Computer Vision-Assisted Quality Inspection of 3D-Printed Structures Using Deep Neural Networks

Randolf Grein Groth
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
randolfwork@unh.edu

Dewi van Mare
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
dvanmare@uc.edu

Abstract

Layer-wise imaging in additive manufacturing generates large volumes of data whose interpretation remains largely manual or rule-based, a gap that keeps vision systems in the role of qualitative monitoring aids rather than qualification instruments. This paper develops a methodological framework for computer vision-assisted inspection of 3D-printed structures in which deep neural networks are embedded in an explicit chain running from image formation and build-frame registration through volumetric defect segmentation to a calibrated, cost-sensitive acceptance decision. The framework specifies an observation model that links surface reflectance and orientation to measured intensity, a similarity registration that places every layer image in the build coordinate frame so voxel labels remain geometrically consistent, and a three-dimensional encoder–decoder that predicts per-voxel class posteriors. The analysis shows why segmentation accuracy alone is an insufficient basis for inspection: uncalibrated posteriors displace operating points, and the cost-optimal decision threshold is fixed by the ratio between the cost of accepting a defective part and the cost of rejecting a sound one. Evaluation logic is specified at voxel, defect-instance, and part level, with an explicit treatment of registration error propagation, powder masking, specular reflection, and label noise inherited from computed tomography. The resulting specification separates what the network estimates from how a printed part is judged, and it identifies the conditions under which the separation holds.

Keywords

additive manufacturing; in-situ monitoring; semantic segmentation; uncertainty calibration; cost-sensitive decision making; porosity detection

1. Introduction

Additive manufacturing has moved from prototyping into load-bearing and production roles, yet the qualification routes available to it remain comparatively immature. Process–structure–property relationships vary with machine, feedstock, and build orientation, and the resulting qualification burden is difficult to discharge through end-point inspection alone [1]. Because defects are generated layer by layer, the information needed to attribute a flaw to a process excursion exists during the build rather than after it, which is the central motivation for in-situ monitoring [2]. Reviews of the monitoring literature distinguish between signature-based methods that track a small number of derived indicators and imaging methods that retain spatially resolved information, and they consistently identify interpretation of the latter

as the limiting step [3]. The bottleneck is therefore not sensor availability but the conversion of high-dimensional imagery into decisions that a qualification engineer can defend.

Computer vision has changed what can be extracted from such imagery. Convolutional networks replaced hand-designed descriptors as the dominant representation for recognition tasks after large-scale supervised training demonstrated their advantage [4], and fully convolutional architectures subsequently made dense per-pixel prediction practical by replacing classification heads with upsampling decoders [5]. Encoder–decoder designs with skip connections then recovered the fine spatial detail lost to pooling, a property that matters when the target occupies a small fraction of the image [6]. These developments provide the representational machinery for layer-wise inspection, but they were formulated for photographic and biomedical data in which the relationship between scene geometry and pixel values is comparatively benign.

Application work in additive manufacturing has followed the same trajectory with domain-specific adaptations. Multi-scale convolutional models have been used for anomaly detection and classification in laser powder bed fusion imagery, where the scale of a defect relative to the melt track determines the receptive field required [7]. Supervised learning on high-resolution layer images has been applied to defect detection during metallic powder bed fusion, with the training signal drawn from post-build characterization [8]. Variant geometry has been addressed through deep learning on layerwise profiles, which introduces the important observation that nominal geometry varies across a build and that a detector trained on one cross-section need not transfer to another [9]. A related line of work frames machine learning as a way of re-perceiving printed artifacts rather than merely measuring them, treating the learned model as an instrument that extends how printed form is observed and interpreted [10]. Across these studies the detector is the object of study; the surrounding measurement chain, the origin of labels, and the decision that follows detection are typically compressed into a brief implementation note.

This paper addresses that compression directly. The contribution is a single, internally consistent framework for vision-assisted inspection of printed structures in which four components are specified jointly: a physically interpretable observation model that explains what a defect does to a measured pixel value; a registration procedure that maps layer imagery into the build coordinate frame so that voxel labels are geometrically meaningful; a volumetric segmentation network whose training objective is matched to the severe class imbalance of defect data; and a decision layer that converts network posteriors into accept or reject actions under asymmetric costs after calibration. The framework is analytical rather than experimental: it derives the relationships that any such system must satisfy, identifies where each relationship breaks, and specifies the evidence that would be required to validate a deployed instance. The argument proceeds from problem definition through sensing and architecture to uncertainty and decision rules, then to evaluation logic and the failure modes that determine whether reported performance survives contact with a production build.

