GETTING THE DATA RIGHT: A PHYSICS-CONSISTENT, CALIBRATED DATASET FOR SEM-BASED DEFECT LOCALIZATION IN PEM FUEL CELLS

Rahul D. Ray
Department of Electrical and Electronics Engineering
Birla Institute of Technology and Science (BITS) Pilani, Hyderabad Campus
Hyderabad, Telangana, India
f20242213@hyderabad.bits-pilani.ac.in

Utkarsh Srivastava
Department of Economics
Birla Institute of Technology and Science (BITS) Pilani, Hyderabad Campus
Hyderabad, Telangana, India
f20240633@hyderabad.bits-pilani.ac.in

Rishi Mohapatra
Department of Chemical Engineering
Birla Institute of Technology and Science (BITS) Pilani, Hyderabad Campus
Hyderabad, Telangana, India
f20240796@hyderabad.bits-pilani.ac.in

Naveen K. Shrivastava
Supervisor, Assistant Professor
Department of Mechanical Engineering
Birla Institute of Technology and Science (BITS) Pilani, Hyderabad Campus
Hyderabad, Telangana, India
naveenks@hyderabad.bits-pilani.ac.in

ABSTRACT

High-quality data is a key bottleneck for vision systems in scientific imaging, yet publicly available datasets for defect localization in proton exchange membrane fuel cells remain scarce. We present a curated grayscale scanning electron microscopy dataset for single-class defect localization consisting of 1,107 images with bounding-box annotations, fixed train/validation/test splits, and a single canonical annotation source to ensure reproducibility. A physics-consistent preprocessing pipeline removes acquisition artifacts, enforces spatial standardization, and applies global intensity normalization to mitigate shortcut learning from non-physical cues. Controlled learnability and augmentation ablations show that even physically plausible transformations, including 90° rotations, can degrade detection performance, highlighting the need for dataset-specific validation rather than heuristic augmentation. By providing a rigorously validated and transparent benchmark for SEM-based defect localization, this dataset supports reliable automated characterization workflows and reduces a key data bottleneck in data-driven materials discovery and diagnostic pipelines.

1 INTRODUCTION

Automated defect localization in scanning electron microscopy (SEM) images remains a persistent challenge in materials science and manufacturing. SEM inspection pipelines are typically constrained by grayscale imagery, low defect-to-background contrast, and limited availability of labeled data,

all of which complicate reliable learning-based localization. Early studies addressed SEM defect analysis in narrowly defined material systems, such as nanofibrous structures, using classical methods or early deep learning models on small datasets Carrera et al. (2016); Napoletano et al. (2018). While these works established feasibility, their limited scale, material specificity, and restricted annotation regimes hindered generalization and reuse.

Subsequent research introduced deep learning–based object detection pipelines for electron microscopy, demonstrating improved localization performance when sufficient annotations were available Shen et al. (2021). However, these datasets were typically constructed for individual experiments, with limited emphasis on preprocessing transparency, validation rigor, or reproducibility. In semiconductor manufacturing, comparative evaluations have shown that performance differences between modern object detectors are often secondary to dataset composition and annotation quality Dehaerne et al. (2022). Larger SEM datasets, such as SEMI-CenterNet, target wafer inspection with multiple defect classes and are not released as fully open, standardized benchmarks De Ridder et al. (2023). Recent surveys consistently identify the lack of publicly available, well-validated SEM datasets as a central bottleneck Lechien et al. (2023); Dehaerne et al. (2025). These studies further highlight the risks of shortcut learning induced by acquisition artifacts and indiscriminate augmentation Geirhos et al. (2020); Lin et al. (2024); Compton et al. (2023). Data-centric frameworks therefore emphasize dataset construction and validation as primary research contributions Whang et al. (2023); Jarrahi et al. (2022); Singh (2023); Mazumder et al. (2023). In the context of proton exchange membrane (PEM) fuel cells, existing machine learning work largely focuses on numerical or simulation-based data rather than microscopy imagery Legala et al. (2022); Zuo et al. (2021); Yuan et al. (2023); Iranzo et al. (2022). To our knowledge, no openly accessible grayscale SEM dataset exists for bounding-box defect localization in PEM materials, motivating the dataset introduced in this work, directly building on recent evidence that data-centric physical consistency can outperform explicit constraint-based approaches in physics-informed machine learning Ray (2025).

