Direct concatenation or naive fusion fails to capture the semantic relationships between modalities (Zhang et al., 2025).

Computational Constraints: Industrial inspection demands real-time performance (typically >20–30 frames per second) on resource-constrained edge hardware. Complex fusion architectures must balance accuracy against latency and power consumption (Zhang et al., 2025; Tan et al., 2026).

Information Preservation: Deep feature extractors progressively downsample spatial resolution, causing fine details essential for defect localization – particularly for small or thin defects – to vanish (National Technical University of Athens, 2025).

This review addresses the intersection of these challenges, focusing specifically on **cross-modal feature alignment** as the critical enabling technology for robust industrial defect detection. Unlike previous surveys that broadly cover sensor fusion in autonomous driving or general computer vision (Brenner et al., 2023; Nam, 2025), we specifically target industrial inspection applications where alignment accuracy, real-time performance, and robustness to modest misregistration are paramount.

1.1 Scope and Contributions

This review provides a systematic examination of cross-modal feature alignment techniques for RGB-thermal and multimodal industrial defect detection. Our key contributions include:

- A comprehensive taxonomy categorizing multimodal fusion methods into data-level, feature-level, and decision-level approaches, with quantitative comparison of their accuracy-efficiency trade-offs on industrial inspection benchmarks.
- Systematic analysis of alignment-aware architectures, including transformer-based cross-attention mechanisms, state-space models (exemplified by MambaAlign), and reconstruction-based feature mapping strategies.
- Critical evaluation of lightweight optimization techniques – quantization, pruning, knowledge distillation – that enable real-time deployment on edge hardware.
- Synthesis of emerging trends including self-supervised alignment, physics-informed fusion, and few-shot defect detection.
- Identification of open challenges and future research directions for next-generation intelligent inspection systems.

The remainder of this paper is organized as follows. Section 2 analyzes the fundamental challenges in multimodal industrial inspection. Section 3 presents a systematic taxonomy of fusion methodologies. Section 4 reviews alignment-aware architectures. Section 5 discusses datasets and evaluation benchmarks. Section 6 examines optimization strategies for edge deployment. Section 7 explores emerging trends, and Section 8 concludes with future research directions.

2. Challenges in Multimodal Industrial Defect Detection

2.1 The Limitations of Single-Modality Inspection

RGB cameras, despite their prevalence, exhibit inherent constraints stemming from their operation within the visible spectrum (Brenner et al., 2023). As illustrated in Figure 1 of Brenner et al., the visible imaging range is notably narrower compared to other spectra, limiting effectiveness to well-lit conditions and surface-level color/texture anomalies (Brenner et al., 2023).

@ 2026 | PUBLISHED BY GJR PUBLICATION, INDIA

Quantitative evidence from industrial applications underscores these limitations. In gear defect detection, traditional vision methods achieve accuracy below 90% under ideal conditions, with performance dropping to 70% or lower in environments with strong noise or variable illumination (Zhang et al., 2025). Crucially, defects with subtle thermal signatures – overheating components, friction-induced hot spots, or thermal dissipation anomalies – remain entirely undetectable to RGB sensors regardless of lighting quality.

Depth sensing addresses geometric defects but introduces its own limitations. Time-of-Flight sensors face interference from sunlight in outdoor applications, while stereo vision struggles with low-texture surfaces (Brenner et al., 2023). Resolution decreases with distance, and fine geometric details may fall below detection thresholds.

2.2 Sensor Misalignment and Registration

The physical separation of sensors on an inspection rig creates inevitable parallax and coordinate system differences. Accurate fusion requires determining intrinsic parameters (focal length, principal point, distortion coefficients) and extrinsic parameters (rotation, translation) for each sensor, then projecting all modalities into a common coordinate frame (Brenner et al., 2023).

This calibration process, while well-understood for static camera pairs, becomes challenging in industrial environments with vibration, thermal expansion, or when sensors are mounted on robotic manipulators. Even after calibration, "modest sensor misregistration" remains common in factory settings (Tan et al., 2026; National Technical University of Athens, 2025).