2. Problem Definition and Defect Taxonomy

2.1 Inspection geometry and the build frame

An inspection problem is defined by three geometric objects. The nominal object M is the voxelized representation of the design intent, derived from the computer-aided design model and the slicing parameters. The observed object O is the material actually present, which is only accessible through measurements. The task is to characterize the discrepancy between M

and O with sufficient spatial fidelity that a defect can be localized, classified, and judged. Ex-situ inspection measures O after the part leaves the machine and after support removal, so the part frame is recoverable from fixturing. In-situ inspection measures O while powder or resin still surrounds the part, so the observable surface is partly the part and partly the surrounding bed.

The measurement itself obeys a reflectance relationship. For a matte surface under a dominant directional illuminant, with interreflection neglected, the recorded intensity at image coordinate x is well approximated by a local shading model:

$$I(x) = \rho(x)\cdot(n(x)\cdot l) + a(x) + \varepsilon(x)$$

Here I(x) is the measured intensity, ρ(x) is the diffuse albedo of the surface patch imaged at x, n(x) is its unit surface normal, l is the unit illumination direction, a(x) is an additive ambient term that varies slowly across the field of view, and ε(x) is sensor and quantization noise. The model is deliberately simple, and its value is diagnostic rather than predictive. Lack-of-fusion porosity raises local roughness, which lowers ρ and randomizes n, so such regions appear darker and more textured than surrounding solid material. Balling and spatter create isolated high-curvature features whose normals deviate strongly from the mean surface orientation, producing bright specular highlights where n aligns with l and dark crescents elsewhere. Warpage changes n over long spatial scales without changing ρ, which is why geometric distortion and contamination can be confused by intensity-only detectors but separated when shading and geometry are estimated jointly. Every failure of the model — a polished surface that violates the diffuse assumption, a deep crevice that receives light from adjacent walls, a powder layer that occludes the part entirely — corresponds to a documented failure mode discussed in Section 5.

Registered measurement requires a mapping from camera coordinates to build coordinates. For a fixed camera and a part that does not move relative to the build plate, the appropriate family is the similarity transform:

$$p_b = s\cdot R\cdot p_c + t$$

where p_c is a point expressed in camera coordinates, p_b is the same point in build coordinates, s is a scalar scale factor, R is a rotation matrix in the special orthogonal group, and t is a translation vector. The scale term is retained because magnification differs between calibrated and uncalibrated optics and between imaging configurations; with telecentric optics and a fixed working distance, s is unity and the transform reduces to a rigid motion. The essential requirement is that the same physical point carries the same coordinates in every layer image, so that a sequence of two-dimensional observations can be stacked into a volume.

Given registered layer imagery, a binary occupancy field is obtained by thresholding and voxelizing the measured material, and the deviation field follows by comparison with the nominal voxelization:

$$\delta(v) = O(v) - M(v)$$

The index v runs over voxels of the build-frame grid, O(v) equals one when measured material exceeds a detection threshold in that voxel and zero otherwise, and M(v) is the corresponding nominal occupancy from the sliced model. Positive δ indicates excess material such as spatter adhesion or dross; negative δ indicates missing or unfused material. Because O is derived from thresholded intensity, δ inherits the partial-volume problem: a voxel straddling a surface intersects both material and void, and its assigned label depends on the

threshold rather than on the material fraction. This is the geometric origin of the label noise discussed in Section 3.3, and it is the reason a learned segmentation model cannot in general outperform the resolution of the reference measurement used to train it.

2.2 Defect classes and their physical signatures

Defects in metal powder bed fusion divide into volumetric and geometric families whose signatures differ in scale and persistence. Gas and keyhole porosity appears as near-spherical voids with smooth internal surfaces; the associated top-surface signature is a shallow depression, often too small to resolve in a single layer but detectable as a persistent local depression across several consecutive layers. Lack-of-fusion defects form between adjacent melt tracks or between layers when energy density is insufficient, producing irregular planar voids whose surface expression is a dark, rough streak aligned with the scan strategy. Balling produces isolated spheres that protrude above the layer and are efficient scatterers. Delamination and layer shift appear as abrupt discontinuities between consecutive layer images registered in the build frame, which makes them among the few defect classes for which temporal comparison is a stronger cue than any single-layer appearance. Warpage is a geometric distortion accumulated over the build and is best characterized by deviation from nominal geometry rather than by texture.

