CONTEXT DETERMINES OPTIMAL ARCHITECTURE IN MATERIALS SEGMENTATION

Mingjian Lu[1] **Pawan K. Tripathi**[1] **Mark Shteyn**[1] **Debargha Ganguly**[1]
Roger H. French[1] **Vipin Chaudhary**[1] **Yinghui Wu**[1]
[1]Case Western Reserve University, Cleveland, Ohio, USA
{mxl1171, pawan.tripathi, mis60, debargha,}@case.edu
{roger.french, vipin, yxw1650}@case.edu
https://github.com/cwru-sdle/materials-data-segmentation-benchmark

ABSTRACT

Segmentation architectures are typically benchmarked on single imaging modalities, obscuring deployment-relevant performance variations: an architecture optimal for one modality may underperform on another. We present a cross-modal evaluation framework for materials image segmentation spanning SEM, AFM, XCT, and optical microscopy. Our evaluation of six encoder-decoder combinations across seven datasets reveals that optimal architectures vary systematically by context: UNet excels for high-contrast 2D imaging while DeepLabv3+ is preferred for the hardest cases. The framework also provides deployment feedback via out-of-distribution detection and counterfactual explanations that reveal which microstructural features drive predictions. Together, the architecture guidance, reliability signals, and interpretability tools address a practical gap in materials characterization, where researchers lack tools to select architectures for their specific imaging setup or assess when models can be trusted on new samples.

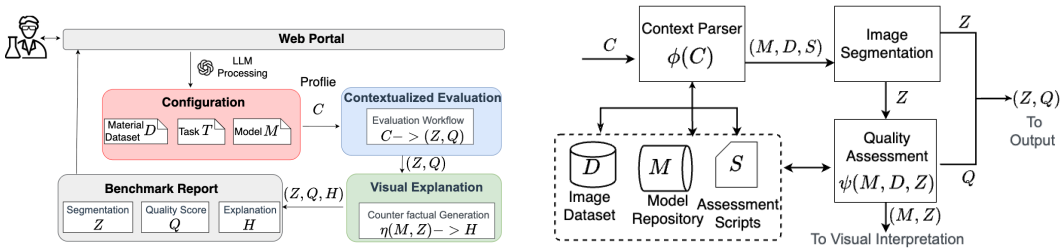
1 INTRODUCTION

Quantitative microstructural analysis underpins materials discovery across manufacturing, energy, and aerospace sectors (Holm et al., 2020; Bostanabad et al., 2018; Choudhary et al., 2022). Deep learning has enabled automated segmentation for tasks ranging from pore detection in additive manufacturing (Stan et al., 2020) to defect identification in electron microscopy (Roberts et al., 2019; Azimi et al., 2018). However, architectures that excel on one imaging modality often underperform on another: in our benchmarks, best-performing models exhibit IoU $\in [0.47, 0.95]$, a 48 percentage-point spread that persists despite identical training protocols. This variation indicates that inductive biases interact non-trivially with modality-specific statistics. Relative model rankings that depend on evaluation domain violate the implicit independent and identically distributed (i.i.d.) assumption underlying standard benchmarks, which report aggregate performance rather than deployment guidance for specific modalities and material systems (Ma et al., 2020; Pokharel et al., 2024).

As characterization pipelines become automated (Hernandez et al., 2024; Ciardi et al., 2024), this gap creates deployment risks. Segmentation models exhibit high-confidence errors under covariate shift, producing semantically plausible but incorrect masks on out-of-distribution samples (due to batch variation, preparation differences, or imaging drift), and these errors propagate undetected into downstream analyses: flawed porosity fractions, grain size distributions, or corrosion rates (Kalutotage et al., 2025). Addressing this requires context-conditional evaluation that provides both architecture guidance and reliability signals for deployment.

We present a configurable evaluation framework addressing this gap through three capabilities:

1. **Cross-modal architecture comparison.** Systematic evaluation testing whether architecture rankings are stable across imaging modalities, or whether optimal inductive biases are context-dependent, using six encoder-decoder combinations: two encoders (ResNet50 (He et al., 2016), SE-ResNeXt101 (Hu et al., 2018)) and three decoders (UNet (Ronneberger et al., 2015), DeepLabv3 (Chen et al., 2017), DeepLabv3+ (Chen et al., 2018)).



(a) Framework overview: three modules for cross-modal configuration, quality feedback, and expert knowledge integration.

(b) Evaluation workflow: configuration-driven data/model selection, segmentation, quality assessment, and report generation.

Figure 1: (a) The three-component evaluation framework maps heterogeneous inputs to standardized profiles, produces segmentation masks with reliability signals, and generates interpretability heatmaps. (b) The quality-controlled evaluation workflow executes this pipeline.

2. **Cross-modal architecture comparison.** Systematic evaluation validating how architectural inductive biases map to specific materials imaging constraints (e.g., surface vs. volumetric), using six encoder-decoder combinations: two encoders (ResNet50 (He et al., 2016), SE-ResNeXt101 (Hu et al., 2018)) and three decoders (UNet (Ronneberger et al., 2015), DeepLabv3 (Chen et al., 2017), DeepLabv3+ (Chen et al., 2018)).
3. **Deployment quality control.** Out-of-distribution detection (Ganguly et al., 2025b) that identifies distribution shift before it manifests as segmentation error, converting silent failures into actionable warnings for scientific workflows (Lu et al., 2023; Meshnick et al., 2023).
4. **Expert-interpretable explanations.** Counterfactual analysis (Vermeire et al., 2022) that identifies the minimal image modifications that change predictions, revealing which microstructural features are decision-critical for validation against domain knowledge (Modarres et al., 2017).