Feature-based registration methods utilizing SIFT, SURF, or similar descriptors can achieve alignment but face two significant hurdles for industrial deployment. First, their computational complexity often precludes real-time operation with moving cameras (Brenner et al., 2023). Second, thermal and depth imagery exhibit different feature characteristics than RGB images, complicating cross-modal matching (Brenner et al., 2023; Nam, 2025).

2.3 Semantic Heterogeneity and Feature Fusion

RGB intensities, thermal radiation values, and depth measurements reside in different physical units and statistical distributions. Direct concatenation at the input level – perhaps the simplest fusion approach – forces networks to learn cross-modal relationships from raw data, requiring substantial training data and capacity (Zhang et al., 2025).

Feature-level fusion, where modality-specific backbones extract representations before combination, preserves modality-specific information but introduces the alignment problem at feature space: how to ensure that features from different modalities correspond to the same spatial locations and semantic concepts (National Technical University of Athens, 2025).

2.4 Computational Constraints and Real-Time Requirements

Industrial inspection lines demand throughput matching production rates – typically 20–30 frames per second or higher (Tan et al., 2026). Edge deployment platforms (e.g., NVIDIA Jetson, embedded GPUs, or custom accelerators) offer limited computational resources compared to server-grade hardware (Zhang et al., 2025).

As noted in recent reviews, "the imbalance between real-time detection requirements and computational resources" presents a persistent challenge (Zhang et al., 2025). Fusion architectures must carefully balance representational capacity against inference latency, often requiring aggressive optimization including quantization, pruning, and hardware-specific acceleration.

2.5 Data Scarcity and Class Imbalance

Industrial defect detection confronts the fundamental challenge of limited anomalous samples. Defects are, by definition, rare events in well-controlled manufacturing processes. Obtaining sufficient defective examples for training supervised models proves difficult, particularly for novel defect types (Zhang et al., 2025; National Technical University of Athens, 2025).

This scarcity compounds with the "curse of dimensionality" introduced by multimodal data – more modalities mean more parameters and greater data requirements for effective training. Class imbalance further complicates learning, as networks may converge to trivial solutions that always predict "normal" (Zhang et al., 2025).

3. A Taxonomy of Multimodal Fusion Methodologies

Multimodal fusion methods in industrial inspection can be categorized along two primary dimensions: the level at which fusion occurs, and the mechanism by which cross-modal relationships are modeled.

@ 2026 | PUBLISHED BY GJR PUBLICATION, INDIA

3.1 Fusion Levels

Data-Level Fusion (Early Fusion): Raw sensor data are combined prior to feature extraction, typically through concatenation, weighted summation, or more sophisticated alignment operations. Data-level fusion preserves all original information and enables unified feature extraction, but requires precise spatial alignment and struggles with heterogeneous data types (Brenner et al., 2023; National Technical University of Athens, 2025).

Early fusion proves most effective when modalities are naturally registered (e.g., RGB and aligned depth from stereo matching) or when computational resources are extremely constrained. However, misalignment at the input level propagates through all subsequent processing, making early fusion brittle to calibration errors (Tan et al., 2026).

Feature-Level Fusion (Intermediate Fusion): Modality-specific backbones extract representations independently, followed by fusion through concatenation, attention mechanisms, or more complex interaction modules. Feature-level fusion dominates recent literature due to several advantages (National Technical University of Athens, 2025):

- Modality-specific architectures can be optimized for each sensor type
- Alignment occurs in learned feature space, offering robustness to modest misregistration
- Intermediate representations preserve spatial structure while abstracting away low-level differences

The primary challenge lies in designing fusion mechanisms that effectively capture cross-modal interactions without excessive computational overhead (Zhang et al., 2025).

Decision-Level Fusion (Late Fusion): Independent detectors process each modality and produce detection outputs (class probabilities, anomaly scores, segmentation maps), which are then combined through voting, averaging, or learned weighting. Decision-level fusion offers maximal flexibility – modalities can use completely different architectures and even operate at different frame rates (Brenner et al., 2023).

However, late fusion sacrifices the opportunity for cross-modal feature interactions that could resolve ambiguities present in individual modalities. An object with ambiguous RGB appearance but clear thermal signature might be missed if RGB detection fails before fusion (Nam, 2025).