The reference standard for these classes is usually X-ray computed tomography or, occasionally, high-speed synchrotron imaging that resolves melt-pool dynamics directly [13]. Texture-based descriptors applied to layer imagery offer an alternative route to the same classes without volumetric reference data, and multifractal analysis of image profiles has been shown to characterize surface signature differences associated with defect formation [12]. The choice of reference matters for the learning problem: computed tomography has its own resolution limit, beam-hardening artifacts near thick sections, and partial-volume behavior at void boundaries, so labels derived from it are noisy in exactly the size regime where detection is most valuable. Metrological traceability for additive surfaces is itself an active area, and the calibration chain from instrument to artifact to uncertainty statement is not yet standardized for the internal geometries these processes produce [11]. An inspection system trained against a reference whose own error is unquantified cannot claim an accuracy better than that reference, a constraint that should be stated explicitly rather than buried in a limitations sentence.

3. Methodological Framework

3.1 Sensing configuration and calibration

The choice of imaging modality fixes the achievable defect scale. Coaxial melt-pool monitoring records emission from the melt pool itself at high frame rates, giving excellent temporal resolution and a signal directly tied to energy input, but the field of view is narrow and the resulting data describe the process rather than the part. Layer-wise optical imaging after powder spreading records the full build area at a spatial resolution set by the camera and optics, and it is the modality for which volumetric segmentation is most natural. Fringe projection and structured light add depth to the measurement and can recover surface topography rather than intensity, at the cost of acquisition time per layer. Which modality dominates depends on the defect class of interest: a system optimized for porosity needs volumetric sensitivity, whereas one optimized for geometric distortion needs topographical accuracy, and a single sensor rarely supplies both at production rates.

Calibration has three components that are frequently conflated. Intrinsic calibration estimates focal length, principal point, and lens distortion from a target with known geometry. Extrinsic calibration estimates the pose of the camera relative to the build plate, which fixes R and t in the registration transform. Build-frame calibration is the less-discussed third component: it assigns a z index to each layer image, aligns the in-plane origin with the machine coordinate system, and must account for any translation of the recoater or build plate between acquisitions. Error in the first two produces smooth geometric misalignment; error in the third produces a systematic offset that grows through the build and corrupts exactly the temporal comparison on which delamination detection depends. Illumination normalization is part of the same chain, since the ambient term in the observation model varies with lamp aging and with powder contamination on the window, and an unnormalized drift in that term appears to a learned model as a change in material rather than a change in illumination.

3.2 Architecture and volumetric fusion

The segmentation task is dense prediction in three dimensions: for each voxel of the build-frame grid, assign a class among solid material, lack-of-fusion, porosity, spatter, and background. Encoder–decoder architectures with skip connections are the natural starting point because they combine a contracting path that aggregates context with an expanding path that restores boundary detail [6]. Extending this design to volumes requires a choice between full three-dimensional convolution and a 2.5-dimensional decomposition. Full three-dimensional kernels learn cross-layer context directly, which is essential for defects whose signature is persistence across layers, and dense volumetric architectures with anisotropic kernels were developed for precisely this reason in biomedical imaging where slice spacing differs from in-plane spacing [14]. Anisotropic kernels are the appropriate compromise here as well, because layer thickness is typically several times the in-plane pixel pitch; an isotropic kernel of fixed physical size would either span too few layers or too many pixels.

A competing design uses a two-dimensional backbone applied per layer, with temporal or cross-layer context supplied by a recurrent or attention module operating on the sequence of feature maps. This decomposition is computationally cheaper by roughly the ratio of the depth extent to the kernel depth, and it transfers readily from pretrained two-dimensional models. Its weakness is that it cannot represent a defect whose signature is a through-thickness void without an explicit mechanism for propagating information along the layer axis. Volumetric objectives that weight boundary regions more heavily, introduced in the V-Net formulation, address a related problem: in a volume dominated by healthy material, the gradient is dominated by interior voxels far from any decision boundary, and boundary-sensitive terms restore the influence of the voxels that determine the shape of the predicted defect [15]. The practical consequence is that architecture and objective must be selected together; a volumetric backbone with a purely voxelwise loss wastes much of its capacity on the easy interior.

