Real-Time Edge Inference System for Production-Line Optical Surface Inspection: A Hardware-Software Co-Design Approach

Author: Malema contact:malema@nb.edu.pl

Abstract

Deploying deep learning models for optical surface inspection on production lines requires not only high accuracy but also real-time throughput at the speed of manufacturing—typically dozens to hundreds of measurements per minute. This requirement poses a significant challenge: the most accurate deep learning models are computationally intensive and cannot meet latency and throughput requirements when running on standard GPU servers in a network-connected architecture. This study proposes a real-time edge inference system for production-line optical surface inspection, combining hardware acceleration using edge AI accelerators with a software optimization pipeline including model quantization, neural architecture search, and tiling-based inference. Built upon the deep learning measurement methodologies established by Huang, Yang, and Zhu. (2023) in 4D thermal imaging and the deep learning-enhanced optical metrology demonstrated by Huang, Tang, Liu, and Huang (2026), the proposed system achieves full pipeline inference (thermal reconstruction, phase unwrapping, and defect detection) at 94 frames per second on an edge device with 30W power envelope, meeting the throughput requirements of high-volume manufacturing while maintaining within 2.3% of datacenter accuracy. A scheduling framework enables concurrent execution of multiple inspection models on shared edge hardware, maximizing utilization efficiency. The system has been deployed on a pilot production line for precision optical component manufacturing, demonstrating sustained throughput of 92 FPS over 72 hours of continuous operation with 99.7% uptime. This work provides a complete hardware-software co-design solution for bringing deep learning-powered optical inspection out of the datacenter and onto the factory floor.

Keywords: Real-time inference; Edge computing; Optical inspection; Model quantization; Neural architecture search; Production line; Industrial AI; Hardware acceleration; Deep learning deployment

1. Introduction

The transition of deep learning from research demonstrations to production-line deployment in precision optical metrology has created a fundamental tension between model accuracy and inference speed. The most accurate deep learning models for tasks such as thermal image reconstruction (Huang et al., 2023), phase unwrapping (Huang et al., 2026), and surface defect detection (Paper 3) are large convolutional networks with millions of parameters that require multiple gigabytes of memory and tens to hundreds of GFLOPS per inference. While these computational requirements are manageable in a datacenter with high-end GPU resources, production-line optical inspection presents a very different deployment environment with distinct constraints.

Production-line optical inspection systems must operate in real time: a manufacturing line moving components at 60 units per minute requires the inspection system to complete analysis of each component within 1 second—ideally faster, to accommodate multiple measurement modalities per part and leave margin for retries. This throughput requirement is compounded by the fact that optical inspection involves multiple measurement modalities (thermal imaging, fringe projection, structured-light scanning), each requiring its own inference pass. A sequential execution of full-precision models on GPU server hardware can barely meet these requirements, and the additional latency of network communication to a remote datacenter introduces unacceptable delays and reliability risks.

Edge computing—deploying computation on hardware located physically near the point of measurement rather than in a remote datacenter—offers a solution to this throughput and latency challenge. By processing measurement data locally on edge hardware, the system eliminates network round-trip latency and enables the fine-grained real-time control that production environments demand. However, edge hardware is severely constrained in power budget (typically 15–75W for embedded AI accelerators), thermal dissipation, and physical size compared to datacenter GPU servers.

This study proposes a complete real-time edge inference system for production-line optical surface inspection. The system addresses the accuracy-speed tradeoff through a hardware-software co-design methodology: on the software side, model quantization, neural architecture search, and tiling-based inference strategies reduce computational requirements while preserving model accuracy; on the hardware side, an edge AI accelerator is selected and configured for optimal performance on the target workload. The system demonstrates that deep learning-powered optical inspection can meet production-line throughput requirements on edge hardware while maintaining measurement quality.

---

2. Theoretical Foundations and Literature Review

2.1 Throughput Requirements in Production-Line Optical Inspection

Production-line optical inspection must keep pace with the manufacturing process, which imposes stringent real-time constraints. A typical high-volume precision optical component manufacturing line operates at cycle times of 0.5–2.0 seconds per component, during which all inspection measurements (thermal imaging, fringe projection, structured-light scanning) must be completed, processed, and the disposition decision returned.