Table 1 compares fusion levels on key criteria:

Criterion	Data-Level	Feature-Level	Decision-Level
Spatial alignment requirement	Critical	Moderate	Low
Cross-modal interaction	Learned implicitly	Explicit modeling	Minimal
Computational cost	Lowest	Moderate	Highest (multiple detectors)
Robustness to misalignment	Lowest	Highest	Moderate
Modality-specific optimization	Limited	Full	Full
Representative examples	Input concatenation	Cross-attention, MambaAlign	Ensemble voting

3.2 Fusion Mechanisms

Concatenation-Based Fusion: The simplest mechanism, concatenating either raw data or feature maps along the channel dimension, followed by convolutional processing. While computationally efficient, concatenation assumes that networks can learn cross-modal relationships from stacked representations – an assumption that may require substantial capacity and data (Zhang et al., 2025).

Attention-Based Fusion: Attention mechanisms enable dynamic weighting of features based on cross-modal relevance. Transformer architectures with multi-head self-attention have demonstrated particular effectiveness for modeling relationships between modality-specific feature sequences (Zhang et al., 2025).

The cross-attention mechanism computes attention from one modality to another, enabling queries from modality A to attend to keys/values from modality B. This provides a natural mechanism for cross-modal feature alignment – features from different modalities that correspond to the same spatial location or semantic concept receive higher attention weights (Zhang et al., 2025).

State-Space Fusion: Recent advances in state-space models, exemplified by the Mamba architecture, offer linear-complexity sequence modeling as an alternative to quadratic-complexity transformer attention. The MambaAlign framework applies state-space refinement to capture "long-range and orientation-aware context" essential for detecting thin or oblique defects such as scratches and cracks (Tan et al., 2026; National Technical University of Athens, 2025).

State-space fusion maintains computational cost close to linear in sequence length while effectively modeling long-range dependencies – critical for defects that span significant portions of the inspection area.

@ 2026 | PUBLISHED BY GJR PUBLICATION, INDIA

Reconstruction-Based Fusion: Reconstruction approaches learn to predict features of one modality from another, or to reconstruct normal patterns from multimodal inputs. The Crossmodal Feature Mapping (CFM) strategy predicts 3D features from 2D features and vice versa, enabling anomaly detection based on reconstruction error (Nam, 2025).

Reconstruction-based fusion proves particularly valuable in few-shot scenarios, where limited defective samples prevent direct supervised learning of anomaly detection (Zhang et al., 2025; Nam, 2025).

4. Alignment-Aware Architectures for Multimodal Fusion

4.1 The Alignment Problem in Multimodal Fusion

The fundamental challenge underlying all fusion approaches is alignment: ensuring that information from different modalities corresponding to the same physical location or object is properly associated. Misalignment – whether due to calibration errors, parallax, or differences in resolution – causes features to reference incorrect spatial positions, degrading detection accuracy and potentially creating false anomalies (Tan et al., 2026).

Traditional approaches address alignment through explicit geometric calibration and registration. Intrinsic and extrinsic parameters are estimated for each sensor, enabling projection into a common coordinate frame (Brenner et al., 2023; Nam, 2025). However, this approach faces limitations in dynamic industrial environments where vibration, thermal expansion, or sensor movement may alter relative positions.

Recent research has shifted toward learned alignment in feature space, where networks develop the capacity to associate corresponding features across modalities even with modest spatial offsets (Tan et al., 2026; National Technical University of Athens, 2025).

4.2 MambaAlign: Alignment-Aware State-Space Fusion

The MambaAlign framework, introduced by Tan et al. (2026), represents a state-of-the-art approach to alignment-aware multimodal fusion for industrial anomaly detection. Developed specifically to address the limitations of existing fusion methods – loss of fine spatial detail, computational overhead, and brittleness to misregistration – MambaAlign introduces several key innovations.