Evaluating six encoder-decoder combinations across seven datasets, we find systematic context-dependence: decoder architecture interacts with task difficulty, suggesting that architectural inductive biases must be matched to image characteristics. UNet dominates high-contrast 2D tasks (SEM, AFM, Optical), while DeepLabv3+ is preferred for the hardest cases (fractography SEM and XCT stress-corrosion) where multi-scale context is critical.

2 METHODOLOGY

To address context-dependent architecture selection, we introduce a three-component evaluation framework for materials image segmentation across heterogeneous modalities (Figure 1a). At a high level, the framework takes dataset specifications, task descriptions, and model choices as input, and produces segmentation masks, reliability signals indicating potential distribution shift, and interpretability heatmaps for expert validation.

More formally, the *Cross-Modal Configuration Module* standardizes heterogeneous input specifications (dataset $\mathcal{D}$, task T, model M) into a configuration C within configuration space $\mathcal{C}$, ensuring consistent evaluation semantics across modalities. The *Quality Feedback Module* takes C to produce segmentation masks Z and reliability signals Q, while the *Expert Knowledge Integration Module* operates on (M, Z) to generate interpretability heatmaps H. These outputs (Z, Q, H) are summarized in a structured feedback report.

These components induce a feedback process wherein reliability signals Q expose distribution shift, prompting configuration revision, while explanations H enable expert validation of learned decision boundaries against domain priors. The loop terminates when Q indicates acceptable calibration and H aligns with domain expectations.

The evaluation pipeline thus factors into three stages: configuration C determines (D, M); inference and quality assessment yield (Z, Q); and the interpreter produces H conditioned on expert-specified regions. The resulting tuple (Z, Q, H) forms a complete evaluation record enabling reproducible, context-specific analysis.

Table 1: Configuration profile schema defining the evaluation space $\mathcal{C}$ for cross-modal benchmarking.

Category	Property	Values	Description
Dataset	Modality	SEM, AFM, XCT, Optical	Resolution: sub-nm to μm (Holm et al., 2020)
	Material	Metals, Polymers, Semicond.	Microstructural features (Choudhary et al., 2022)
	Metadata	Ontological descriptors	Provenance tracking (Rajamohan et al., 2025)
Task	Objective	"Identify grain boundaries"	Natural language specification
	Metrics	IoU, Dice, pixel accuracy	Quality measures (Cho, 2024)
	QC signals	AUROC, F_1, precision	OOD indicators (Ganguly et al., 2025b)
Model	Encoders	ResNet50, SE-ResNeXt101	Backbone (He et al., 2016; Hu et al., 2018)
	Decoders	UNet, DeepLabv3(+)	Seg. head (Ronneberger et al., 2015; Chen et al., 2018)

Cross-Modal Configuration Module. The configuration module implements π by aggregating dataset specifications $\mathcal{D}$ with feature representations D_F and task descriptors T, producing a configuration profile $C \in \mathcal{C}$ that serves as the canonical input to downstream evaluation. Profiles are standardized JSON documents (Table 1) specifying three categories:

Dataset Configuration. Specifies properties of datasets D, including (1) imaging modalities (e.g., SEM, AFM, XCT, Optical), (2) material systems (e.g., metals, polymers, semiconductors), and (3) auxiliary metadata and provenance such as materials ontologies and experiment descriptors (Rajamohan et al., 2025).

Task Configuration. Contains (1) natural language statements of a task description T that describes segmentation goals (e.g., "Identify crystal impingement sites"), and (2) performance measures (Cho, 2024) and scripts for assessing segmentation quality and model performance.

Model Specification. Declares the segmentation model $M = F_2 \circ F_1$ as the composition of an encoder F_1 and decoder F_2. In this study, encoders include ResNet50 (He et al., 2016) and SE-ResNeXt101 (Hu et al., 2018; Xie et al., 2017), while decoders include UNet (Ronneberger et al., 2015), DeepLabv3 (Chen et al., 2017), and DeepLabv3+ (Chen et al., 2018).

The configuration profile C specifies the Cartesian product of candidate encoders and decoders for systematic evaluation. While C can be authored directly as structured input (JSON), future extensions could compile natural language specifications into C via NLP pipelines, further supporting reproducibility across material-imaging-task contexts.

Quality-Controlled Evaluation with Feedback Signals. Given configuration C, this module executes the evaluation workflow (Figure 1b) via two components: **(1) Query parser** $\phi : \mathcal{C} \to \mathcal{Q} \times \mathcal{J}$: A deterministic compilation function that maps configuration C to an executable query set $\mathcal{Q}$ and job specification $\mathcal{J}$, which (i) loads the dataset D and its representation D_F, (ii) retrieves encoder F_1 and decoder F_2 specifications and assembles the segmentation model $M = F_2 \circ F_1$, and (iii) generates training and evaluation scripts. **(2) Performance evaluator** $\psi : M \times \mathcal{D} \times \mathcal{C} \to (Z, Q)$: Computes the segmentation map $Z = M(D)$ and derives quality feedback $Q = (q_{\text{AUROC}}, q_{F_1}, q_{\text{prec}})$ via the OOD detection procedure specified in C (Ganguly et al., 2025b), which utilizes self-supervised embeddings and typicality estimation to flag outliers, where Q quantifies model calibration under potential covariate shift and indicates when retraining or domain adaptation may be warranted.