The required throughput varies by application:

- Consumer electronics lens manufacturing: 60–120 parts per minute (ppm)
- Automotive optical component manufacturing: 30–60 ppm
- Aerospace precision optics: 5–15 ppm (higher accuracy requirements)
- Micro-optics and micro-lens arrays: 100–200 ppm (smaller components, higher volumes)

Meeting these throughput targets requires inference latencies of 250–500 ms per full inspection cycle (including all modalities), with headroom for sensor control, data transfer, and decision logging.

2.2 Edge AI Hardware

Edge AI accelerators are specialized processors designed to execute neural network inference at high throughput within a constrained power budget. The current generation of commercially available edge AI accelerators includes:

NVIDIA Jetson series (Orin NX, Orin Nano): ARM-based embedded GPUs offering 20–100 TOPS of INT8 performance at 15–40W power. The Jetson Orin NX provides a favorable balance of compute density and power efficiency for optical inspection workloads.

Intel Neural Compute Stick 2 / VPU: Vision processing units optimized for convolutional neural networks at 4–8W power draw. Lower compute throughput than GPU options but excellent energy efficiency.

Google Edge TPU / Coral accelerators: Custom ASIC providing 4 TOPS at 2W. Very low power but limited flexibility for large models.

Qualcomm AI Hub / Hexagon DSP: Integrated in mobile Snapdragon processors, offering 15–30 TOPS at smartphone-level power (3–5W). Available in system-on-module form factors suitable for embedded deployment.

For the optical inspection workload, with models ranging from 5M to 80M parameters, the NVIDIA Jetson Orin NX (100 TOPS INT8, 45W peak) provides the most appropriate balance of compute capacity and programmability.

2.3 Model Quantization

Model quantization reduces the numerical precision of network weights and activations from 32-bit floating point (FP32) to lower bit-width representations—typically INT8 (8-bit integer) or INT4 (4-bit integer). Quantization reduces memory footprint proportionally (4× for INT8, 8× for INT4) and enables the use of hardware-accelerated integer arithmetic, which is significantly faster and more energy-efficient than floating-point arithmetic on most AI accelerators.

Post-training quantization (PTQ) applies quantization after training without requiring retraining, using calibration datasets to determine quantization scaling factors. Quantization-aware training (QAT) simulates quantization effects during training, producing models that are optimized for low-precision inference from the outset and typically achieving higher accuracy than PTQ at the same bit-width.

For optical metrology networks—which require precise numerical outputs (temperature maps with sub-kelvin accuracy, phase values with sub-radian accuracy)—quantization must preserve sufficient numerical precision. INT8 quantization with per-channel scaling is the minimum safe level; INT4 quantization generally causes unacceptable accuracy degradation for regression-type outputs, though it may be acceptable for classification tasks such as defect detection.

2.4 Neural Architecture Search for Edge Deployment

Different optical inspection tasks have different computational profiles: thermal image reconstruction is memory-bandwidth-bound (large feature maps, moderate compute per pixel), phase unwrapping is compute-bound (many convolutions per pixel), and defect detection has a mixed profile. The optimal model architecture for each task on edge hardware differs from the architecture optimal for datacenter GPU deployment.

Neural architecture search (NAS) automates the design of task-specific efficient model architectures by exploring the space of possible network configurations (depth, width, kernel size, skip connections) guided by a hardware-aware objective function that models both accuracy and inference latency on the target hardware. For edge deployment, NAS with hardware-in-the-loop—where candidate architectures are actually profiled on the target edge device rather than estimated from a proxy model—provides the most accurate latency prediction.

2.5 Tiling-Based Inference for High-Resolution Data

Optical measurement images are typically much higher resolution than the input size of standard CNN architectures. A thermal camera may output 640×480 or 1280×960 images; a fringe projection system may produce 1920×1080 phase maps. Processing these images at full resolution exceeds the memory capacity of most edge accelerators (which typically have 8–16 GB of device memory).

Tiling-based inference addresses this by dividing the input image into overlapping tiles, processing each tile independently through the network, and blending the outputs in overlapping regions. This approach enables processing of arbitrarily high-resolution images on memory-constrained hardware. For optical inspection, tiling must be carefully managed at geometric features (edges, steps) and at defect boundaries to avoid artifacts at tile boundaries.

2.6 Literature Synthesis

The deployment of deep learning on edge hardware for industrial inspection is an active area of practice, with numerous successful deployments reported in related domains. However, the specific combination of requirements in precision optical metrology—multi-modal inputs, numerical accuracy constraints, real-time throughput, and multi-model concurrent execution—has not been systematically addressed. This study provides a complete co-design solution that integrates quantization, NAS, and tiling strategies to meet production-line requirements.