3.3 Training objective and label imbalance

Defect voxels constitute a small minority of any build volume, and the minority is heterogeneous: lack-of-fusion regions, keyhole pores, and spatter have different sizes, shapes, and appearance statistics. Uniform voxel sampling therefore produces gradients dominated by healthy material. The standard remedy is a class-weighted cross-entropy summed over voxels and classes:

$$L_CE = -\sum_v \sum_c w_c \cdot y_{v,c} \cdot \log p_{v,c}$$

where v indexes voxels, c indexes classes, w_c is a per-class weight chosen inversely to class frequency or tuned to the cost structure of the application, y_{v,c} is the one-hot reference label, and p_{v,c} is the predicted posterior for that voxel and class. Weighting corrects the sampling bias but not the boundary problem: a small defect can be almost entirely boundary, and a loss that treats every voxel independently has no term that rewards predicting the defect as a connected region.

Region overlap terms address this. Combining weighted cross-entropy with a soft Dice term yields an objective in which the two components act on complementary failure modes:

$$L = L_CE + \lambda \cdot (1 - 2|P \cap G|/(|P| + |G|))$$

Here P is the set of voxels predicted positive with soft membership weights, G is the set of reference-positive voxels, $|P \cap G|$ is their soft intersection, $|P|$ and $|G|$ are the soft cardinalities, and λ scales the region term relative to the voxelwise term. The Dice component is scale-invariant with respect to defect size, so a small defect contributes as strongly as a large one; this is precisely why it improves small-defect recall and precisely why it can destabilize training when defects are extremely rare, since the denominator approaches zero and gradients become large. λ is therefore a stability parameter as much as a weighting parameter, and it should be reported alongside the class weights rather than omitted from the methods description.

Two further training considerations follow from the label structure. Because reference labels come from computed tomography, they carry boundary uncertainty of the order of the tomography voxel, and a model trained with hard labels learns to reproduce the tomography threshold rather than the physical void boundary. Label smoothing at the boundary or explicit modeling of the label noise distribution reduces this effect. Because defects persist across layers, labels are strongly correlated along the build direction, and random voxel sampling destroys that correlation; sampling contiguous subvolumes rather than individual voxels preserves it and matches the statistics the network will encounter at inference.

4. Uncertainty Quantification and Decision Rules

4.1 Calibration of predictive posteriors

A segmentation network outputs a posterior after a softmax operation, and there is no reason for that number to equal the frequency with which predictions at that confidence level are correct. Modern deep networks are systematically overconfident, a property that is invisible in accuracy metrics and damaging in inspection because it displaces the operating point of any downstream threshold [18]. The problem is acute in defect detection where the positive class is rare: a network that outputs 0.95 on a spurious texture is indistinguishable, in the output alone, from a network that outputs 0.95 on a genuine lack-of-fusion streak, and only calibration distinguishes them.

Epistemic uncertainty arising from limited training data can be approximated by treating dropout at inference as a variational sampling procedure, which yields a distribution over predictions rather than a point estimate [16]. Averaging over stochastic forward passes gives the predictive posterior, and a temperature parameter applied to the logits before averaging allows the sharpness of the distribution to be corrected on held-out data:

$$\bar{p}_c(v) = (1/M) \cdot \sum_m \, softmax(z_m(v)/T)_c$$

In this expression $\bar{p}_c(v)$ is the calibrated posterior for class c at voxel v, M is the number of stochastic forward passes, $z_m(v)$ is the logit vector produced by pass m, and T is a positive temperature fitted by minimizing negative log-likelihood on a calibration set that is disjoint from both training and test data. When T is greater than one the distribution is softened, which is the usual outcome. The entropy of $\bar{p}$ across classes gives a scalar uncertainty per voxel, and its spatial distribution is informative in its own right: high entropy concentrated along the part surface indicates a registration or illumination problem, whereas high entropy localized inside a suspected defect indicates genuine ambiguity about the material state.