2 Method and Dataset Construction

2.1 Dataset Construction

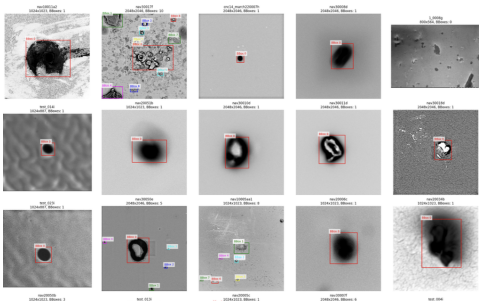
We construct a curated SEM defect detection dataset consisting of 206 images with bounding box annotations defined in absolute pixel coordinates. All annotations are stored in a single canonical JSON file, which serves as the sole authoritative annotation source throughout dataset construction, validation, and release. Framework-specific formats (e.g., YOLO-style labels) are derived deterministically from this representation for internal experiments. Representative examples of the processed SEM images and a summary of dataset partitioning are shown in Fig. 1.

It is important to note that the 206 images described above correspond to the set of unique raw SEM micrographs prior to preprocessing and data augmentation. This raw image set serves as the canonical source for all subsequent dataset transformations. The final dataset used for model training and benchmarking is derived from this raw corpus through standardized preprocessing and training-only data augmentation, as detailed in the following sections.

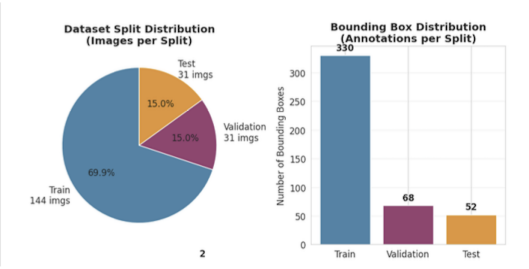
2.2 Preprocessing, Standardization, and Augmentation

To mitigate shortcut learning from acquisition-specific artifacts, a deterministic spatial preprocessing step is applied to all SEM images that removes the bottom 6% of each image corresponding to the scale-bar region. Bounding box annotations are programmatically adjusted to match the cropped image dimensions, with intersecting boxes clipped and invalid boxes discarded. This process reduces the annotation count from 451 to 450, indicating negligible impact on dataset composition. All images are then resized to a fixed resolution of 640×640 pixels using Lanczos interpolation, with bounding box coordinates scaled accordingly. While absolute aspect ratios are not preserved, spatial alignment between images and annotations remains exact in the transformed coordinate space. Pixel intensities are normalized using min–max scaling and stored in `uint8` format for compatibility with standard vision pipelines.

Following preprocessing, the dataset—consisting of 206 unique SEM images—is split into training, validation, and test subsets prior to any data augmentation to ensure strict separation between splits. Dataset expansion is performed exclusively on the training subset using a fixed set of deterministic,



(a) Representative SEM images after deterministic bottom-cropping (6%) to remove acquisition-specific artifacts (scale bars), applied uniformly across the dataset.



(b) Visualization of dataset partitioning, showing the proportion of images in the training, validation, and test splits (pie chart) alongside the distribution of bounding box annotations across splits (bar chart).

Figure 1: Dataset overview including preprocessing effects and train–validation–test partitioning.

label-preserving augmentation transforms. Each training image is replicated under predefined combinations of photometric variations, including brightness, contrast, and gamma adjustments within $\pm 15\%$, optional mild Gaussian noise ($\sigma \leq 0.01$), and light Gaussian blurring, as well as discrete rotations by integer multiples of $90°$ $\{0°, 90°, 180°, 270°\}$. Bounding box coordinates are transformed analytically to maintain exact spatial alignment. After augmentation, the final dataset contains 1,107 images in total, comprising 1,045 training images and 31 images each for validation and testing. Validation and test subsets remain entirely unaugmented, ensuring unbiased and reproducible evaluation.

2.3 Validation and Integrity Check

Annotation integrity is verified through automated and manual checks. Automated validation confirms the absence of out-of-bounds boxes, duplicate annotations, split leakage, and schema violations. Manual inspection of randomly sampled images further confirms correct spatial alignment between images and annotations. To assess robustness under augmentation, we apply a calibrated validation framework that separates hard structural requirements from diagnostic analyses. Statistical tests confirm preservation of bounding box geometry under augmentation, with no truncation or annotation loss observed. Augmented samples remain structurally distinct from their originals, ensuring statistical independence. A summary of dataset scale and annotation geometry after preprocessing is reported in Table A.2.1.

