IJTIS

International Journal of Technological Innovation for Society (IJTIS)
https://journals.telkomuniversity.ac.id/ijtis

Vol. 1, No.1, 1–9, 2025
Publisher: Telkom University Journals

RESEARCH ARTICLE

Beyond Clean Accuracy: Corruption Robustness and Failure Modes of CNN and Transformer Models for Magnetic Tile Defect Segmentation

Andi Shafira Dyah Kurniasari[1*], **Karisa Ardelia Hanifah**[1] **and Khalis Sofi**[1]

[1] *School of Computing, Telkom University, Bandung, Indonesia*
[*] *Corresponding author*

Abstract
Reliable magnetic-tile defect segmentation requires evaluation beyond clean images because model behavior may change under degraded acquisition conditions. This study compared U-Net + ResNet18, DeepLabV3+ + MobileNetV2, and SegFormer-B0 on 199 independent scene groups using group-aware five-fold cross-validation. Four test-only corruptions, Gaussian Noise, Gaussian Blur, Brightness Increase, and Contrast Reduction, were evaluated at five severity levels. Clean Macro Dice showed no significant model effect (Friedman $p = 0.1693$). In contrast, mean corrupted Dice differed significantly ($p = 0.001118$); SegFormer-B0 achieved the highest aggregate *mDice-C* (0.2935) and lowest mRPD (53.02%), with significantly higher paired mean corrupted Dice than both CNN baselines after Holm correction. This advantage was not universal: under Gaussian Noise, SegFormer-B0 had the lowest mean defect Dice, but at S5 its defect-free FPAR was only 0.0357, compared with 0.4753 for U-Net and 0.7869 for DeepLabV3+, despite an IFAR of 0.880. These findings show that clean segmentation performance alone does not fully characterize reliability and that corruption-aware assessment should jointly consider defect overlap, false-alarm frequency, and false-positive spatial extent.

Keywords: common corruptions, defect segmentation, magnetic tile inspection, robustness, SegFormer, semantic segmentation

Corresponding author:
andishafiradyah@telkomuniversity.ac.id

Received: 15 August 2026
Accepted: 23 August 2026
Available online: 27 August 2026

1. Introduction

1.1. Background

Automated surface-defect inspection is central to industrial quality control, but magnetic-tile defects vary in morphology, scale, orientation, and contrast, while acquisition conditions can also alter image appearance. The Magnetic Tile Surface Defects Dataset provides pixel-level annotations that have supported learning-based localization and segmentation research [1]. Subsequent studies have improved magnetic-tile segmentation through multiscale representation, attention mechanisms, and lightweight architectures [2, 3]. However, these advances have been evaluated primarily under standard benchmark conditions.

Clean segmentation performance does not necessarily indicate how a model behaves when image quality deteriorates. Sensor noise, blur, illumination changes, and reduced contrast can alter both defect appearance and normal surface texture without changing the physical state of the tile. Common-corruption studies in semantic segmentation have shown that degradation can substantially reduce predictive performance and that robustness varies across architectures and corruption types [4]. For industrial inspection, reliability also includes behavior on defect-free products because corrupted inputs may generate false-positive regions that increase unnecessary rejection or manual reinspection. Robustness assessment should therefore consider both defect localization and false-positive behavior under degraded inputs.

Cite as: Kurniasari et al.(2025). Beyond Clean Accuracy: Corruption Robustness and Failure Modes of CNN and Transformer Models for Magnetic Tile Defect Segmentation. International Journal of Technological Innovation for Society , 1(1) (2025) 1–9.

1.2. Research Problem and Gap

Existing magnetic-tile studies have primarily emphasized segmentation accuracy, feature representation, localization quality, or computational efficiency, leaving corruption-induced reliability differences insufficiently characterized in this domain. Evidence from general semantic-segmentation benchmarks cannot be assumed to transfer directly because magnetic-tile images differ in visual structure, foreground prevalence, defect morphology, and inspection objectives. Although U-Net, DeepLabV3+, and SegFormer represent different feature-processing strategies, prior robustness results for Transformer-based segmentation, including SegFormer on Cityscapes-C, provide motivation for comparison rather than evidence of a predetermined advantage [5].
A second gap concerns failure characterization. Dice is appropriate for measuring defect-region overlap on defect-bearing scenes, but it does not describe how often or how extensively a model predicts defects on normal tiles. Broader reliability research similarly distinguishes predictive accuracy from other dimensions of reliable model behavior [6]. In addition, the dataset contains multiple acquisitions of the same underlying scenes, making scene-level independence important for avoiding leakage across experimental roles. This study therefore addresses the limited scene-independent characterization of corruption robustness and false-positive failure behavior among representative CNN- and Transformer-based magnetic-tile segmentation models.

1.3. Research Objective and Questions

This study compares U-Net + ResNet18, DeepLabV3+ + MobileNetV2, and SegFormer-B0 using 199 independent magnetic-tile scene groups under group-aware five-fold cross-validation. Four test-only corruptions, Gaussian Noise, Gaussian Blur, Brightness Increase, and Contrast Reduction, are evaluated at five severity levels. Defect-bearing scenes are assessed primarily with scene-level Macro Dice, while defect-free scenes are evaluated using the False Positive Area Ratio (FPAR) and Image False Alarm Rate (IFAR). The study addresses the following questions:
RQ1. How do the three evaluated architectures perform on clean magnetic-tile defect segmentation?
RQ2. How do the four corruptions and their severity levels affect segmentation performance?
RQ3. Are there statistically significant differences among the architectures in aggregate corruption robustness?
RQ4. How do the evaluated corruptions affect the frequency and spatial extent of false-positive predictions on defect-free magnetic tiles?