Alignment-Aware Design: Rather than assuming perfect registration, MambaAlign explicitly accommodates modest misalignment through a top-down reconstruction mechanism. Low-level feature channels are reconstituted after high-level semantic exchange, allowing the network to recover precise spatial localization even when initial feature maps are slightly misaligned (Tan et al., 2026; National Technical University of Athens, 2025).

State-Space Refinement: The framework captures long-range, orientation-aware context using state-space recurrences, which prove particularly effective for detecting thin or oblique defects such as scratches and cracks. Unlike transformer attention with quadratic complexity, state-space models maintain linear computational cost while effectively modeling long-range dependencies (National Technical University of Athens, 2025).

Cross Mamba Interaction: Semantic guidance between modalities is exchanged only at high-level feature stages through lightweight cross-recurrence, avoiding the computational burden of dense cross-attention at all scales. This selective interaction preserves computational efficiency while enabling cross-modal information flow (Tan et al., 2026).

Quantitative Performance: Extensive experiments across three RGB-plus-auxiliary-modality (RGB-X) datasets demonstrate substantial improvements:

- Image-level AUROC improved by approximately **4.8%** over prior methods
- Pixel-level AUROC improved by approximately **5.0%**
- Area under per-region overlap curve improved by roughly **6.5%** (Tan et al., 2026; National Technical University of Athens, 2025)

Crucially, these gains come without excessive computational overhead. The model sustains **close to 30 frames per second** at moderate resolutions with controlled memory usage, making it practical for deployment in real production lines (Tan et al., 2026; National Technical University of Athens, 2025).

4.3 Crossmodal Feature Mapping for 3D Anomaly Detection

The Crossmodal Feature Mapping (CFM) approach, presented at CVPR 2024 and reviewed by Nam (2025), addresses multimodal fusion from a reconstruction perspective. Designed for 3D anomaly detection combining RGB and point cloud data, CFM introduces Deep Feature Reconstruction (DFR) where features of one modality are predicted from the other.

Bidirectional Prediction: The framework simultaneously predicts 3D features from 2D RGB features and 2D features from 3D point cloud features. This bidirectional mapping ensures that both modalities contribute to anomaly detection –

@ 2026 | PUBLISHED BY GJR PUBLICATION, INDIA

an anomaly visible only in one modality creates prediction errors that propagate through the reconstruction process (Nam, 2025).

AND Aggregation: Anomaly scores from both directions are combined through AND-style aggregation, requiring consensus across modalities to declare an anomaly. This strategy proves particularly effective for detecting subtle defects that may be ambiguous in individual modalities but consistent cross-modally (Nam, 2025).

Layers Pruning: To accelerate inference, CFM employs layers pruning – using only shallow layer features during extraction. This reduces computational overhead while maintaining detection performance, addressing the real-time requirements of industrial deployment (Nam, 2025).

4.4 Transformer-Based Cross-Modal Alignment

Transformer architectures have demonstrated particular effectiveness for cross-modal feature alignment due to their ability to model long-range dependencies and learn flexible attention patterns (Zhang et al., 2025).

In gear defect detection applications, transformer-based cross-modal feature alignment mechanisms establish "global correlation mapping" between visual and vibration signals, enabling joint embedding that captures relationships across modalities (Zhang et al., 2025). The attention mechanism learns which regions of the visual input correspond to which vibration patterns, effectively aligning modalities without explicit geometric calibration.

This approach proves valuable when modalities operate at different spatial or temporal resolutions – vibration sensors provide time-series data that must be aligned with spatial regions in visual imagery. Attention mechanisms learn to associate temporal patterns with spatial locations, enabling comprehensive defect characterization (Zhang et al., 2025).

4.5 Comparative Analysis of Alignment-Aware Architectures

Table 2 compares representative alignment-aware fusion architectures:

Architecture	Alignment Mechanism	Fusion Level	Key Innovation	Performance Gain	FPS	Source
MambaAlign	State-space refinement + top-down reconstruction	Feature	Alignment-aware design, linear complexity	+4.8–6.5% AUROC	~30	Tan et al. (2026); National Technical University of Athens (2025)
Crossmodal Feature Mapping	Bidirectional feature prediction	Feature	Reconstruction-based, AND aggregation	State-of-the-art on MVTec3D	Not reported	Nam (2025)
Transformer Cross-Attention	Multi-head attention	Feature	Global correlation mapping	Cross-modal joint embedding	Varies	Zhang et al. (2025)
Feature Concatenation (Baseline)	None (assumes alignment)	Feature	–	Baseline	Highest	Zhang et al. (2025)

5. Datasets and Evaluation Benchmarks

5.1 Industrial Inspection Datasets

The development and evaluation of multimodal fusion methods require appropriately designed datasets capturing both normal and defective samples across multiple modalities. Several datasets have emerged as community benchmarks.

