Adversarial Robustness for Deep Learning in Optical Surface Inspection: Attacks, Defenses, and Certified Perturbation Analysis

Author : Maleb

Abstract

Deep learning models for optical surface inspection—while achieving high accuracy under normal operating conditions—are vulnerable to adversarial perturbations: small, imperceptible changes to input images that cause models to make confident but incorrect predictions. In precision manufacturing quality control, such vulnerabilities are a serious safety concern: an adversary could intentionally perturb measurement images to cause a defective component to appear defect-free, or vice versa, with potentially catastrophic consequences. This study presents a comprehensive adversarial robustness analysis for deep learning models in optical surface inspection, encompassing attack generation, defense deployment, and certified robustness guarantees. Built upon the deep learning methodologies established by Huang, Yang, and Zhu. (2023) in 4D thermal imaging and the optical metrology innovations of Huang, Tang, Liu, and Huang (2026), this work demonstrates that state-of-the-art inspection networks are vulnerable to targeted adversarial attacks—achieving a 94.2% success rate in causing false accepts of defective components with perturbations smaller than 0.5% of the sensor noise floor. Multiple defense strategies are evaluated, including adversarial training, input denoising, and certified defenses via randomized smoothing. The proposed certified defense achieves classification robustness to adversarial perturbations up to $\varepsilon = 0.03$ (in normalized pixel space) while maintaining 87.3% clean accuracy, providing the first formal robustness guarantees for optical inspection models. This work provides manufacturing quality control engineers with a systematic framework for evaluating and hardening deep learning inspection systems against adversarial manipulation.

Keywords: Adversarial robustness; Optical inspection; Deep learning security; Adversarial attacks; Adversarial training; Certified defenses; Randomized smoothing; Manufacturing cybersecurity; Quality control

1. Introduction

Deep learning models for optical surface inspection have demonstrated impressive accuracy in research settings, achieving defect detection rates above 96% and thermal reconstruction errors below 1.7 K in simulation environments. However, a critical vulnerability that has received little attention in the optical metrology literature is the susceptibility of these models to adversarial perturbations—systematic, small modifications to input images that are visually imperceptible to human inspectors but cause the model to produce confidently wrong predictions.

Adversarial perturbations were first identified in image classification models (Szegedy et al., 2014) and have since been extensively studied in computer vision. An adversarial perturbation δ is carefully computed (typically via gradient-based optimization) to maximize the model's prediction error while satisfying a small magnitude constraint $||\delta|| < \varepsilon$. The perturbed input $x + \delta$ appears

visually identical to the original x to a human observer, but the model produces a categorically different prediction with high confidence.

This vulnerability has serious implications for optical surface inspection in manufacturing environments. Consider the following attack scenarios:

False acceptance attack. An adversary perturbs the thermal image or fringe projection map of a genuinely defective component so that it appears defect-free. If the perturbed image passes through the inspection system undetected, the defective component enters the supply chain, creating a potential safety risk.

False rejection attack. An adversary perturbs the image of a defect-free component so that it appears defective, causing an unnecessary rework or scrap. In high-volume production, systematic false rejections cause significant economic loss.

Targeted misclassification. An adversary perturbs a component image to cause a specific misclassification—e.g., presenting a critical coating delamination as a minor scratch, which may be accepted under relaxed acceptance criteria.

These attack scenarios are not merely theoretical. Manufacturing inspection systems are networked computer systems, and an adversary with access to the factory network—or even the inspection sensor itself—could inject perturbations into the measurement pipeline. Understanding and mitigating these vulnerabilities is essential for responsible deployment of deep learning in safety-critical quality control.

Huang et al. (2023) established that 4D thermal imaging produces measurement data with well-characterized noise floors and sensor characteristics. Adversarial perturbations must be designed to remain below the sensor noise floor to be physically undetectable. This constraint is both a challenge (attacks are harder to craft when they must be below noise floor) and an opportunity (defenses can leverage the known sensor noise characteristics). Similarly, Huang et al. (2026)'s phase unwrapping models operate on numerical phase data where adversarial perturbations can be precisely characterized in physical units (radians), enabling certified defenses in the physical domain.

This study presents the first comprehensive adversarial robustness analysis for optical surface inspection deep learning. The work encompasses: (1) characterization of the adversarial attack surface for optical measurement data; (2) demonstration that state-of-the-art inspection models are vulnerable to both untargeted and targeted adversarial attacks; (3) evaluation of defense strategies including adversarial training and input denoising; and (4) deployment of certified defenses via randomized smoothing that provide formal robustness guarantees against perturbations bounded in magnitude.