All integrity and validation checks are performed on both the raw pre-augmentation dataset and the final augmented training set, with validation and test subsets remaining unaugmented to preserve unbiased evaluation.

3 Baseline Learnability

Following the calibrated dataset validation described in Section 2.3, we evaluate baseline learnability to determine whether the constructed dataset supports non-trivial supervised defect localization under conservative experimental conditions. This evaluation is intended to verify dataset suitability rather than to optimize model performance or establish benchmark results.

3.1 Experimental Setup

Baseline evaluation is performed using two standard object detection models, **Faster R-CNN** and **RetinaNet**, each with a **ResNet-50-FPN** backbone. Both models are trained on the fixed training split and evaluated on the held out test set of 31 images, following the dataset preprocessing and validation procedures described in Appendix A.5 and A.1.

All inputs are single channel grayscale SEM micrographs. To ensure compatibility with ImageNet pretrained CNN backbones, grayscale images are converted to three channel format by replicating

intensities across RGB channels during both training and evaluation. Hyperparameters are held fixed across models. During evaluation, model weights are frozen, and no fine tuning, confidence threshold optimization, or test time augmentation is applied.

3.2 Evaluation Metrics

Performance is reported using the following metrics: (a) AP@0.5, (b) AP@0.5:0.95, (c) Mean Intersection over Union (IoU), (d) Recall@0.5, and (e) average per-image detection rate (diagnostic).

Average Precision is computed from precision recall curves by sweeping confidence thresholds over the full range of [0,1]. For point metrics, including Recall@0.5, Mean IoU, and average per image detection rate, a fixed confidence threshold of 0.5 is used uniformly across models without optimization. Confidence sweeping is used exclusively for constructing precision recall curves and computing AP. All point metrics are therefore evaluated at the same fixed threshold. Reported values are intended for relative learnability assessment within this study and are not meant for direct comparison with COCO benchmarks or state of the art results. s applied.

3.3 Quantitative Results

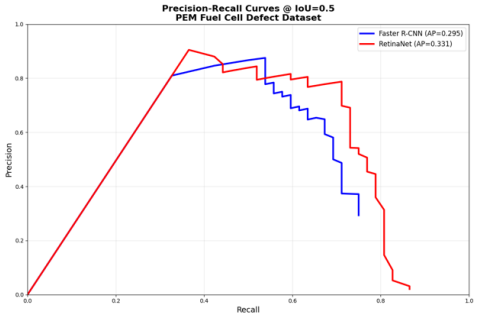


Figure 2: Precision-Recall Curve@IoU = 0.5

Table 1: Model Learnability Assessment

Metric	Faster R-CNN	RetinaNet
Avg. Det. Rate	0.83	0.86
Mean IoU	0.7882	0.7725
Recall@0.5	0.6346	0.7115
AP@0.5	0.2949	0.3306
AP@0.5:0.95	0.1739	0.1989
Recall@0.5	0.6346	0.7115

Both models achieve non-zero Average Precision (Table 1), confirming non-random learnability. RetinaNet demonstrates a consistent advantage in precision–recall balance (0.3306 vs. 0.2949) and recall (0.7115 vs. 0.6346), while Mean IoU values ($\sim$ 0.78) remain comparable across models, indicating reasonably accurate localization when detections are made. Average per-image detection rates exceed 0.8 for both models (0.83 for Faster R-CNN, 0.86 for RetinaNet), suggesting that defects are not systematically missed. The curves exhibit smooth, monotonic trade-offs without instability (Figure 2), indicating consistent evaluation behavior. RetinaNet maintains higher precision across a broader recall range, consistent with the quantitative results in Table 1. These curves provide qualitative confirmation of dataset learnability rather than evidence of optimized performance.

4 Data-Centric Augmentation Ablation Study

Following the baseline learnability assessment described in Section 3, we conduct a controlled ablation study to examine how physically plausible augmentation strategies interact with model learnability and dataset characteristics. The study serves a diagnostic purpose, aiming to characterize augmentation sensitivity under strict experimental protocols.

4.1 Experimental Setup

A controlled ablation study is conducted to evaluate three distinct augmentation strategies and isolate their specific effects. The study employs a single object detection architecture, Faster R-CNN with a ResNet-50-FPN backbone. All experiments use identical train/validation/test splits, fixed hyperparameters, and a consistent evaluation protocol to ensure comparability across conditions. Three training conditions are evaluated. G1 applies no augmentation and serves as a lower-bound reference. G2 applies photometric perturbations (brightness, contrast, gamma, additive noise ($\sigma \leq$

0.01), and blur within $\pm 15\%$ ranges). G3 extends G2 with additional 90° rotations from the set $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$. A summary of the ablation design is provided in Table B.1.1.