MVTec AD and MVTec 3D-AD: The MVTec Anomaly Detection dataset and its 3D extension have become de facto standards for industrial anomaly detection research. MVTec AD provides RGB images of 15 object categories with pixel-level anomaly annotations. MVTec 3D-AD adds high-resolution 3D sensor data, enabling evaluation of RGB-D fusion methods (Nam, 2025).

RGB-X Datasets: The MambaAlign evaluation utilizes three RGB-plus-auxiliary-modality datasets spanning different industrial scenarios, though specific dataset names are not disclosed in available materials (Tan et al., 2026; National Technical University of Athens, 2025). These datasets include paired RGB and thermal/depth imagery with various defect types.

Custom Industrial Datasets: Many research efforts rely on proprietary datasets collected from specific production lines. The gear defect detection review notes that dataset scarcity remains a significant barrier, with limited publicly available multimodal industrial inspection datasets (Zhang et al., 2025; National Technical University of Athens, 2025).

@ 2026 | PUBLISHED BY GJR PUBLICATION, INDIA

5.2 Evaluation Metrics

Detection Metrics: Image-level Area Under the Receiver Operating Characteristic curve (AUROC) measures the model's ability to distinguish normal from anomalous images regardless of threshold choice. Pixel-level AUROC evaluates localization accuracy – whether the model identifies the precise spatial extent of defects (Tan et al., 2026).

Per-Region Overlap (PRO): The PRO curve evaluates how well predicted anomaly regions overlap with ground truth annotations, providing a more stringent localization metric than pixel AUROC alone. MambaAlign achieves approximately 6.5% improvement in PRO compared to prior methods (Tan et al., 2026).

Computational Metrics: Frames per second (FPS), parameters (millions), FLOPs (billions), and memory usage (MB) quantify deployment feasibility. For industrial inspection, achieving target frame rates (typically >20–30 FPS) on target hardware is often as important as detection accuracy (Zhang et al., 2025; Tan et al., 2026).

5.3 Dataset Limitations and Future Needs

The limited availability of high-quality multimodal industrial datasets constrains research progress (Brenner et al., 2023; National Technical University of Athens, 2025). Key gaps include:

- **Diverse defect types:** Many datasets focus on specific defect categories, limiting evaluation of generalization to novel anomaly types.
- **Realistic misalignment:** Controlled datasets with perfect registration may not reflect factory conditions where modest misalignment is common.
- **Temporal information:** Video datasets enabling motion-based detection remain scarce.
- **Multi-modal variety:** RGB-thermal-depth triple-modality datasets are particularly rare (Brenner et al., 2023; Nam, 2025).

6. Optimization Strategies for Edge Deployment

6.1 Lightweight Architectures

Industrial deployment on edge hardware demands computationally efficient models. Lightweight architectures reduce parameters and FLOPs while maintaining detection accuracy.

MobileNet and Inverted Residuals: Depthwise separable convolutions, introduced in MobileNet, factorize standard convolution into depthwise (spatial filtering per channel) and pointwise (channel combination) operations, reducing computation by 8–9× (Zhang et al., 2025). Inverted residual blocks with linear bottlenecks further improve efficiency while preserving representational capacity.

ShuffleNet and Channel Shuffle: ShuffleNet introduces channel shuffle operations to reduce inter-channel dependence, enabling efficient group convolutions without sacrificing cross-channel information flow. This design achieves favorable accuracy-efficiency trade-offs for resource-constrained deployment (Zhang et al., 2025).