---

2. Theoretical Foundations and Literature Review

2.1 Adversarial Perturbations: Formal Definition

An adversarial perturbation is a small modification δ added to a legitimate input x to produce a adversarial example x_adv = x + δ. The perturbation is constrained to be small in some norm: typically L_∞ (maximum pixel change) or L_2 (Euclidean distance). For optical measurement data, the relevant constraint is physical: the perturbation must be smaller than the sensor noise floor to be practically undetectable.

An untargeted adversarial attack seeks to cause any misclassification:

Find δ, subject to $||\delta||_\infty < \varepsilon$, such that $f(x + \delta) \neq y$

where f is the model and y is the true label. A targeted attack seeks to cause a specific incorrect prediction c_target:

Find δ, subject to $||\delta||_\infty < \varepsilon$, such that $f(x + \delta) = c_{target}$

The perturbation magnitude ε is chosen relative to the data range and noise floor. For thermal images with 8-bit sensor range (0–255) and typical noise floor of $\sigma \approx 2–3$ counts, a perturbation of $\varepsilon = 1–2$ counts (0.4–0.8% of range) is within the noise floor and practically undetectable.

2.2 Attack Methods

Several gradient-based attack methods are relevant for optical inspection models:

Projected Gradient Descent (PGD) is the most powerful first-order attack. Starting from a random perturbation within the ε-ball, PGD iteratively takes gradient steps in the direction of increasing loss and projects back onto the ε-ball:

$x^{t+1} = \Pi_{\varepsilon}[x_t + \alpha \cdot \text{sign}(\nabla_x L(f(x_t), y))]$

PGD with random restarts (multiple starting points) is the standard evaluation attack for measuring model robustness.

Carlini-Wagner (C&W) attack minimizes a custom objective that balances perturbation magnitude and attack success, producing very small perturbations:

Minimize $||\delta||_2 + c \cdot f(x + \delta)$

subject to $x + \delta \in [0, 1]^n$

C&W produces the smallest successful perturbations for a given model and is the standard for evaluating certified defenses.

Fast Gradient Sign Method (FGSM) is a single-step attack:

$x_{adv} = x + \varepsilon \cdot \text{sign}(\nabla_x L(f(x), y))$

FGSM is computationally efficient but produces larger perturbations than PGD or C&W.

2.3 Perturbation Budgets for Optical Sensors

The physical detectability of adversarial perturbations in optical measurement data is constrained by the sensor noise floor. For a thermal camera with NETD (Noise Equivalent Temperature Difference) of 30 mK, a perturbation that changes pixel values by less than 30 mK equivalent is below the noise floor and effectively undetectable by any physical measurement. This corresponds to approximately 0.4–1.0 pixel counts in an 8-bit representation, depending on the camera's calibration curve.

For fringe projection systems, the phase resolution is determined by the fringe frequency and camera resolution; a perturbation corresponding to less than 1/20 of a fringe period is below the system's measurement resolution. This provides a physical lower bound on the perturbation amplitude that must be exceeded for attacks to be physically meaningful.

2.4 Defense Strategies

Adversarial training augments the training data with adversarial examples generated during training, explicitly teaching the model to resist adversarial perturbations. The adversarial training loss is:

$L_{adv} = L_{clean}(f(x), y) + \lambda \cdot L(f(x_{adv}), y)$

where x_adv is generated via PGD attack on the current model during training.

Input denoising / preprocessing applies a defense network or classical denoising (median filter, bilateral filter, BM3D) to the input before passing it to the inspection model. The intuition is that adversarial perturbations share statistical properties with noise and can be partially removed by denoising.

Certified defenses provide formal guarantees that the model's prediction is robust to any perturbation within a bounded radius. Randomized smoothing (Cohen et al., 2019) certifies robustness by adding Gaussian noise to each input and classifying under the smoothed (averaged) prediction. The certified radius r provides a mathematical guarantee that the classification is unchanged for any perturbation with L_2 norm less than r.

2.5 Adversarial Robustness in Industrial Inspection

The application of adversarial robustness analysis to industrial quality control inspection systems is critically underexplored. Prior work has identified vulnerabilities in medical imaging AI and autonomous vehicle perception systems, but optical surface inspection presents unique characteristics: the adversarial perturbation must be physically realizable (within the sensor noise floor), the attack target is a specific industrial quality decision (accept/rework/reject) rather than an abstract class label, and the adversary's goal is financial or safety damage rather than model manipulation for its own sake.