Expert Knowledge Integration through Counterfactual Feedback. Deployment in scientific workflows requires attribution of predictions to input features, enabling validation that learned representations encode domain-relevant structure. Our framework generates counterfactual explanations focused on designated areas of interest via the interpreter function $\eta : M \times Z \to H$. Given an input image D, its prediction $Z = M(D)$, and a user annotation mask A marking the region of interest, we construct the perturbed input $D^{(i,j)} = D + \delta_{ij}$ for each spatial location (i, j), where δ_{ij} is a localized perturbation kernel, and compute the counterfactual prediction $Z^{(i,j)} = M(D^{(i,j)})$. The influence heatmap H is computed as

$$H_{i,j} = \left\| (Z \odot A) - (Z^{(i,j)} \odot A) \right\|_1 ,$$

where $H_{i,j}$ measures the ℓ_1 sensitivity of predictions within A to perturbations at location (i, j). This formulation restricts attribution to expert-specified structures, yielding a spatially-localized sensitivity map that identifies which input regions exert maximal influence on predictions within the annotated area.

3 Experiments and Case Analysis

Experimental Setup. We conduct a systematic evaluation across encoder–decoder combinations to assess performance across materials research scenarios and tasks. All segmentation models M use ImageNet-pretrained encoders and identical training settings: batch size 16, learning rate 10^{-4} with cosine annealing, and early stopping with patience 20. Code is available at (Lu et al., 2025) for reproducibility.

Configurations. We adopt the configuration schema and example values in Table 1. Dataset details and additional settings to aid reproducibility are reported in (Lu et al., 2025).

Experimental Results. We report our findings below.

Table 2: Cross-modal segmentation performance. SE-ResNeXt101 is the optimal encoder across all configurations; we report only the best-performing decoder. Difficulty reflects intrinsic task complexity rather than achieved IoU. (Full results are in Table 3.)

Modality	Dataset	Material	Task	Difficulty	IoU	Dice
AFM	Crystallites (Lu et al., 2023)	Fluoroelastomer	Spherulite/Lamellae	Moderate	**0.901**	0.961
SEM	Contact Corr.	Si Solar Cell	Corroded/Intact	Easy	**0.951**	0.962
	Fractography†	Ti-6Al-4V	Defects	Hard	**0.636**	0.490
	Carbon SEM (Pokharel et al., 2024)	Carbon	Carbon regions	Moderate	**0.762**	0.621
XCT	Al Pitting (Kalutotage et al., 2025)	Al Wire	Pitting corrosion	V. Hard	**0.752**	0.570
	Al-Mg-Si SC† (Ciardi et al., 2024)	Al-Mg-Si	Stress corrosion	V. Hard	**0.472**	0.550
Optical	L-PBF (Hernandez et al., 2024)	316L Steel	Melt pool/Spatter	Hard	**0.652**	0.628

†DeepLabv3+ decoder; all others use UNet.

Exp-1: Cross-Modal Architecture Recommendations. Table 2 reports IoU and Dice scores across seven datasets spanning four imaging modalities. Our evaluation reveals three natural use-case categories, each sharing common data characteristics (high-contrast boundaries, temporal dynamics, or multi-scale volumetric structure) that benefit from consistent architecture choices. We summarize the findings below.

Use Case 1: High-Contrast Surface Characterization. For datasets with sharp boundaries and high contrast (AFM, surface SEM), SE-ResNeXt101 paired with UNet achieves IoU above 0.90, including 0.951 for semiconductor contact corrosion and 0.901 for polymer crystallite analysis. UNet's skip connections preserve fine spatial gradients critical for accurate boundary delineation, while SE-ResNeXt101's channel attention amplifies contrast at material interfaces.

Use Case 2: Dynamic Process Monitoring. For real-time characterization tasks (L-PBF additive manufacturing, Carbon SEM), the same SE-ResNeXt101 + UNet combination yields IoU between 0.65–0.76. The lower performance vs. Use Case 1 reflects challenges like temporal variation in melt-pool morphology rather than architectural mismatch. The attention mechanism in SE-ResNeXt101 helps emphasize transient, high-contrast regions, while UNet's lower latency better suits high-throughput in-situ monitoring.

Use Case 3: Multi-Scale Volumetric Analysis. For 3D volumetric tasks requiring multi-scale reasoning (XCT corrosion, fractography), switching to DeepLabv3+ as the decoder improves performance, achieving IoU of 0.47–0.75. This includes challenging cases like stress corrosion cracking (IoU 0.472), pitting corrosion (IoU 0.752), and fractography defects (IoU 0.636). DeepLabv3+'s atrous spatial pyramid pooling captures defect patterns across multiple scales, and its boundary refinement module better handles the irregular morphologies of sub-surface crack networks.

Summary. Our contextualized evaluation reveals three key findings that static, aggregate benchmarks would obscure:

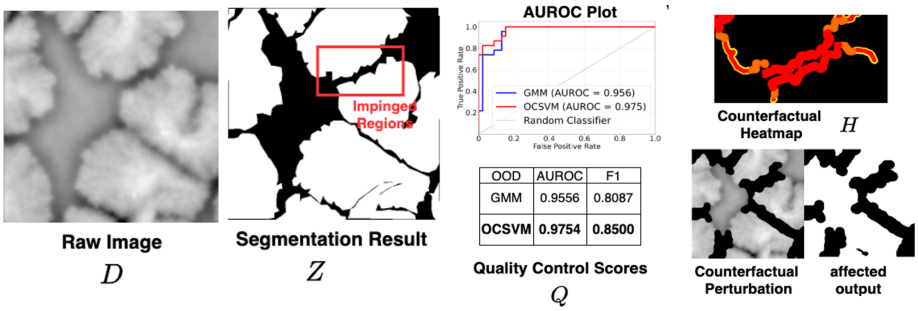


Figure 2: Complete feedback loop demonstration for AFM crystallite analysis: (a) Multi-channel input D, (b) Segmentation output Z, (c) Quality feedback signals Q (AUROC, F1) indicating deployment readiness, (d) Interpretable feedback H enabling expert validation of decision-relevant regions.