Quantitative Comparisons: Recent reviews demonstrate that lightweight models such as MobileNet and ShuffleNet reduce parameters by 40–60% while maintaining mean Average Precision essentially unchanged compared to standard architectures (Zhang et al., 2025).

6.2 Quantization

Quantization reduces numerical precision from 32-bit floating point (FP32) to lower bit-widths (FP16, INT8), decreasing memory bandwidth and enabling faster arithmetic on hardware with native low-precision support.

FP16 Quantization: Modern edge GPUs include Tensor Cores optimized for FP16 computation, typically achieving 2–4× speedup with minimal (often <1%) accuracy degradation for well-trained models (Zhang et al., 2025).

INT8 Quantization: Further reducing precision to 8-bit integers enables deployment on CPUs and specialized edge AI accelerators. Accuracy degradation can reach 2–5% for detection tasks, particularly affecting small defect detection, but careful calibration minimizes this impact (Zhang et al., 2025).

6.3 Pruning and Knowledge Distillation

Network Pruning: Structured pruning removes entire channels or layers based on importance criteria (e.g., batch normalization scaling factors). Pruning 30–50% of channels can reduce computation substantially with limited accuracy loss, particularly when followed by fine-tuning (Zhang et al., 2025).

Knowledge Distillation: A compact "student" network learns to mimic the outputs of a larger "teacher" model, achieving efficiency approaching the student with accuracy approaching the teacher. Distillation proves particularly valuable for deploying complex fusion architectures on edge hardware (Zhang et al., 2025).

@ 2026 | PUBLISHED BY GJR PUBLICATION, INDIA

7. Emerging Trends and Future Directions

7.1 Self-Supervised Alignment

Manual annotation of cross-modal correspondences is expensive and often impractical. Self-supervised approaches learn alignment from unlabeled multimodal data by exploiting natural constraints – temporal consistency in video, spatial continuity, or cross-modal prediction tasks.

Contrastive learning frameworks that maximize mutual information between aligned modalities while minimizing it for non-aligned pairs offer a promising direction for learning alignment without explicit supervision (Zhang et al., 2025).

7.2 Physics-Informed Fusion

Incorporating physical principles governing the relationship between modalities offers a path to more robust and data-efficient fusion. Thermal diffusion models, for example, predict how heat should propagate through materials – deviations between predicted and observed thermal patterns indicate anomalies (National Technical University of Athens, 2025).

Physics-informed fusion combines data-driven learning with domain knowledge, potentially reducing data requirements while improving generalization to novel defect types.

7.3 Few-Shot and Zero-Shot Defect Detection

The scarcity of defective samples in well-controlled manufacturing processes motivates few-shot and zero-shot approaches. Reconstruction-based methods that detect anomalies as deviations from learned normal patterns require only normal samples for training, bypassing the need for defective examples (Zhang et al., 2025; Nam, 2025).

Recent advances demonstrate that few-shot generation models based on contrastive learning achieve accuracy in 10-shot scenarios reaching 90% of fully supervised models (Zhang et al., 2025). Extending these approaches to multimodal settings remains an active research direction.

7.4 Edge-Cloud Collaborative Deployment

Hybrid deployment strategies leverage edge devices for real-time low-latency inference while offloading complex analysis to cloud servers. A lightweight detector on the edge identifies regions of interest or potential anomalies; high-resolution multimodal analysis of these regions occurs in the cloud (Zhang et al., 2025).

Dynamic scheduling algorithms allocate computational resources based on current workload, defect likelihood, and available bandwidth, optimizing the accuracy-efficiency trade-off across the inspection system.

7.5 Explainable Multimodal Fusion

Industrial adoption of AI-based inspection requires trust – operators must understand why a system flagged a particular region as anomalous. Explainable AI techniques for multimodal fusion aim to attribute detection decisions to specific modalities and spatial regions (Zhang et al., 2025).

Attention weights from transformer-based fusion provide one form of explanation, indicating which modalities and which spatial locations contributed to the detection. Future work should develop standardized explainability metrics and interfaces for industrial inspection.