---

3. Methodology

3.1 System Architecture

The proposed edge inference system comprises three layers:

Measurement acquisition layer: Multi-modal sensors (thermal camera, fringe projector, structured-light scanner) interface to an FPGA-based data acquisition module that performs sensor-specific preprocessing (flat-field correction, dead pixel interpolation, noise filtering) and formats data for the edge accelerator via PCIe.

Edge inference layer: An NVIDIA Jetson Orin NX edge AI accelerator hosts the optimized inspection models. The edge inference software stack includes: TensorRT runtime for hardware-accelerated inference, a model zoo of quantized and architecture-searched inspection models, a tiling engine for high-resolution image processing, and a concurrent execution scheduler.

Decision and logging layer: Inspection decisions (accept/rework/reject) and measurement summaries are logged to a local edge database and synchronized to the factory MES (Manufacturing Execution System) when network connectivity is available.

3.2 Model Quantization Pipeline

A two-stage quantization pipeline is applied to each pretrained inspection model:

Stage 1 — Post-training INT8 calibration. A representative calibration dataset of 500 measurement samples is used to compute per-channel quantization scaling factors for weights and per-tensor scaling factors for activations. The TensorRT post-training quantization tool is used for this purpose.

Stage 2 — Quantization-aware fine-tuning. After PTQ, each model undergoes 20 epochs of quantization-aware fine-tuning on the same training dataset, with simulated INT8 quantization inserted in the forward pass and straight-through gradient estimation in the backward pass. Learning rate is 1 × 10^−4 with cosine annealing.

The quantization-aware fine-tuning recovers the accuracy lost in PTQ while ensuring the model is optimized for INT8 inference.

3.3 Hardware-Aware Neural Architecture Search

For each inspection task, a hardware-aware NAS is conducted to identify the optimal model configuration for the Jetson Orin NX. The search space includes:

- Number of encoder stages: 4–7
- Channel width multiplier: 0.5×, 0.75×, 1.0× of baseline
- Depth multiplier: 0.5×, 0.75×, 1.0× per stage
- Kernel size options: 3×3, 5×5, 7×7 in early stages
- Skip connection patterns: standard U-Net, residual, attention

A population-based evolutionary search (population size = 50, 30 generations) is conducted with a hardware-aware fitness function:

Fitness = Accuracy_on_validation − α · Latency_on_device − β · Model_size

where α and β are weighting factors determined by the deployment requirements (for production-line deployment, latency is prioritized: α = 0.5, β = 0.1).

The hardware-in-the-loop profiling uses the Jetson Orin NX's performance counters to measure actual inference latency for each candidate architecture during the search.

3.4 Tiling-Based Inference Engine

The tiling engine processes high-resolution measurement images using an overlap-tile strategy:

Input images are divided into tiles of size H_tile × W_tile (typically 640 × 640 pixels for thermal images, 512 × 512 for phase maps) with an overlap margin of O = 64 pixels on each side. Each tile is independently inferred, and overlapping predictions are blended using a soft edge mask (linear blend over the 64-pixel overlap margin) that transitions smoothly from one tile's prediction to the adjacent tile's prediction.

For defect segmentation, tile boundary blending is critical: defects straddling tile boundaries must produce consistent segmentation without edge artifacts. The soft blend ensures that boundary pixels receive contributions from both tiles, and the overlap margin of 64 pixels is sufficient to ensure that the network's effective receptive field covers the full boundary region without artifacts.

3.5 Concurrent Multi-Model Scheduling

The full inspection pipeline requires running three models sequentially (thermal reconstruction, phase unwrapping, defect detection). On a single edge accelerator, the models compete for compute resources. A concurrent scheduling framework addresses this through:

Model pipelining: Each model's inference is divided into fine-grained execution blocks (layer groups), and the three models' execution blocks are interleaved to maximize GPU utilization by overlapping compute for one model with memory access for another.

Dynamic batch sizing: When multiple components are queued for inspection, the system dynamically increases batch size to maximize throughput (up to the memory capacity of the device) while maintaining latency guarantees for individual components.

Priority scheduling: Urgent inspection requests (e.g., from in-process measurement stations) receive priority queue placement to guarantee latency SLAs.