1. **No single architecture dominates.** While SE-ResNeXt101 consistently outperforms ResNet50 as an encoder, optimal decoder choice depends on task characteristics: UNet excels for high-contrast surface tasks, while DeepLabv3+ is preferred for volumetric analysis.
2. **Performance is context-sensitive.** The same architecture (SE-ResNeXt101 + UNet) achieves IoU of 0.901 on crystallite analysis but only 0.652 for manufacturing monitoring, a 25 percentage-point spread that would be masked by aggregate reporting.
3. **Task difficulty guides decoder selection.** Easy and moderate tasks favor UNet; hard and very hard tasks, especially those involving complex morphologies or volumetric defects, benefit from DeepLabv3+'s multi-scale context.

Exp-2: End-to-End Feedback Loop Demonstration. We illustrate a complete feedback cycle for AFM crystallite analysis, from task specification to expert validation:

Step 1 (Configuration): The user specifies the task ("Identify crystal impingement sites in polycrystalline fluoroelastomer"), yielding a configuration for AFM multi-channel data (Figure 2a).

Step 2 (Model Evaluation): The framework evaluates all encoder-decoder combinations. SE-ResNeXt101 + UNet achieves the best performance with IoU = 0.901 and Dice = 0.961 (Figure 2b).

Step 3 (Quality Assessment): The OOD detector produces deployment-readiness scores: AUROC > 0.95 and F1 > 0.80 under both GMM and OCSVM estimators (Figure 2c). Since AUROC exceeds the 0.95 threshold, the model is flagged as ready for deployment.

Step 4 (Report Generation): The framework consolidates segmentation masks, quality scores, and interpretability heatmaps into a structured evaluation report (Figure 2a–d).

This walkthrough demonstrates how the framework enables an iterative workflow: from task specification to architecture selection to quality assessment to expert validation, supporting rapid iteration for accelerated materials characterization.

Exp-3: Expert Knowledge Integration. The interpretability heatmaps (Figure 2d) confirm that the model focuses on scientifically meaningful features rather than spurious correlations. High-influence regions (red/orange) correspond precisely to crystal impingement sites, where crystallites meet and form boundaries, matching expert annotations in the original image (Figure 2a) and boundary regions in the segmentation output (Figure 2b).

Perturbation analysis validates this interpretation: blurring the high-influence regions significantly degrades predictions, while perturbing background regions has minimal effect. This confirms that the model encodes sensitivity to topographic gradients and phase discontinuities at grain boundaries, the physically meaningful features for crystallographic analysis, rather than dataset-specific artifacts. Such validation provides confidence that the chosen architecture generalizes appropriately for boundary characterization tasks.

4 CONCLUSIONS

We present a cross-modal evaluation framework for materials image segmentation that provides architecture guidance, deployment-readiness signals, and interpretable explanations. Our evaluation across seven datasets reveals that optimal architectures vary systematically by context: UNet excels for high-contrast surface imaging while DeepLabv3+ is preferred for volumetric analysis. By combining quality control with expert-interpretable feedback, the framework supports reliable deployment in scientific workflows where silent failures carry real cost.

ACKNOWLEDGMENTS

This work is supported by the Department of Energy's National Nuclear Security Administration under Award Number DE-NA0004104 (MDS^3-COE), and NSF Grant 2117439.We extend our sincere thanks to Christine Orme (Lawrence Livermore National Laboratory, LLNL) for providing the AFM dataset; Jean-Baptiste Forien, Brian Giera, and Ibo Matthews (LLNL) for the L-PBF HSC dataset; John Lewandowski (CWRU) for the Fractography and Al-Mg-Si SC XCT datasets; Max Liggett and Kris Davis (University of Central Florida); and Philip Noell (Sandia National Laboratories) for the Al Pitting XCT dataset.

REFERENCES

Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

Seyed Majid Azimi, Dominik Britz, Michael Engstler, Mario Fritz, and Frank Mücklich. Advanced steel microstructural classification by deep learning methods. *Scientific reports*, 8(1):2128, 2018.

Ramin Bostanabad, Yichi Zhang, Xiaolin Li, Tucker Kearney, L Catherine Brinson, Daniel W Apley, Wing Kam Liu, and Wei Chen. Computational microstructure characterization and reconstruction: Review of the state-of-the-art techniques. *Progress in Materials Science*, 95:1–41, 2018.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

Weicong Chen, Vikash Singh, Zahra Rahmani, Debargha Ganguly, Mohsen Hariri, and Vipin Chaudhary. K^4: Online log anomaly detection via unsupervised typicality learning. *arXiv preprint arXiv:2507.20051*, 2025.

Yeong-Jun Cho. Weighted intersection over union (wiou) for evaluating image segmentation. *Pattern Recognition Letters*, 185:101–107, 2024.

Kamal Choudhary, Brian DeCost, Chi Chen, Anubhav Jain, Francesca Tavazza, Ryan Cohn, Cheol Woo Park, Alok Choudhary, Ankit Agrawal, Simon JL Billinge, et al. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*, 8(1):59, 2022.