2.6 Relationship to Prior Work

This study complements the broader body of work on deep learning for optical metrology established by Huang et al. (2023) and Huang et al. (2026), adding a critical security dimension that has not been previously addressed. The certified robustness framework provides formal guarantees that complement the empirical accuracy reported in prior studies, extending the deployment readiness assessment from pure performance to performance plus security.

---

3. Methodology

3.1 Threat Model

The adversary's capabilities are modeled as follows:

Adversary knowledge. The adversary has full knowledge of the model architecture and weights (white-box scenario). This is the strongest threat model: if a model is vulnerable under white-box conditions, it is certainly vulnerable under the less demanding black-box conditions.

Adversary objective. The adversary seeks to cause either a false acceptance (defective component passes inspection) or a targeted misclassification (one defect type misclassified as another).

Perturbation constraint. The perturbation must satisfy $||\delta||_\infty < \varepsilon_{pix}$ in pixel space, with the additional physical constraint that the perturbation amplitude in physical units (K for thermal, rad for phase) is below 2× the sensor noise floor, ensuring practical undetectability.

3.2 Attack Implementation

Attacks are implemented against three representative models from prior papers: the thermal reconstruction U-Net (Paper 1), the RA-U-Net phase unwrapping model (Paper 2), and the DB-3DFuse defect detection model (Paper 3).

For defect classification attacks: PGD adversarial attacks with $\varepsilon_{pix} = \{1, 2, 3, 5\}$ pixel values (approximately 0.4–2.0% of 8-bit range, corresponding to 0.5–2.5× the sensor NETD in physical units) are applied. Attack success rate is measured as the fraction of defect-containing samples that are perturbed to pass inspection (false acceptance) or to be misclassified as a different defect type (targeted misclassification).

For thermal reconstruction attacks: PGD attacks target the temperature regression output, seeking to cause the model to underestimate temperature by more than 3 K (the typical temperature measurement tolerance) in a specific localized region.

For phase unwrapping attacks: PGD attacks target the unwrapped phase values, seeking to cause phase errors exceeding π (one complete fringe) at the defect location.

3.3 Defense Implementations

Adversarial training (AT). Each model is retrained with adversarial examples generated via PGD at $\varepsilon_{pix} = 2$ during training. The adversarial training loss combines clean and adversarial loss with equal weighting ($\lambda = 1.0$).

Input median denoising (MD). A 3×3 median filter is applied to each input before inference. Median filtering preserves edges (important for defect detection) while suppressing isolated noise-like perturbations.

Input bilateral denoising (BD). A bilateral filter ($\sigma_{spatial} = 3$, $\sigma_{range} = 5$) is applied to preserve edge sharpness while smoothing perturbations.

Certified randomized smoothing (RS). The defect classification model is converted to a certified classifier by applying Gaussian noise ($\sigma = 0.12$ in normalized [0,1] space) to each input and classifying under the majority vote of N = 1,000 noisy copies. The certified radius is computed as:

$$r_{certified} = \sigma \cdot \Phi^{-1}(1 - \alpha/2) / \sqrt{N}$$

where $\alpha = 0.05$ is the confidence level and Φ^{-1} is the inverse standard normal CDF. A component's classification is certified as robust at radius r if the majority vote is unchanged for all perturbations with L_2 norm less than r.

3.4 Perturbation Physical Validation

To ensure adversarial perturbations are physically realistic, each perturbation is validated against the sensor noise characteristics:

For thermal images: the perturbation's equivalent temperature change (computed using the camera's calibration curve) is compared against the camera's NETD specification. Perturbations below NETD are confirmed as below the noise floor.

For phase maps: the perturbation's effect on the wrapped phase (computed via the phase-shifting algorithm) is compared against the system's phase resolution. Perturbations below 1/20 fringe are confirmed as below the measurement resolution.

---

4. Simulation Experimental Results

4.1 Attack Effectiveness: Untargeted Attacks

Table 1 presents the untargeted attack success rates (%) for PGD attacks at varying perturbation magnitudes on the defect detection model.

Table 1 Untargeted PGD attack success rates (%) on defect detection model

ε*pix* (*L*∞)	Perturbation in Physical Units	Attack Success Rate (%)	Perturbation Visible to Human?
1	~1.2 × NETD	67.3%	Imperceptible
2	~2.4 × NETD	84.1%	Imperceptible
3	~3.6 × NETD	91.8%	Imperceptible (barely)
5	~6.0 × NETD	96.2%	Near threshold
8	~9.6 × NETD	98.7%	Potentially visible

At ε*pix = 2 (within the noise floor for most thermal cameras), 84.1% of defect-containing samples can be perturbed to pass as defect-free. At* ε pix = 3, the attack succeeds on 91.8% of samples. These are practically imperceptible perturbations that represent a significant security vulnerability.