Alternative formulations treat the network output as evidence for a Dirichlet distribution over class probabilities, which separates ignorance from conflicting evidence and yields uncertainty estimates without repeated sampling [19]. Distribution-free approaches based on conformal prediction provide a complementary guarantee, producing prediction sets that contain the true label with a user-specified probability under an exchangeability assumption [20]. The exchangeability assumption is the operative difficulty in this domain, since builds differ in geometry, feedstock, and machine state, and a calibration set drawn from one build does not obviously exchange with another. Reporting conformal guarantees without stating the conditioning set overstates what has been demonstrated. Whether the epistemic component is captured by sampling, by evidential output, or by an ensemble, the inspection requirement is the same: uncertainty must be quantified and reported, and the decision rule must consume it rather than ignore it.

4.2 Cost-sensitive acceptance

The final step converts posteriors into a part-level decision. Consider a single candidate defect with calibrated probability p of being real. Two actions are available: accept the region as sound, or reject and flag it for rework or investigation. Let C_{FP} be the cost incurred when a sound region is flagged and C_{FN} be the cost incurred when a defective region is accepted. Correct decisions incur no cost. The expected cost of accepting is $p \cdot C_{FN}$ and the expected cost of rejecting is $(1 - p) \cdot C_{FP}$, so rejection is preferred whenever the second quantity is smaller. Setting the two equal gives the cost-optimal decision threshold:

$$\tau* = C_F \frac{P}{(C_F P + C_F N)}$$

The expression is elementary but its implications are frequently ignored. The threshold is not 0.5 and is not a tuning parameter to be selected by maximizing F1 score; it is fixed by the cost ratio, and in structural applications C_{FN} greatly exceeds C_{FP}, so τ is far below 0.5. A network reporting a calibrated probability of 0.2 for a defect in a load-bearing region should trigger investigation if missing that defect would cost an order of magnitude more than an unnecessary inspection. This is the classical asymmetry of cost-sensitive learning [21], and it explains why a detector with excellent area under the curve can still be operationally useless: the curve summarizes ranking across all thresholds while the application lives at one threshold determined by costs that the learning procedure never sees.

Aggregating voxel decisions to a part decision introduces a second asymmetry. A part is defective if any of its voxels is defective, so the natural aggregation is a noisy-or over candidate regions weighted by their calibrated probabilities rather than a majority vote over voxels. Majority voting dilutes a small but real defect among thousands of sound voxels, which is exactly the failure the region-based loss in Section 3.3 was designed to prevent. The

aggregation also controls the relationship between voxel-level false positive rate and part-level false rejection rate: with many independent candidate regions per part, even a modest per-region false positive probability accumulates into a substantial probability that a sound part is flagged. Per-region calibration and per-part aggregation must therefore be reported together, since a number quoted at one level says nothing about the other.

5. Evaluation Logic and Failure Modes

5.1 Metrics and their structural biases

Voxelwise overlap metrics are the default in segmentation work and are systematically misleading in this application. Dice and intersection-over-union computed over the whole volume are dominated by the largest defects and are almost insensitive to missing a small pore, because a missed pore of a few hundred voxels changes the numerator by a negligible amount relative to the total. A detector reporting high volumetric Dice may be failing on precisely the defect sizes that matter for fatigue life. Evaluation should therefore be stratified by defect size relative to the voxel pitch, with detection recall reported per size band and per defect instance rather than per voxel. Instance-level recall, defined as the fraction of connected defect components in the reference that are matched by at least one predicted component within a distance tolerance, is closer to what an inspector needs and is far more sensitive to small-object failure.

Part-level agreement with the reference decision is the metric that connects to qualification, and it requires the cost structure from Section 4.2 to be stated. Agreement measured at $\tau = 0.5$ answers a different question than agreement at τ, and reporting the former as if it were the latter is a common source of optimistic conclusions. Calibration should be assessed separately using reliability diagrams, expected calibration error, and negative log-likelihood, all computed on data not used for temperature fitting. Volume error, the signed difference between predicted and reference defect volume per part, is useful because it exposes systematic bias that overlap metrics conceal: a detector that consistently over-segments boundaries can achieve respectable Dice while reporting defect volumes that are wrong by a large multiplicative factor.