Thomas G. Ciardi, Arafath Nihar, Rounak Chawla, Olatunde Akanbi, Pawan K. Tripathi, Yinghui Wu, Vipin Chaudhary, and Roger H. French. Materials data science using CRADLE: A distributed, data-centric approach. *MRS Communications*, 14(4):601–611, August 2024. ISSN 2159-6867. doi: 10.1557/s43579-024-00616-6.

Taylor Denouden, Rick Salay, Krzysztof Czarnecki, Vahdat Abdelzad, Buu Phan, and Sachin Vernekar. Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. *arXiv preprint arXiv:1812.02765*, 2018.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Debargha Ganguly and Debayan Gupta. Machine learning explainability from an information-theoretic perspective. In *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems*, 2022. URL https://openreview.net/forum?id=SqTLQ5LjQWp.

Debargha Ganguly, Debayan Gupta, and Vipin Chaudhary. Visual concept networks: A graph-based approach to detecting anomalous data in deep neural networks. In *International Conference on Pattern Recognition and Artificial Intelligence*, pp. 219–232. Springer, 2024a.

Debargha Ganguly, Srinivasan Iyengar, Vipin Chaudhary, and Shivkumar Kalyanaraman. PROOF OF THOUGHT : Neurosymbolic program synthesis allows robust and interpretable reasoning. In *The First Workshop on System-2 Reasoning at Scale, NeurIPS'24*, 2024b. URL https://openreview.net/forum?id=Pxx3r14j3U.

Debargha Ganguly, Sumit Kumar, Ishwar Balappanawar, Weicong Chen, Shashank Kambhatla, Srinivasan Iyengar, Shivkumar Kalyanaraman, Ponnurangam Kumaraguru, and Vipin Chaudhary. Labeling copilot: A deep research agent for automated data curation in computer vision. *arXiv preprint arXiv:2509.22631*, 2025a.

Debargha Ganguly, Warren Richard Morningstar, Andrew Seohwan Yu, and Vipin Chaudhary. Forte : Finding outliers with representation typicality estimation. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=7XNgVPxCiA.

Debargha Ganguly, Vikash Singh, Sreehari Sankar, Biyao Zhang, Xuecen Zhang, Srinivasan Iyengar, Xiaotian Han, Amit Sharma, Shivkumar Kalyanaraman, and Vipin Chaudhary. Grammars of formal uncertainty: When to trust LLMs in automated reasoning tasks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025c. URL https://openreview.net/forum?id=QfKpJ00t2L.

Debargha Ganguly, Sreehari Sankar, Biyao Zhang, Vikash Singh, Kanan Gupta, Harshini Kavuru, Alan Luo, Weicong Chen, Warren Richard Morningstar, Raghu Machiraju, and Vipin Chaudhary. Trust the typical. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=vfbeleLBWv.

Mark S Graham, Walter HL Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and Jorge Cardoso. Denoising diffusion models for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2948–2957, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Kristen J. Hernandez, Thomas G. Ciardi, Rachel Yamamoto, Mingjian Lu, Arafath Nihar, Jayvic Cristian Jimenez, Pawan K. Tripathi, Brian Giera, Jean-Baptiste Forien, John J. Lewandowski, Roger H. French, and Laura S. Bruckman. L-PBF High-Throughput Data Pipeline Approach for Multi-modal Integration. *Integrating Materials and Manufacturing Innovation*, July 2024. ISSN 2193-9772. doi: 10.1007/s40192-024-00368-0.

Elizabeth A Holm, Ryan Cohn, Nan Gao, Andrew R Kitahara, Thomas P Matson, Bo Lei, and Srujana Rao Yarasi. Overview: Computer vision and machine learning for microstructural characterization and analysis. *Metallurgical and Materials Transactions A*, 51:5985–5999, 2020.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Maliesha S. Kalutotage, Thomas G. Ciardi, Pawan K. Tripathi, Liangyi Huang, Jayvic Cristian Jimenez, Philip J. Noell, Laura S. Bruckman, Roger H. French, and Alp Sehirlioglu. Automated Image Segmentation and Processing Pipeline Applied to X-Ray Computed Tomography Studies of Pitting Corrosion in Aluminum Wires. *Advanced Engineering Materials*, n/a(n/a):2401699, 2025. ISSN 1527-2648. doi: 10.1002/adem.202401699.

Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. Why normalizing flows fail to detect out-of-distribution data. *Advances in neural information processing systems*, 33:20578–20589, 2020.

Mingjian Lu, Sameera Nalin Venkat, Jube Augustino, David Meshnick, Jayvic Cristian Jimenez, Pawan K Tripathi, Arafath Nihar, Christine A Orme, Roger H French, Laura S Bruckman, et al. Image processing pipeline for fluoroelastomer crystallite detection in atomic force microscopy images. *Integrating Materials and Manufacturing Innovation*, 12(4):371–385, 2023.

Mingjian Lu, Pawan K. Tripathi, et al. Materials data segmentation benchmark. https://github.com/cwru-sdle/materials-data-segmentation-benchmark, 2025. Accessed: June 2025.

Wufei Ma, Elizabeth J Kautz, Arun Baskaran, Aritra Chowdhury, Vineet Joshi, Bulent Yener, and Daniel J Lewis. Image-driven discriminative and generative machine learning algorithms for establishing microstructure–processing relationships. *Journal of Applied Physics*, 128(13), 2020.

David C. Meshnick, Nahal Shahini, Debargha Ganguly, Yinghui Wu, Roger H. French, and Vipin Chaudhary. Enhancing scientific image classification through multimodal learning: Insights from chest x-ray and atomic force microscopy datasets. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 2211–2220, 2023. doi: 10.1109/BigData59044.2023.10386478.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Ieee, 2016.