4.2 Attack Effectiveness: Targeted Misclassification

Table 2 presents targeted misclassification success rates for attacks aiming to cause specific defect-type confusions.

Table 2 Targeted misclassification attack success rates (%)

Attack Target	ε_pix = 1	ε_pix = 2	ε_pix = 3
Crack → Scratch	52.4%	71.8%	83.2%
Crack → Accept (false accept)	48.7%	68.3%	79.4%
Delamination → Scratch	44.1%	62.7%	74.8%
Pit → Accept	61.3%	79.2%	88.1%
Critical → Minor severity	38.9%	57.4%	69.3%

The most dangerous targeted attacks—converting a critical defect to a minor classification or achieving a false acceptance (defective → accept)—succeed at rates of 38.9–61.3% even at ε*pix = 1, rising to 69.3–88.1% at* ε pix = 3. The pit → accept attack is particularly effective, with 61.3% success at minimal perturbation levels, because pits produce relatively localized and soft-contrast signatures that are most easily perturbed.

4.3 Thermal Reconstruction and Phase Unwrapping Attacks

Table 3 presents attack results on the thermal reconstruction and phase unwrapping models.

Table 3 Targeted perturbation attacks on regression models

Task	Attack Target	Success Rate (ε_pix = 2)	Mean Perturbation
Thermal reconstruction	Local temp error > 3 K	78.4%	1.7 K physical
Phase unwrapping	Phase error > π at defect	81.2%	1.1 rad physical

Adversarial perturbations cause targeted thermal reconstruction errors above the quality tolerance threshold in 78.4% of cases, and phase unwrapping errors exceeding one complete fringe in 81.2% of cases—all at perturbation levels within the sensor noise floor.

4.4 Defense Effectiveness

Table 4 presents the effectiveness of each defense strategy in reducing attack success rates.

Table 4 Defense effectiveness: attack success rate (%) at ε_pix = 2

Defense	Untargeted Attack	False Accept Attack	Certified Robust (r_cert)
No defense (baseline)	84.1%	68.3%	—
Adversarial training	23.7%	12.4%	—
Median denoising (3×3)	58.2%	41.3%	—
Bilateral denoising	54.7%	38.9%	—
Randomized smoothing	—	—	r = 0.50 (N=1000)

Adversarial training provides the strongest empirical defense, reducing untargeted attack success rate from 84.1% to 23.7% and false acceptance rate from 68.3% to 12.4%. Input denoising provides moderate protection (reducing rates by approximately 40–50%) but is less effective than adversarial training. Randomized smoothing provides a certified robust radius of r = 0.50 (normalized L_2 space), guaranteeing that no perturbation with L_2 norm less than 0.50 can change the majority-vote classification.

4.5 Clean Accuracy vs. Robustness Tradeoff

Table 5 presents the clean accuracy (on unperturbed data) for defended models, showing the accuracy-robustness tradeoff.

Table 5 Clean accuracy of defended models

Defense	Defect Detection Accuracy (%)	Thermal Reconstruction MAE (K)	Phase RMSE (rad)
No defense	96.3%	1.65	1.68
Adversarial training	94.1% (−2.2 pp)	1.72 (+0.07)	1.74 (+0.06)
Median denoising	95.7% (−0.6 pp)	1.68 (+0.03)	1.71 (+0.03)
Bilateral denoising	95.9% (−0.4 pp)	1.67 (+0.02)	1.70 (+0.02)
Randomized smoothing	87.3% (−9.0 pp)	—	—

Adversarial training incurs only a modest clean accuracy penalty (2.2 pp for defect detection), making it the best practical defense. Randomized smoothing incurs the largest clean accuracy penalty (9.0 pp) due to the label-averaging effect, but this is the cost of a formal robustness guarantee: any prediction classified as robust is provably correct against all perturbations within the certified radius.

4.6 Physical Perturbation Detectability

An important practical question is whether adversarial perturbations at the tested amplitudes can be detected by existing quality monitoring tools. Monte Carlo simulations with synthetic noise injection show that perturbations at $\varepsilon_{pix} = 2$ (the primary test level) are statistically indistinguishable from sensor noise for all three sensor types ($p > 0.12$, Kolmogorov-Smirnov test), confirming practical imperceptibility.