Augmentations are applied on-the-fly during training. All models are trained for 20 epochs using identical optimization settings (SGD with learning rate 0.005, momentum 0.9, weight decay 0.0005, and StepLR scheduling). Training is performed on the fixed training split (1,045 images), and evaluation is conducted on the held-out test set of 31 images. No test-time fine-tuning, confidence threshold optimization, or test-time augmentation is applied.

Input handling follows the same grayscale-to-RGB channel replication described in Section 3.1.

4.2 Evaluation Metrics

Average Precision (AP) is computed from precision–recall curves using a custom but consistent evaluation pipeline, applied uniformly across all augmentation conditions and following the metric definitions described in Section 3.2. AP is calculated by sweeping confidence thresholds over the full range of [0,1].

For point metrics, including Recall@0.5 and Mean IoU, a fixed confidence threshold of 0.5 is used consistently across all conditions, as defined in Section 3.2. This threshold is applied without optimization to ensure fair comparison across augmentation strategies. The reported metrics are intended for relative learnability assessment rather than direct comparison to COCO benchmarks or state-of-the-art detection results.

4.3 Quantitative Results

Table 2: Quantitative Results of Augmentation Ablation Study

Metric	G1: No Augmentation	G2: Photometric Only	G3: Photometric + Rotations
Mean IoU	0.7881	0.8004	0.7898
Recall@0.5	0.7115	0.6731	0.6923
AP@0.5	0.3416	0.3373	0.1939
AP@0.5:0.95	0.2083	0.2127	0.1176

Defying the expected monotonic improvement from increased augmentation, even the most aggressive strategy (G3) significantly degrades Average Precision compared to baseline (G1) and photometric-only (G2) conditions (Table 2). Photometric augmentation (G2) alone yields modest localization gains, increasing AP@0.5:0.95 from 0.2083 to 0.2127 and Mean IoU from 0.7881 to 0.8004. Conversely, adding 90° rotations (G3) causes a marked decline in AP metrics despite stable Recall@0.5 and Mean IoU. Specifically, G3 drops to 0.1939 AP@0.5 and 0.1176 AP@0.5:0.95, compared to the baseline (G1) values of 0.3416 and 0.2083, respectively (Table 2).

Precision–recall curves support this observation. The rotation-augmented condition (G3) exhibits consistently lower precision across most recall levels, indicating degraded precision–recall balance rather than loss of detectability (Figure B.2.1). Training dynamics provide complementary evidence, with G3 showing consistently higher training and validation losses throughout all 20 epochs, indicating increased optimization difficulty (Figures B.2.2 and B.2.3). Although 90° rotations are physically plausible for PEM fuel cell imagery, the observed degradation suggests that defect orientations and background textures are not statistically symmetric, and that rotation augmentation disrupts informative structure under limited data regimes (approximately 1,000 training images).

Overall, these results fulfill the diagnostic objective of the ablation study, demonstrating that physically plausible augmentations are not necessarily statistically beneficial and underscoring the need for dataset-specific augmentation validation.

5 Discussion

This work adopts a data-centric perspective, treating dataset construction, preprocessing, validation, and characterization as primary contributions rather than auxiliary details (Section 2). Analysis and

augmentation studies show that dataset design materially influences model behavior in scientific imaging. Physics-consistent preprocessing prevents shortcut learning by removing scale bars and acquisition artifacts (Section 2.1). Spatial characterization reveals non-uniform defect distributions independent of absolute image scale (Section 2.2). Baseline experiments confirm non-trivial supervised localization (Section 3). Ablation studies show that physically plausible transformations, such as 90° rotations, can degrade performance (Section 4), demonstrating that physical admissibility does not guarantee statistical benefit and underscoring the need for dataset-specific validation.

This study is intentionally scoped to data-centric analysis. It introduces no new architectures, performs no hyperparameter tuning, and does not address multi-class defect taxonomy. Evaluations exclude test-time augmentation and threshold optimization, and results are not intended for comparison with large-scale benchmarks such as COCO or to establish state-of-the-art performance.