Mohammad Hadi Modarres, Rossella Aversa, Stefano Cozzini, Regina Ciancio, Angelo Leto, and Giuseppe Piero Brandino. Neural network for nanoscience scanning electron microscope image recognition. *Scientific reports*, 7(1):13282, 2017.

Warren Morningstar, Cusuh Ham, Andrew Gallagher, Balaji Lakshminarayanan, Alex Alemi, and Joshua Dillon. Density of states estimation for out of distribution detection. In *International Conference on Artificial Intelligence and Statistics*, pp. 3232–3240. PMLR, 2021.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023.

Bishwas Pokharel, Deep Shankar Pandey, Anjuli Sapkota, Bhimraj Yadav, Vasanta Gurung, Mandira Pradhananga Adhikari, Loknath Regmi, and Nanda B Adhikari. A comparative study of state-of-the-art deep learning models for semantic segmentation of pores in scanning electron microscope images of activated carbon. *IEEE Access*, 12(50217-50243):5, 2024.

Balashanmuga Priyan Rajamohan, Alexander C Harding Bradley, Van D Tran, Jonathan E Gordon, Hayden W Caldwell, Redad Mehdi, Gabriel Ponon, Quynh D Tran, Ozan Dernek, Jarod Kaltenbaugh, et al. Materials data science ontology (mds-onto): Unifying domain knowledge in materials and applied data science. *Scientific Data*, 12(1):628, 2025.

Graham Roberts, Simon Y Haile, Rajat Sainju, Danny J Edwards, Brian Hutchinson, and Yuanyuan Zhu. Deep learning for semantic segmentation of defects in advanced stem images of steels. *Scientific reports*, 9(1):12744, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.

Tiberiu Stan, Zachary T Thompson, and Peter W Voorhees. Optimizing convolutional neural networks to perform semantic segmentation on large materials imaging datasets: X-ray tomography and serial sectioning. *Materials Characterization*, 160:110119, 2020.

Tom Vermeire, Dieter Brughmans, Sofie Goethals, Raphael Mazzine Barbossa De Oliveira, and David Martens. Explainable image classification with evidence counterfactual. *Pattern Analysis and Applications*, 25(2):315–335, 2022.

Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022.

A APPENDIX

A.1 ADDITIONAL EXPERIMENT DETAILS TO AID REPRODUCIBILITY

Details of Datasets. Our framework is evaluated on seven diverse materials datasets spanning multiple imaging modalities, material systems, and segmentation challenges. Each dataset represents real-world materials characterization scenarios with expert-validated ground truth annotations.

Imaging Modalities: We cover four primary characterization techniques: (1) *Scanning Electron Microscopy (SEM)* providing high-resolution surface and cross-sectional analysis with submicron scale features; (2) *Atomic Force Microscopy (AFM)* offering multi-channel topographical, phase, and amplitude information at 512×512 resolution; (3) *X-ray Computed Tomography (XCT)* enabling 3D volumetric analysis of internal structures at 1–2μm/voxel resolution; and (4) *Optical microscopy* for dynamic process monitoring at high temporal resolution ($\sim$1kHz).

Material Systems: The datasets encompass diverse material classes critical to various applications: metals including 316L stainless steel for additive manufacturing, Ti-6Al-4V for aerospace applications, and aluminum alloys for structural components; polymers such as fluoroelastomers (Kel-F, FK800) for chemical resistance applications; semiconductors including silicon solar cells for photovoltaic systems; and carbon-based materials for advanced applications.

Segmentation Challenges: Tasks span the full spectrum of materials characterization needs: (1) *Microstructure analysis* including crystalline spherulites and lamellar structures in polymers; (2) *Defect detection* encompassing manufacturing defects in fractography, keyhole porosity, and lack-of-fusion defects; (3) *Corrosion assessment* covering contact corrosion in solar cells, pitting corrosion in aluminum, and stress corrosion cracking; and (4) *Dynamic phenomena* such as melt pool and spatter detection in laser powder bed fusion processes.

Difficulty Classification: Each dataset is classified into five difficulty levels based on segmentation complexity: *Easy* cases feature high contrast and clear boundaries (e.g., contact corrosion in SEM); *Moderate* cases involve well-defined structures with some texture complexity (e.g., AFM crystallites); *Hard* cases require distinguishing subtle features or handling class imbalance (e.g., fractography defects); and *Very Hard* cases involve complex 3D structures with low contrast and significant noise (e.g., stress corrosion in XCT). This classification enables systematic evaluation of architecture performance across varying complexity levels.

Evaluation metrics. Segmentation performance is evaluated using Intersection over Union (IoU) (Everingham et al., 2010), Dice coefficient (Milletari et al., 2016), and a weighted variant of IoU to account for class imbalance and structural heterogeneity across datasets.

Let D denote an input image, M a segmentation model, and $Z = M(D)$ the predicted segmentation map. Let Z^* denote the corresponding ground-truth segmentation.

The standard Intersection over Union (IoU) is defined as

$$\mathrm{IoU} = \frac{|Z \cap Z^*|}{|Z \cup Z^*|}. \tag{1}$$

The Dice coefficient is computed as

$$\mathrm{Dice} = \frac{2|Z \cap Z^*|}{|Z| + |Z^*|}, \tag{2}$$

which emphasizes overlap consistency and is commonly used alongside IoU in materials image segmentation.