8. Conclusion

This review has systematically examined cross-modal feature alignment techniques for robust defect detection in multimodal industrial inspection. The central challenge – aligning information from heterogeneous sensors while maintaining computational efficiency for real-time edge deployment – has driven substantial innovation in fusion architectures, alignment mechanisms, and optimization strategies.

Key findings include:

1. **Alignment-aware fusion** represents a paradigm shift from assuming perfect registration to explicitly accommodating misalignment in feature space. The MambaAlign framework demonstrates that state-space refinement with top-down reconstruction achieves substantial accuracy improvements (4.8–6.5% AUROC gains) while sustaining real-time performance (~30 FPS) (Tan et al., 2026; National Technical University of Athens, 2025).
2. **Fusion level selection** involves fundamental trade-offs between alignment requirements, computational cost, and cross-modal interaction. Feature-level fusion with learned alignment offers the best balance for industrial applications, enabling modality-specific optimization while capturing cross-modal relationships (Zhang et al., 2025; National Technical University of Athens, 2025).

@ 2026 | PUBLISHED BY GJR PUBLICATION, INDIA

3. **Reconstruction-based approaches** exemplified by Crossmodal Feature Mapping prove particularly valuable for few-shot scenarios, detecting anomalies as deviations from learned normal patterns without requiring extensive defective samples (Nam, 2025).
4. **Hardware-aware optimization** – lightweight architectures, quantization, pruning – is essential for achieving real-time performance on edge platforms, with parameter reductions of 40–60% achievable while maintaining accuracy (Zhang et al., 2025).
5. **Dataset limitations** remain a significant barrier to progress, with few publicly available multimodal industrial datasets capturing diverse defect types and realistic misalignment conditions (Brenner et al., 2023; Nam, 2025).

The field is rapidly converging toward a unified design philosophy: alignment-aware architectures that learn cross-modal correspondences in feature space, maintain linear computational complexity through state-space or sparse attention mechanisms, and preserve fine spatial details through hierarchical reconstruction. These principles, combined with self-supervised learning and physics-informed constraints, will enable increasingly capable intelligent inspection systems.

Future research should address remaining challenges: developing standardized multimodal datasets with realistic misalignment, advancing few-shot learning for novel defect types, creating explainable fusion systems that build operator trust, and establishing theoretical foundations for cross-modal alignment in industrial settings. As manufacturing continues its transformation toward Industry 4.0, robust multimodal inspection will play an increasingly critical role in ensuring product quality, reducing waste, and enabling autonomous quality control.

References

1. Brenner, M., Reyes, N. H., Susnjak, T., & Barczak, A. L. C. (2023). RGB-D and thermal sensor fusion: A systematic literature review. *IEEE Access*, *11*, 82410–82442.
2. Nam, W. (2025). [Paper review] Multimodal industrial anomaly detection by crossmodal feature mapping. Seoul National University DSBA Seminar.
3. National Technical University of Athens. (2025). Recent advances in sensor fusion monitoring and control strategies in laser powder bed fusion: A review. *Machines*, *13*(9), 820.
4. Tan, P. X., Hoang, D.-C., et al. (2026). MambaAlign: Alignment-aware state-space fusion for RGB-X industrial anomaly detection. *Journal of Computational Design and Engineering*, *13*(1).
5. Wang, Z., Jia, C., He, W., Zhang, X., & Feng, H. (2025). Integrated multi-modal flexible sensors and AI-driven fusion modeling for internal and external quality detection of agricultural products. *Trends in Food Science & Technology*.
6. Zhang, D., Zhou, S., Zheng, Y., & Xu, X. (2025). Review on application of machine vision-based intelligent algorithms in gear defect detection. *Processes*, *13*(10), 3370.

CITATION

Kofa, H. A., Imam, B. A., Ayagi, S. H.& Abubakar, S. B. (2026). Cross-Modal Feature Alignment for Robust Defect Detection in Industrial Inspection: A Comprehensive Review of RGB-Thermal and Multimodal Fusion Techniques. Global Journal of Research in Engineering & Computer Sciences, 6(4), 30–38.
https://doi.org/10.5281/zenodo.21283386

@ 2026 | PUBLISHED BY GJR PUBLICATION, INDIA