6 Conclusion

We present a validated grayscale SEM dataset for single-class defect localization in fuel cell materials. The dataset features physics-consistent preprocessing, reproducible splits, integrity checks, and a calibrated validation framework. Baseline studies demonstrate that physically plausible augmentations can degrade performance when statistical assumptions are violated, underscoring the need for dataset-specific validation. This dataset provides a reliable foundation for future work on architectures, multi-class annotation, and domain-specific inductive biases.

References

Diego Carrera, Fabio Manganini, Giacomo Boracchi, and Ettore Lanzarone. Defect detection in sem images of nanofibrous materials. *IEEE Transactions on Industrial Informatics*, 13(2):551–561, 2016.

Rhys Compton, Lily Zhang, Aahlad Puli, and Rajesh Ranganath. When more is less: Incorporating additional datasets can hurt performance by introducing spurious correlations. In *Machine learning for healthcare conference*, pp. 110–127. PMLR, 2023.

Vic De Ridder, Bappaditya Dey, Enrique Dehaerne, Sandip Halder, Stefan De Gendt, and Bartel Van Waeyenberge. Semi-centernet: a machine learning facilitated approach for semiconductor defect inspection. In *38th European Mask and Lithography Conference (EMLC 2023)*, volume 12802, pp. 220–228. SPIE, 2023.

Enrique Dehaerne, Bappaditya Dey, and Sandip Halder. A comparative study of deep-learning object detectors for semiconductor defect detection. In *2022 29th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, pp. 1–2. IEEE, 2022.

Enrique Dehaerne, Bappaditya Dey, Victor Blanco, and Jesse Davis. Scanning electron microscopy-based automatic defect inspection for semiconductor manufacturing: a systematic review. *Journal of Micro/Nanopatterning, Materials, and Metrology*, 24(2):020901–020901, 2025.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Alfredo Iranzo, Baltasar Toharias, Christian Suárez, Felipe Rosa, and Javier Pino. Dataset and mesh of the cfd numerical model for the modelling and simulation of a pem fuel cell. *Data in Brief*, 41: 107987, 2022.

Mohammad Hossein Jarrahi, Ali Memariani, and Shion Guha. The principles of data-centric ai (dcai). *arXiv preprint arXiv:2211.14611*, 2022.

Thibault Lechien, Enrique Dehaerne, Bappaditya Dey, Victor Blanco, Sandip Halder, Stefan De Gendt, and Wannes Meert. Automated semiconductor defect inspection in scanning electron microscope images: a systematic review. *arXiv preprint arXiv:2308.08376*, 2023.

Adithya Legala, Jian Zhao, and Xianguo Li. Machine learning modeling for proton exchange membrane fuel cell performance. *Energy and AI*, 10:100183, 2022.

Manxi Lin, Nina Weng, Kamil Mikolaj, Zahra Bashir, Morten BS Svendsen, Martin G Tolsgaard, Anders N Christensen, and Aasa Feragen. Shortcut learning in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 623–633. Springer, 2024.

Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, et al. Dataperf: Benchmarks for data-centric ai development. *Advances in Neural Information Processing Systems*, 36:5320–5347, 2023.

Paolo Napoletano, Flavio Piccoli, and Raimondo Schettini. Anomaly detection in nanofibrous materials by cnn-based self-similarity. *Sensors*, 18(1):209, 2018.

Rahul D Ray. The physics constraint paradox: When removing explicit constraints improves physics-informed data for machine learning. *arXiv preprint arXiv:2512.22261*, 2025.

Mingren Shen, Guanzhao Li, Dongxia Wu, Yuhan Liu, Jacob RC Greaves, Wei Hao, Nathaniel J Krakauer, Leah Krudy, Jacob Perez, Varun Sreenivasan, et al. Multi defect detection and analysis of electron microscopy images with deep learning. *Computational Materials Science*, 199:110576, 2021.

Prerna Singh. Systematic review of data-centric approaches in artificial intelligence and machine learning. *Data Science and Management*, 6(3):144–157, 2023.

Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. Data collection and quality challenges in deep learning: A data-centric ai perspective. *The VLDB Journal*, 32(4):791–813, 2023.

Hao Yuan, Shaozhe Zhang, Xuezhe Wei, and Haifeng Dai. Fault diagnosis of proton exchange membrane fuel cell based on nonlinear impedance spectrum. *Automotive Innovation*, 6(4):597–610, 2023.

Jian Zuo, Hong Lv, Daming Zhou, Qiong Xue, Liming Jin, Wei Zhou, Daijun Yang, and Cunman Zhang. Long-term dynamic durability test datasets for single proton exchange membrane fuel cell. *Data in Brief*, 35:106775, 2021.