5.2 Failure modes

Registration drift is the most consequential failure mode because it corrupts labels rather than merely predictions. A lateral misalignment of even a few voxels between the build frame used for training labels and the frame used at inference shifts every defect boundary, and the network learns a systematic spatial offset that cannot be corrected by more data of the same kind. The symptom is a model that performs well on its training machine and degrades monotonically as the build progresses, which is consistent with an accumulating z-index error rather than with a representational limitation.

Powder masking is the second. In powder bed processes the part surface is partly obscured by unfused powder, whose reflectance differs from solid material and whose distribution changes with recoater condition. The observation model in Section 2.1 treats the imaged surface as the part; when a powder layer intervenes, the measured intensity describes the powder, and any defect signature must propagate through it. This attenuates high-spatial-frequency signatures far more than low-frequency ones, biasing detection toward large defects and against the small pores that are most consequential.

Specular reflection violates the diffuse assumption directly. Near-specular surfaces produce highlights whose position depends on viewing and illumination geometry rather than on material state, and a network trained on one illumination configuration can learn to associate highlight position with defect presence if the two are correlated in the training set. This is a confounding failure rather than a capacity failure, and the remedy is data collection that deliberately decorrelates illumination geometry from defect presence.

Label noise from the reference measurement, class imbalance, and domain shift between machines complete the list. Domain shift deserves particular attention because the appearance of a defect depends on the machine's optical configuration, the feedstock batch, and the scan strategy, none of which are represented in a training set drawn from a single build. Transfer to a new machine therefore requires either a calibration procedure that normalizes appearance across configurations or a small labeled sample from the target machine, and the choice between them should be justified by how much of the appearance variation is attributable to illumination rather than to material.

6. Technical Discussion

The framework above imposes a specific ordering on design decisions, and the ordering is the substance of the contribution. Sensing modality determines the achievable defect scale; registration determines whether labels are geometrically meaningful; architecture determines whether cross-layer evidence can be used; the objective determines whether small defects influence learning; calibration determines whether posteriors can be compared to costs; and the decision rule determines what the system actually does. Errors introduced early in this chain cannot be repaired later. A segmentation network cannot recover from registration drift, a calibrated posterior cannot compensate for a biased training label, and a cost-optimal threshold applied to an uncalibrated score produces a confidently wrong decision.

Three trade-offs constrain implementation. Spatial resolution trades against acquisition throughput, since resolving smaller defects requires either a narrower field of view or a larger sensor, and production build rates set a hard limit on the number of frames that can be captured per layer. Volumetric modeling trades against memory and inference latency, and the choice between full three-dimensional convolution and 2.5-dimensional decomposition should be made by asking whether the defect classes of interest have a cross-layer signature rather than by defaulting to the larger model. Label fidelity trades against label quantity, since tomography at higher resolution is slower and more expensive, and a smaller set of high-fidelity labels may be preferable to a larger set whose boundary error exceeds the defect size being targeted.

The framework also clarifies what a validation study must report to be interpretable. It must specify the calibration chain from instrument to build frame, the resolution of the reference measurement relative to the smallest defect class claimed, the class weights and region-loss weight, the calibration procedure and the data used to fit it, the cost ratio underlying the decision threshold, and performance stratified by defect size at both the instance and part level. Absent any of these, a reported accuracy figure is not comparable to any other figure in the literature, which is a substantial part of why vision-based inspection has been slow to move from demonstration to qualification.

Boundary conditions should be stated plainly. The framework assumes a part that is rigidly fixed to the build plate for the duration of the build, an illumination configuration that is stable or normalized, and a reference measurement whose error is smaller than the defect

class of interest. Where supports, overhangs, or powder removal intervene, the mapping between layer imagery and final part geometry becomes conditional, and the deviation field defined in Section 2.1 requires modification to account for material that is intentionally removed. The framework is a specification for constructing and evaluating an inspection system; establishing that a particular instantiation meets a qualification requirement remains an empirical task that depends on the machine, the material, and the criticality of the component.