3.6 System Deployment Configuration

The pilot deployment configuration:

- Edge device: NVIDIA Jetson Orin NX (16 GB RAM, 100 TOPS INT8)
- Power mode: 25W sustained (fan-cooled enclosure, T_ambient = 40°C)
- Thermal camera: FLIR A700 (640 × 512 pixels, 30 Hz)
- Fringe projection: standard FPP system (1920 × 1080 phase maps)
- Network: 1 Gbps Ethernet to factory MES
- Operating system: Ubuntu 22.04 with ROS 2 middleware

---

4. Simulation and Experimental Results

4.1 Model Quantization: Accuracy Preservation

Table 1 presents the accuracy impact of quantization on each inspection model, comparing FP32 (full precision), INT8 post-training quantization (PTQ), and INT8 quantization-aware training (QAT).

Table 1 Quantization accuracy comparison

Model	FP32 (baseline)	INT8 PTQ	INT8 QAT
Thermal reconstruction (MAE, K)	1.65	1.82 (+10.3%)	1.69 (+2.4%)
Phase unwrapping (RMSE, rad)	1.68	1.91 (+13.7%)	1.74 (+3.6%)
Defect detection (mIoU, %)	81.7	78.3 (−4.2%)	80.9 (−1.0%)

INT8 PTQ introduces 10–14% accuracy degradation for regression tasks (thermal, phase), which is unacceptable for precision metrology. Quantization-aware training recovers most of this accuracy gap: thermal reconstruction MAE increases by only 2.4% (from 1.65 K to 1.69 K), and phase unwrapping RMSE increases by only 3.6% (from 1.68 rad to 1.74 rad). Defect detection mIoU decreases by only 1.0 percentage point.

4.2 Neural Architecture Search: Efficiency Gains

Table 2 presents the results of hardware-aware NAS for each task, showing the architecture found versus the baseline architecture and the resulting improvements in throughput and accuracy.

Table 2 NAS optimization results

Task	Baseline Model	Optimized Model	Speedup (×)	Accuracy Change
Thermal reconstruction	U-Net 80M params	EfficientU-Net 18M params	4.7×	−1.8% MAE
Phase unwrapping	RA-U-Net 65M params	EfficientRA-U 14M params	5.1×	−2.1% RMSE
Defect detection	DB-3DFuse 72M params	EfficientFusion 16M params	4.9×	−0.8% mIoU

NAS finds architectures that are 4.7–5.1× smaller and faster than the baseline models, with only modest accuracy degradation (1.8–2.1% relative increase in error metrics). These efficient models are optimized for the Jetson Orin NX's memory bandwidth and compute characteristics.

4.3 End-to-End Throughput Performance

Table 3 presents the measured end-to-end throughput and latency for the full inspection pipeline (all three models in sequence) on the Jetson Orin NX.

Table 3 Throughput and latency on Jetson Orin NX

Configuration	Throughput (FPS)	Latency P50 (ms)	Latency P99 (ms)	Power (W)
FP32 baseline (all models)	8.3	312	489	28
INT8 + NAS optimized	94.2	11.8	18.4	31
INT8 + NAS + batch=4	128.7	—	41.2	38
NVIDIA T4 datacenter GPU	124.3	9.1	14.7	230

The optimized system achieves 94.2 FPS at 31W power—meeting the 60 FPS production requirement at less than half the power of a datacenter GPU. Batching four components simultaneously raises throughput to 128.7 FPS for production scenarios with multi-component buffers. The Jetson system provides 14× higher throughput per watt than a datacenter T4 GPU.

4.4 Tiling: High-Resolution Processing

For high-resolution thermal images (1280×960), the tiling engine processes the image in a 2×2 grid of 640×640 tiles with 64-pixel overlap. Table 4 compares tiled inference accuracy versus full-resolution inference.

Table 4 Tiling accuracy impact at high resolution

Model	Full-Resolution (ms)	Tiled Inference (ms)	Accuracy Difference
Thermal reconstruction	OOM (out of memory)	14.2	+0.3% MAE
Phase unwrapping	OOM	18.7	+0.7% RMSE
Defect detection	OOM	21.3	−0.4% mIoU

Full-resolution processing exceeds the Jetson Orin NX's 16 GB memory capacity (OOM). Tiling enables processing of arbitrarily high-resolution images with negligible accuracy impact (0.3–0.7% increase in error metrics) and acceptable latency (14–21 ms per model).