7 APPENDIX

A ADDITIONAL DATASET ANALYSIS AND VALIDATION

This appendix provides extended dataset analyses, validation protocols, and algorithmic specifications that support reproducibility and transparency. The material presented here complements the main paper by documenting secondary analyses and integrity checks that are not required for understanding the core dataset construction and baseline evaluation, but are included to facilitate reuse, auditing, and methodological extension.

A.1 STANDARDIZED PREPROCESSING AND ANNOTATION VALIDATION PIPELINE

We first present the complete deterministic preprocessing and annotation validation pipeline used to transform raw SEM micrographs into a training-ready dataset. The pipeline enforces spatial and intensity standardization while preserving annotation geometry, and logs potential integrity issues without excluding valid samples. This explicit specification is provided to support exact reimplementation and dataset auditing. Figure A.1.1 details the full procedure, including spatial normalization, bounding box scaling, intensity normalization, and annotation integrity checks.

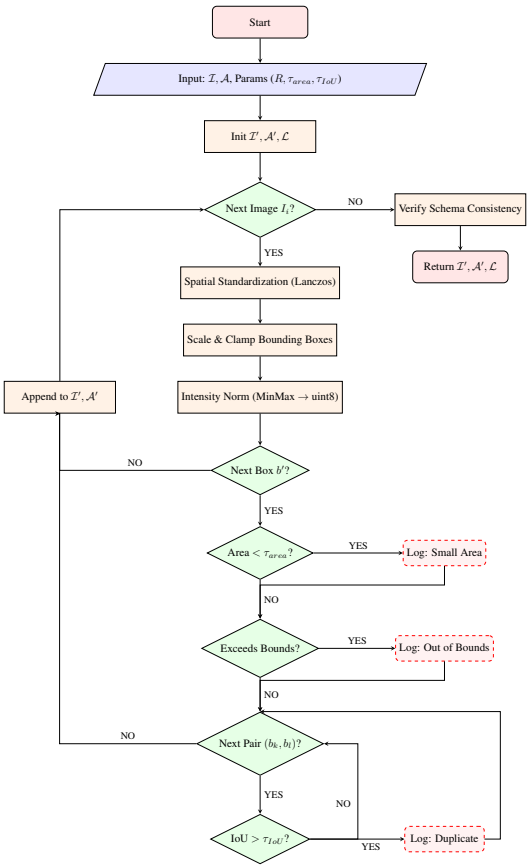


Figure A.1.1: Standardized Preprocessing and Annotation Validation Pipeline. I denotes the input SEM image and A its bounding-box annotations. Spatial standardization is applied using resizing operator R (Lanczos interpolation), producing processed outputs I' and updated annotations A'. Each transformed box b' is scaled and clamped to image bounds, with small-area cases flagged if $\mathrm{Area}(b') < \tau_{\mathrm{area}}$. Annotation validity is verified via out-of-bounds checks, and duplicate boxes are detected by pairwise overlap testing, logging duplicates when $\mathrm{IoU}(b_k, b_l) > \tau_{\mathrm{IoU}}$. L records all diagnostic integrity events (small boxes, truncation, duplicates) without excluding valid samples.

A.2 Dataset Scale and Annotation Geometry

Table A.2.1 summarizes dataset scale and bounding box geometry after preprocessing and prior to data augmentation. These statistics highlight the sparsity of defects, wide intra-image variation, and geometric consistency of annotations, providing context for the detection difficulty and data-scarce regime addressed in this work.

Table A.2.1: Dataset scale, defect frequency, and bounding box geometry after preprocessing and before augmentation.

Metric	**Value**
Total Images	206
Total Bounding Boxes	450
Boxes per Image (Mean $\pm$ Std)	2.18 ± 2.54
Boxes per Image (Range)	0–14
Bounding Box Area (Median)	5,365.5 px^2
Aspect Ratio (Median)	1.04
Mean Width / Height	131.6 px / 126.0 px

A.3 Spatial Distribution of Defect Locations

To examine spatial trends in defect occurrence, we analyze the distribution of defect centroids across the image plane. Figure A.3.1 visualizes aggregated defect locations using both absolute pixel coordinates and normalized coordinates.

The absolute-coordinate heatmap reflects acquisition-space bias, while the normalized-coordinate view enables comparison across images with identical post-processed resolution. The results reveal mild lateral asymmetry and non-uniform spatial density, while maintaining low overall defect concentration. These patterns may reflect systematic process or imaging effects; however, no causal interpretation is asserted.