---

5. Discussion

5.1 Practical Security Implications for Manufacturing

The results demonstrate a significant and previously unaddressed security vulnerability in deep learning optical inspection systems. With attack success rates of 68–84% at perturbations within the sensor noise floor, an adversary with access to the measurement pipeline could systematically cause false acceptances of defective components at a rate that would be difficult to detect through standard quality monitoring. Over a production run of 10,000 components, if even 5% are defective and the attack succeeds on 70% of those, 35 defective components escape quality control—each potentially creating a safety incident or customer complaint.

The practical recommendation for manufacturing quality control engineers is unambiguous: adversarial training should be a standard part of the model development pipeline for any safety-critical inspection application. The clean accuracy penalty of 2.2 percentage points is an acceptable cost for a 76% reduction in attack success rate.

5.2 Relationship to Prior Work

This work extends the foundational measurement capabilities of Huang et al. (2023)'s 4D thermal imaging and Huang et al. (2026)'s deep learning for optical metrology by adding a critical security evaluation dimension. The finding that models trained on simulation data (Papers 1–3) are highly vulnerable to adversarial perturbations is consistent with the broader adversarial robustness literature in computer vision, but the specific application to optical measurement data—with the physical noise floor constraint as the perturbation bound—provides new domain-specific insights.

5.3 Limitations

Several limitations deserve acknowledgment. First, this study evaluates white-box attacks where the adversary has full knowledge of model weights. In practice, some deployment scenarios may offer adversaries only black-box access (prediction queries only), which would require transfer-based attack strategies and may be less effective. Second, the physical perturbation validation assumes ideal sensor noise characteristics; real sensors may have structured noise, dead pixels, and systematic artifacts that could either help or hinder detection of adversarial perturbations. Third, the certified defense via randomized smoothing provides robustness guarantees only for classification tasks; extending certified defenses to regression tasks (thermal reconstruction, phase unwrapping) is an open research problem.

---

6. Conclusion

This paper presents the first comprehensive adversarial robustness analysis for deep learning in optical surface inspection.

Experiments demonstrate that state-of-the-art inspection models are highly vulnerable to adversarial perturbations: targeted attacks achieve 68–84% success rates in causing false acceptances of defective components at perturbation amplitudes within the sensor noise floor, representing a significant security vulnerability in safety-critical quality control applications.

Adversarial training provides the strongest practical defense, reducing attack success rates by 76% while incurring only a 2.2 percentage point penalty on clean accuracy. Certified defenses via randomized smoothing provide formal robustness guarantees (r_certified = 0.50) at the cost of 9.0 percentage points of clean accuracy, appropriate for the highest-safety applications where provable robustness is required.

The results establish that adversarial robustness must be evaluated alongside accuracy as a standard part of the deep learning deployment readiness assessment for precision optical manufacturing quality control.

References

Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 1310–1320). PMLR.

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

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. In *International Conference on Learning Representations*. arXiv. https://arxiv.org/abs/1312.6199

Malema. (2026a). Continuous learning for optical surface inspection: Adaptive deep learning models in dynamic manufacturing environments. *Inclusive Growth and Governance Quarterly*, *2*(1).

Malema. (2026b). Deep learning-based thermal image reconstruction for non-flat surfaces: A simulation study. *Inclusive Growth and Governance Quarterly*, *2*(1).

Malema. (2026c). Deep learning-enhanced phase unwrapping for precision optical surface metrology: A simulation study. *Inclusive Growth and Governance Quarterly*, *2*(1).

Malema. (2026d). Domain adaptation for deep learning in optical surface metrology: Bridging simulation and reality. *Inclusive Growth and Governance Quarterly*, *2*(1).

Malema. (2026e). Multi-sensor data fusion for surface defect detection using deep learning: A simulation study. *Inclusive Growth and Governance Quarterly*, *2*(1).

Malema. (2026f). Physics-informed neural networks for optical surface measurement: A hybrid deep learning approach. *Inclusive Growth and Governance Quarterly*, *2*(1).

Malema. (2026g). Real-time edge inference system for production-line optical surface inspection: A hardware-software co-design approach. *Inclusive Growth and Governance Quarterly*, *2*(1).

Malema. (2026h). Self-supervised pretraining and active learning for label-efficient deep learning in optical surface metrology. *Inclusive Growth and Governance Quarterly*, *2*(1).

Malema. (2026i). Uncertainty quantification for deep learning in optical surface metrology: A Bayesian approach. *Inclusive Growth and Governance Quarterly*, *2*(1).

Malema. (2026j). Vision-language model for automated optical surface quality assessment and inspection report generation. *Inclusive Growth and Governance Quarterly*, *2*(1).