To mitigate the dominance of large homogeneous regions and better reflect segmentation quality on small or critical structures, we additionally consider the weighted Intersection over Union (wIoU) (Cho, 2024). Let Z_i and Z_i^* denote the predicted and ground-truth labels at spatial index i, and let w_i be a non-negative weight. The weighted IoU is defined as

$$\mathrm{wIoU} = \frac{\sum_i w_i \, |Z_i \cap Z_i^*|}{\sum_i w_i \, |Z_i \cup Z_i^*|}. \tag{3}$$

In our experiments, weights w_i are derived from class-frequency statistics to emphasize underrepresented structures, ensuring that performance comparisons remain robust across datasets with varying class imbalance and segmentation difficulty.

Full encoder–decoder evaluation results. To ensure transparency in architecture selection and to support the context-dependent conclusions reported in the main text, Table 3 reports the complete segmentation performance for all evaluated encoder–decoder combinations across datasets. These results constitute the full empirical basis from which the best-performing configurations summarized in Table 2 are selected.

A.2 Out-of-Distribution for Quality Control

Why OOD? Out-of-Distribution detection is a non-negotiable problem towards safe and deployable machine learning. It provides the first line of defense, preventing silent failures in deployed machine learning systems. It bounds AI capabilities by recognizing model knowledge limitations via training data, and allows safe fallback, enabling human oversight when needed. The term OOD detection first emerged in late-2016 with the seminal baseline work by Hendrycks & Gimpel (2016), which established that well-trained neural networks tend to assign higher softmax scores to in-distribution examples.

Density-based methods for OOD detection model the distribution of in-distribution data and flag low-density regions as OOD. These include approaches using variational autoencoders (Denouden et al., 2018), normalizing flows (Kirichenko et al., 2020), and more recently, diffusion models (Graham et al., 2023). Morningstar et al. (2021) provide analysis of when density estimation succeeds and fails for OOD detection. Compared to prior work in OOD literature (Yang et al., 2022; Morningstar et al., 2021; Ganguly et al., 2024a) Forte (Ganguly et al., 2025b) presents distinct advantages.

Why Forte? Forte was chosen because it utilizes self-supervised representations to capture semantic features, built with a plug-and-play design - it works with any feature extractor, including intermediate layer features from material science specific neural networks. The per-point summary statistics incorporate elements of manifold estimation to account for local topology. This minimizes deployment overhead, eliminates additional model training requirements and provides zero-shot performance. Moreover, Forte requires no class labels, no exposure to OOD data during training and no restrictions to the architecture of predictive or generative models, with strong domain generalizations shown to synthetically generated images (Ganguly et al., 2025a), MRI images (Ganguly et al., 2025b), HPC log anomaly detection (Chen et al., 2025) and Large Language Model Guard-railing including jailbreak and toxicity detection(Ganguly et al., 2026). This is enabled because the problem of out-of-distribution detection is very similar to algorithmic challenges such as anomaly detection, open-set recognition, novelty detection. Forte's underlying principle of typicality estimation also finds uses in syntactic typicality of neurosymbolic program synthesis (Ganguly et al., 2024b), as a proxy of uncertainty quantification (UQ)(Ganguly et al., 2025c).

AFM crystallites case study. We emphasize that quality control is not evaluated as a systematic or cross-domain study in this work. Instead, we provide a focused case study on the AFM crystallites dataset to illustrate how out-of-distribution (OOD) detection can be used to surface low-quality segmentation results within a segmentation benchmarking pipeline.

In this setting, OOD samples correspond to segmentation outputs that deviate from the typical boundary morphology of crystalline spherulites. Such deviations most commonly arise from incomplete, fragmented, or distorted crystallite boundaries, which are difficult to reliably detect using aggregate segmentation metrics alone. While these predictions may still achieve moderately high IoU scores, they represent semantically poor segmentations from a materials characterization perspective.

Forte assigns low typicality scores to these atypical outputs by comparing their latent representations against the distribution learned from in-distribution AFM samples. As a result, segmentation failures associated with boundary inaccuracies are effectively flagged despite not being extreme outliers under standard performance metrics. Representative OOD failure cases identified by this process are shown in Figure 3.

Table 3: Complete encoder–decoder segmentation results across all datasets. All evaluated configurations are reported. For each dataset, the best-performing configuration (based on IoU) is highlighted in bold.

Dataset	Encoder	Decoder	IoU	Dice
L-PBF	ResNet50	UNet	0.578	0.551
	ResNet50	DeepLabv3	0.342	0.387
	ResNet50	DeepLabv3+	0.471	0.597
	SE-ResNeXt101	UNet	**0.652**	**0.628**
	SE-ResNeXt101	DeepLabv3	0.353	0.395
	SE-ResNeXt101	DeepLabv3+	0.494	0.612
Fractography	ResNet50	UNet	0.593	0.516
	ResNet50	DeepLabv3	0.568	0.487
	ResNet50	DeepLabv3+	0.576	0.501
	SE-ResNeXt101	UNet	0.627	0.548
	SE-ResNeXt101	DeepLabv3	0.636	0.450
	SE-ResNeXt101	DeepLabv3+	**0.636**	**0.490**
AFM Crystallites	ResNet50	UNet	0.897	0.902
	ResNet50	DeepLabv3	0.808	0.843
	ResNet50	DeepLabv3+	0.894	0.937
	SE-ResNeXt101	UNet	**0.901**	**0.961**
	SE-ResNeXt101	DeepLabv3	0.892	0.959
	SE-ResNeXt101	DeepLabv3+	0.895	0.951
Al-Mg-Si SC	ResNet50	UNet	0.285	0.294
	ResNet50	DeepLabv3	0.398	0.476
	ResNet50	DeepLabv3+	0.412	0.489
	SE-ResNeXt101	UNet	0.311	0.311
	SE-ResNeXt101	DeepLabv3	0.463	0.545
	SE-ResNeXt101	DeepLabv3+	**0.472**	**0.550**
Al Pitting	ResNet50	UNet	0.715	0.541
	ResNet50	DeepLabv3	0.231	0.348
	ResNet50	DeepLabv3+	0.253	0.374
	SE-ResNeXt101	UNet	**0.752**	**0.570**
	SE-ResNeXt101	DeepLabv3	0.245	0.365
	SE-ResNeXt101	DeepLabv3+	0.268	0.389
Contact Corrosion	ResNet50	UNet	0.923	0.911
	ResNet50	DeepLabv3	0.842	0.834
	ResNet50	DeepLabv3+	0.776	0.728
	SE-ResNeXt101	UNet	**0.951**	**0.962**
	SE-ResNeXt101	DeepLabv3	0.856	0.914
	SE-ResNeXt101	DeepLabv3+	0.526	0.628
Carbon	ResNet50	UNet	0.736	0.609
	ResNet50	DeepLabv3	0.524	0.649
	ResNet50	DeepLabv3+	0.586	0.672
	SE-ResNeXt101	UNet	**0.762**	**0.621**
	SE-ResNeXt101	DeepLabv3	0.547	0.671
	SE-ResNeXt101	DeepLabv3+	0.599	0.686