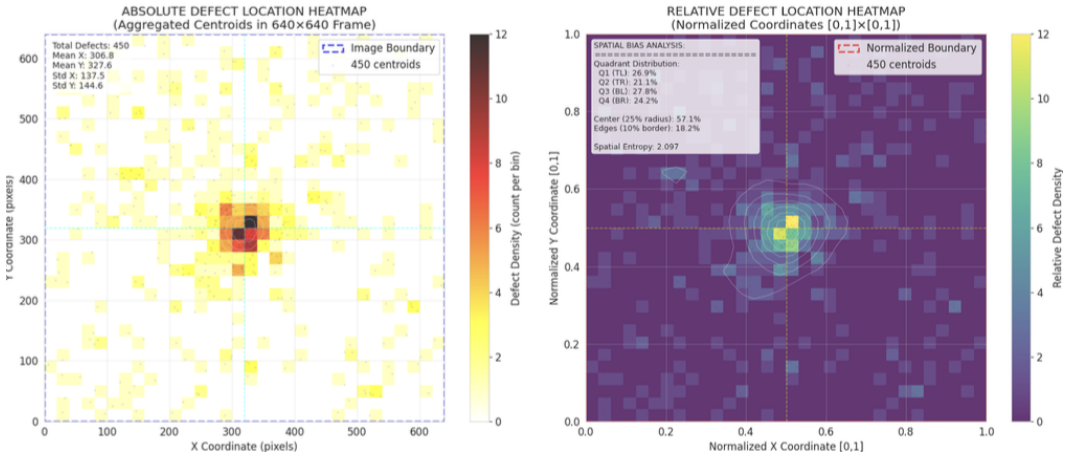


Figure A.3.1: Aggregated spatial distribution of defect centroids. Left: absolute pixel coordinates. Right: normalized coordinates.

A.4 Dataset-Level Process and Morphology Indicators

Table A.4.1 reports summary indicators describing spatial uniformity, defect morphology, and positional bias. These metrics are intended as qualitative descriptors rather than formal process diagnostics.

Table A.4.1: Dataset-level indicators of defect uniformity, morphology, and spatial bias.

Category	**Finding**	**Interpretation**
Uniformity Proxy	0.223 (Low)	Mild spatial non-uniformity
Defect Morphology	97.6% rounded	Predominantly isotropic defects
Center vs Edge	Ratio = 2.10	Center-biased occurrence
Defect Density	5.33 ppm	Sparse defect regime
Lateral Bias	Right-side preference (55.6%)	Possible alignment asymmetry

A.5 Calibrated Validation Framework

To balance dataset rigor with inclusivity, we adopt a calibrated validation framework that separates mandatory structural requirements from informative diagnostic analyses. Hard requirements enforce annotation correctness and dataset integrity, while diagnostic checks provide insight into augmentation behavior without acting as exclusion criteria.

Table A.5.1 summarizes the validation categories, evaluation metrics, and their intended roles.

Table A.5.1: Calibrated validation framework for dataset integrity and augmentation behavior.

Category	**Test Modality**	**Metric(s)**	**Purpose**
Geometric Preservation	Statistical	KS-test p-values	Enforce geometry invariance
Aspect Ratio Stability	Diagnostic	Percentage change	Shape consistency
Truncation Detection	Geometric	Box loss rate	Prevent clipping
Statistical Independence	Similarity	Mean SSIM	Avoid duplicates
Physical Plausibility	Image analysis	Gradients, contrast	Guide augmentation limits
Dataset Integrity	Structural	Leakage, duplication	Ensure correctness

B Data-Centric Augmentation Ablation Details

This appendix provides supplementary results for the baseline learnability and augmentation ablation experiments (Sections 3 and 4), including ablation group definitions and supporting training diagnostics such as precision–recall curves and loss trajectories to ensure transparency and reproducibility.

B.1 ablation study design

Table B.1.1: Supplementary summary of augmentation conditions used in the controlled ablation study in Section 4.