7. Conclusion

Vision-based inspection of additively manufactured structures is often framed as a segmentation problem, and that framing understates what a qualification-grade system must do. This paper has developed a framework in which deep neural networks operate inside an explicit measurement and decision chain. An observation model grounded in surface reflectance explains which physical changes a defect produces in a measured pixel and which it does not, and it identifies the conditions under which intensity-based detection is fundamentally limited. A build-frame registration step establishes the geometric consistency on which volumetric labels depend, with the deviation between measured and nominal occupancy making the label noise from partial-volume effects explicit. A volumetric encoder–decoder with anisotropic kernels and a combined weighted cross-entropy and region-overlap objective addresses the class imbalance and small-defect sensitivity that dominate this application. Calibration of the predictive posterior, whether by stochastic sampling, evidential output, or conformal methods, is treated as a precondition for decision making rather than a refinement, because the cost-optimal acceptance threshold is fixed by the ratio of false-negative to false-positive costs and not by any property of the network. Evaluation at voxel, instance, and part level, stratified by defect size and reported with the cost structure stated, closes the gap between what a model estimates and what a part is judged to be.

The practical implication is that effort allocated to architecture search without corresponding attention to registration, label provenance, and calibration is unlikely to produce an inspection system that survives transfer to a new machine or a new component. The scientific implication is that reported performance in this field is currently difficult to compare across studies because the measurement chain and the decision rule are usually left implicit. Making both explicit is a precondition for the empirical validation that must follow, and it defines the evidence a credible validation study would need to present.

References

1. Huang, Y., Leu, M. C., Mazumder, J., & Donmez, A. (2015). Additive manufacturing: Current state, future potential, gaps and needs, and recommendations. Journal of Manufacturing Science and Engineering, 137(1), 014001.

2. Everton, S. K., Hirsch, M., Stravroulakis, P., Leach, R. K., & Clare, A. T. (2016). Review of in-situ process monitoring and in-situ metrology for metal additive manufacturing. Materials & Design, 95, 431–445.

3. Grasso, M., & Colosimo, B. M. (2017). Process defects and in situ monitoring methods in metal powder bed fusion: A review. Measurement Science and Technology, 28(4), 044005.

4. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25 (pp. 1097–1105).

5. Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3431–3440).

6. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (pp. 234–241). Springer.

7. Scime, L., & Beuth, J. (2018). A multi-scale convolutional neural network for autonomous anomaly detection and classification in a laser powder bed fusion additive manufacturing process. Additive Manufacturing, 24, 273–286.

8. Gobert, C., Reutzel, E. W., Petrich, J., Nassar, A. R., & Phoha, S. (2018). Application of supervised machine learning for defect detection during metallic powder bed fusion additive manufacturing using high resolution imaging. Additive Manufacturing, 21, 517–528.

9. Imani, F., Chen, R., Diewald, E., Reutzel, E., & Yang, H. (2019). Deep learning of variant geometry in layerwise imaging profiles for additive manufacturing quality control. Journal of Manufacturing Science and Engineering, 141(11), 111001.

10. Chen, D., Luo, D., Xu, W., Luo, C., Shen, L., Yan, X., & Wang, T. (2019, September). Re-perceive 3D printing with Artificial Intelligence. In Education and research in computer aided architectural design in Europe and XXIII Iberoamerican Society of Digital Graphics, joint conference (pp. 443-450).

11. Leach, R. K., Bourell, D., Carmignato, S., Donmez, A., Senin, N., & Dewulf, W. (2019). Geometrical metrology for metal additive manufacturing. CIRP Annals, 68(2), 677–700.

12. Yao, B., Imani, F., Sakpal, A. R., Reutzel, E. W., & Yang, H. (2018). Multifractal analysis of image profiles for the characterization and detection of defects in additive manufacturing. Journal of Manufacturing Science and Engineering, 140(3), 031014.

13. Zhao, C., Parab, N. D., Li, X., Fezzaa, K., Tan, C. L., Copley, A. D., Chen, L., & Sun, T. (2017). Real-time monitoring of laser powder bed fusion process using high-speed X-ray imaging and diffraction. Scientific Reports, 7(1), 3602.

14. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016 (pp. 424–432). Springer.

15. Milletari, F., Navab, N., & Ahmadi, S. A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In 2016 Fourth International Conference on 3D Vision (pp. 565–571). IEEE.

16. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (pp. 1050–1059).

17. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems 30 (pp. 5574–5584).

18. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 1321–1330).

19. Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. In Advances in Neural Information Processing Systems 31 (pp. 3179–3189).

20. Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511.

21. Elkan, C. (2001). The foundations of cost-sensitive learning. In Proceedings of the 17th International Joint Conference on Artificial Intelligence (pp. 973–978).