A.3 Explainability and Interpretability

The widespread adoption of deep learning in high-stakes domains has created an urgent need for explainable AI (XAI) methods that can illuminate the decision-making processes of complex models (Samek et al., 2021; Adadi & Berrada, 2018). Explainability serves multiple purposes: building user trust, meeting regulatory requirements, debugging models, and enabling scientific discovery (Arrieta et al., 2020). XAI methods can be categorized along several dimensions (Nauta et al., 2023; Vilone &

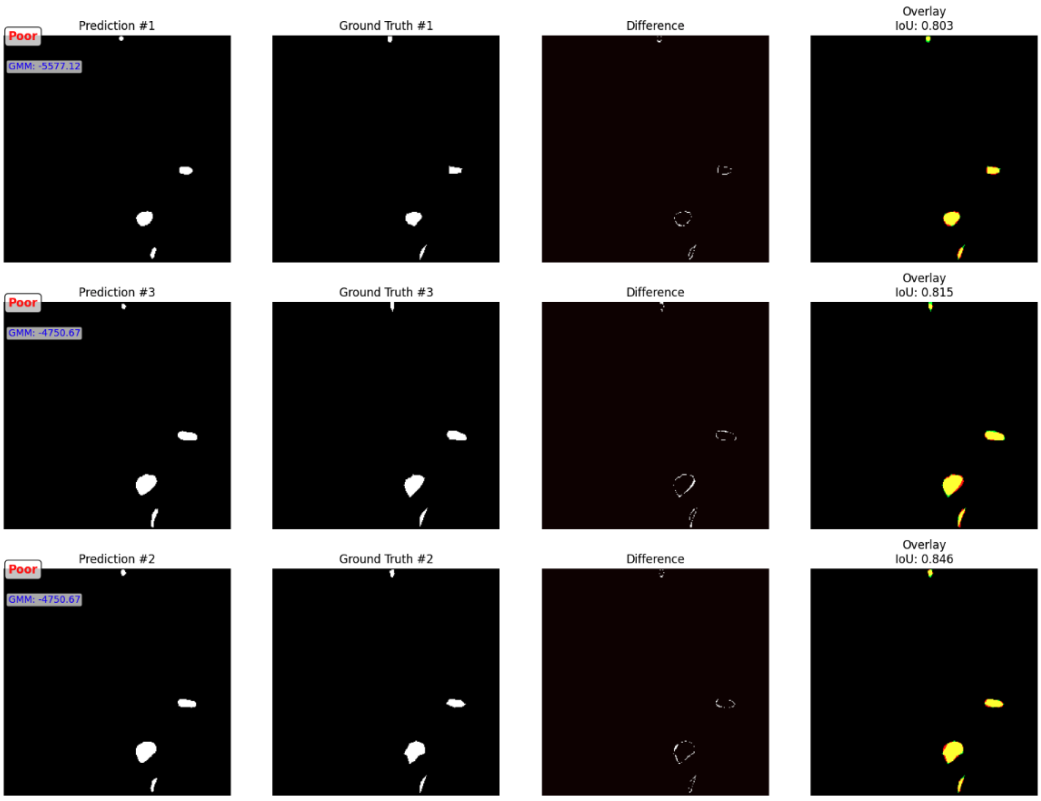


Figure 3: AFM crystallites segmentation failure cases flagged by OOD-based quality control. Each row shows a representative prediction assigned a low Forte score. From left to right: predicted segmentation, ground-truth annotation, pixel-wise difference, and overlay visualization with the corresponding IoU score. Although the IoU values remain moderately high, the predictions exhibit systematic errors in crystallite boundary delineation. These deviations from typical boundary morphology lead to low Forte scores, demonstrating the utility of OOD-based quality control for identifying subtle but semantically meaningful segmentation failures.

Longo, 2021). The task of evaluating which explainability method is best suited given a problem statement is an active area of research (Ganguly & Gupta, 2022).

While quality control identifies when predictions are unreliable (e.g., the boundary failures in Figure 3), it does not reveal why valid decisions are made. To address this, we employ counterfactual analysis (Figure 2d) to verify that the model attends to physically meaningful features, such as grain boundaries, rather than spurious correlations. Our choice of explainability method is guided by domain expert feedback on the qualitative visualizations.