Condition	**Augmentation Strategy**	**Key Parameters**	**Purpose**
G1: Baseline	No augmentation	Original images only	Establish lower-bound learnability
G2: Photometric	Brightness, contrast, gamma, noise, blur	($\pm 15\%$, $\sigma \leq 0.01$)	Assess robustness to imaging variations
G3: Photometric + Rotations	G2 augmentations + 90° rotations	$\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$	Evaluate impact of orientation diversity

B.2 ABLATION STUDY RESULTS VISUALIZATION: P-R CURVE, TRAINING AND VALIDATION LOSS OVER EPOCHS

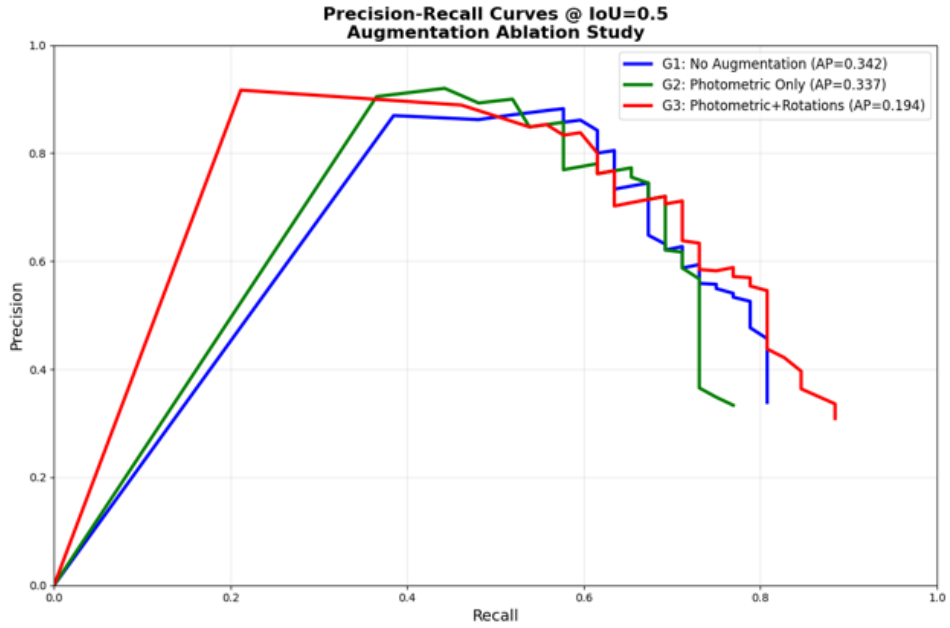


Figure B.2.1: Precision-Recall curve at IoU = 0.5

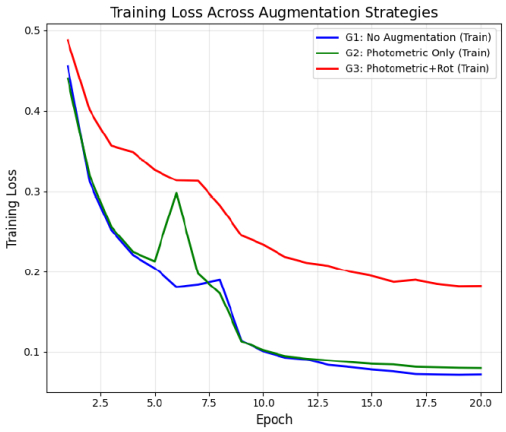


Figure B.2.2: Training loss trajectory over 20 epochs for the augmentation ablation conditions.

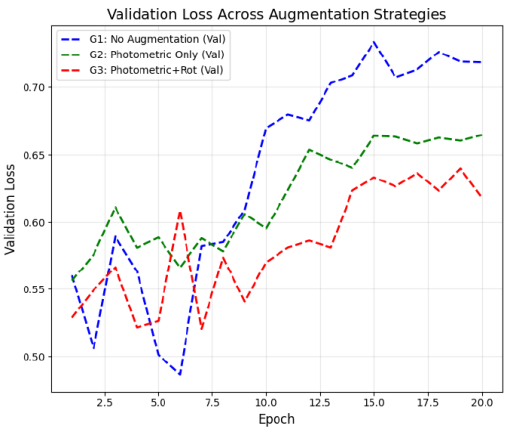


Figure B.2.3: Validation loss trajectory over 20 epochs for the augmentation ablation conditions.

C LLM USAGE DISCLOSURE

Large Language Models (LLMs) were used in limited capacity during the preparation of this research. Specifically, LLMs were used to check grammar and refine sentence structure after the initial draft was completed, primarily to correct awkward expressions and maintain consistency in writing style. However, all core research ideas, analytical methodologies, interpretations of the results, and conclusions were developed entirely by the authors. The LLM did not contribute to any creative content or academic judgments. This use of LLMs was conducted within limits that do not compromise the originality or academic integrity of the research.