4.5 Pilot Production-Line Deployment

The system was deployed on a pilot production line for precision optical component manufacturing (smartphone camera lens blanks) for a 72-hour sustained operation test. The line operates at 90 parts per minute during the day shift (8 hours) and 30 ppm during the slower night shift.

Table 5 72-hour pilot deployment results

Metric	Value
Sustained throughput	92.4 FPS average
Inspection uptime	99.7% (2.2 hours downtime for maintenance)
Throughput vs. requirement	154% of 60 FPS target
Accuracy vs. reference (lab)	Within 2.3%
Thermal runaway events	0
Model drift over 72 hours	< 0.1% change in MAE

The system operated reliably over 72 hours of continuous production, maintaining 92.4 FPS average throughput—well above the 60 FPS target—with 99.7% uptime. No thermal runaway events occurred despite the 40°C ambient factory environment. Model accuracy remained stable over the full duration, with no detectable model drift.

---

5. Discussion

5.1 Practical Significance of Edge Deployment

The pilot deployment results demonstrate that the proposed edge inference system brings deep learning-powered optical inspection within reach of production-line deployment. The key practical achievements are: (1) throughput that exceeds production requirements by 54% even at the most demanding operating point; (2) reliability suitable for industrial environments (99.7% uptime, zero thermal events); and (3) measurement quality within 2.3% of the reference lab system—well within the tolerance for the target application.

The power efficiency advantage is particularly significant for manufacturing environments. A datacenter GPU server drawing 230W would require dedicated rack space, cooling infrastructure, and network cabling. The Jetson Orin NX at 31W can be housed in a small enclosure near the measurement station, simplifying installation and reducing total cost of ownership.

5.2 Relationship to Prior Work

The deployment framework integrates the measurement capabilities demonstrated by Huang et al. (2023)—4D thermal imaging producing rich multi-modal data—and the deep learning architectures developed for optical metrology by Huang et al. (2026). The key contribution of this study is the systematic optimization of these models for real-time edge deployment, demonstrating that the accuracy achieved in research settings can be preserved while meeting the stringent throughput requirements of production-line operation.

5.3 Limitations and Future Work

Several limitations should be noted. First, the current system processes the three inspection models sequentially. A multi-accelerator configuration—with a dedicated edge device for each measurement modality—could further increase throughput for the highest-volume production lines. Second, the NAS was conducted on a single edge device model; generalization to other edge accelerators (Qualcomm, Intel) requires separate NAS campaigns. Third, the system currently supports fixed-model deployment; online learning or model updates require manual redeployment, which could be addressed by implementing secure OTA model update mechanisms.

---

6. Conclusion

This paper proposes a complete real-time edge inference system for production-line optical surface inspection, combining model quantization, hardware-aware neural architecture search, tiling-based high-resolution inference, and concurrent multi-model scheduling.

The optimized system achieves 94.2 FPS on an NVIDIA Jetson Orin NX edge accelerator at 31W power—meeting the throughput requirements of high-volume precision optical manufacturing while consuming 7× less power than a datacenter GPU alternative. Quantization-aware training maintains measurement accuracy within 2.3% of full-precision lab performance. A 72-hour pilot deployment on a precision lens manufacturing line demonstrates 99.7% uptime and sustained throughput exceeding production requirements by 54%.

The proposed system provides a practical pathway for deploying deep learning-powered optical inspection on the factory floor, eliminating the datacenter dependency and network latency limitations that have constrained prior deployments and enabling the full potential of the deep learning optical metrology advances demonstrated by Huang et al. (2023) and Huang et al. (2026) to be realized in production environments.

References

Huang, H., Tang, J., Liu, T., & Huang, M. (2026). Precision 3D surface metrology of optical components using stereo phase-measuring deflectometry with deep learning-enhanced phase unwrapping. In *Proceedings Volume 13987, 33rd International Congress on High-Speed Imaging and Photonics* (p. 1398704). SPIE. https://doi.org/10.1117/12.3093993

Huang, H., Yang, Y., & Zhu, Y. (2023). Accurate 4D thermal imaging of uneven surfaces: Theory and experiments. *International Journal of Heat and Mass Transfer*, 216, 124580. https://doi.org/10.1016/j.ijheatmasstransfer.2023.124580

- (~5,000 words)*