1.4. Contributions

The study makes three main contributions:

1. A scene-independent, leakage-conscious evaluation protocol based on 199 scene groups and group-aware five-fold cross-validation, with the scene group treated as the experimental and statistical unit.
2. A controlled test-only corruption benchmark across three representative CNN- and Transformer-based segmentation configurations, combined with scene-level statistical comparison of clean and mean corrupted Dice.
3. A complementary reliability analysis that combines defect-scene segmentation with FPAR and IFAR on defect-free scenes, enabling corruption-dependent failure modes to be identified beyond defect-overlap performance alone.

2. Literature Review

2.1. Magnetic-Tile Surface Defect Segmentation

Magnetic-tile surface inspection is challenging because defects differ in morphology and scale, while normal surfaces may contain texture variations that complicate localization. The Magnetic Tile Surface Defects Dataset introduced by Huang et al. [1] provided pixel-level annotations for several defect categories and has become a common benchmark for learning-based inspection. Subsequent research has increasingly shifted from conventional detection toward pixel-level localization and segmentation. Representative studies have explored multilevel convolutional features, instance segmentation, attention-enhanced encoder-decoder models, and multiscale architectures to improve the representation of small or irregular defects [2], [3], [5].
More recent work has also incorporated self-attention and Transformer-based representations. Liu et al. [7] used cross self-attention to connect high- and low-level features, while Jiang et al. [8] proposed CINFormer, which combines multistage CNN features with a Transformer encoder for industrial surface-defect segmentation. Collectively, these studies show continued progress in feature fusion, localization accuracy, and computational efficiency. However, their primary emphasis remains predictive performance under the respective benchmark protocols. Consequently, strong clean segmentation results do not by themselves establish whether the same models remain reliable when image quality is systematically degraded.

2.2. Segmentation Architectures and Corruption Robustness

The architectures compared in this study represent three distinct segmentation strategies. U-Net uses an encoder-decoder structure with skip connections to combine contextual and high-resolution spatial information [9]; its ResNet18 encoder introduces residual feature extraction [10]. DeepLabV3+ combines atrous multiscale contextual processing with decoder-based boundary refinement [11], while MobileNetV2 provides an efficient convolutional encoder based on inverted residual blocks and linear bottlenecks [12]. SegFormer differs from these CNN baselines by using a hierarchical Mix Transformer encoder and a lightweight multiscale decoder [5]. These architectural differences provide a useful basis for examining whether models that appear comparable under clean evaluation also respond similarly to degraded inputs.
Common-corruption benchmarking treats robustness as a distinct evaluation dimension from clean predictive performance. Hendrycks and Dietterich [13] formalized this approach by testing models on multiple non-adversarial corruption families and severity levels without retraining

on the evaluation corruptions. For semantic segmentation, Kamann and Rother [4] showed that common image corruptions can substantially reduce performance and that degradation patterns depend on both architecture and corruption type. Their findings also demonstrate why a single aggregate score is insufficient: a model that is comparatively robust to one degradation may remain highly vulnerable to another.
Transformer-based segmentation provides additional motivation for such comparisons. The original SegFormer study reported favorable zero-shot robustness on Cityscapes-C [5], while broader work on Vision Transformers has suggested that attention-based representations can differ from convolutional models in their response to perturbations and domain shifts. Nevertheless, these observations do not establish a universal Transformer advantage. Robustness remains dependent on architecture, dataset, task, training configuration, and corruption type. For magnetic-tile inspection, corruption behavior therefore requires direct domain-specific evaluation rather than inference from clean rankings or results reported on general-scene datasets.

2.3. Reliability Beyond Clean Segmentation and Research Position

Reliable segmentation cannot be characterized only by foreground overlap. Broader semantic-segmentation research has distinguished conventional predictive accuracy from other reliability dimensions such as robustness, calibration, and out-of-distribution behavior [6]. This distinction is particularly important in industrial inspection because defect-bearing and defect-free images represent different operational concerns. On defect-bearing scenes, overlap metrics such as Dice describe how accurately the defect region is localized. On defect-free scenes, however, reliability concerns whether spurious defect predictions occur and how extensive those predictions become.
False-positive frequency and spatial extent are not interchangeable. A model may produce very small positive regions on many normal images, yielding frequent alarms with limited affected area, whereas another may trigger less frequently but generate extensive false-positive segmentation when it fails. The present study therefore uses the False Positive Area Ratio (FPAR) and Image False Alarm Rate (IFAR) as complementary measures for defect-free scenes. Their mathematical definitions are provided in the evaluation methodology; within the literature context, their role is to distinguish how much of a normal surface is incorrectly segmented from how often a normal image triggers a false alarm [6].
The reviewed literature therefore leaves three connected gaps. First, corruption robustness remains insufficiently characterized for representative CNN- and Transformer-based magnetic-tile segmentation models. Second, defect-overlap performance alone cannot describe corruption-induced false-positive behavior on defect-free tiles. Third, the presence of multiple acquisitions from the same underlying magnetic-tile scenes requires evaluation designs that preserve scene-level independence when comparing models statistically. The present study addresses these gaps through a unified scene-aware evaluation of clean segmentation, controlled test-time corruption, and complementary false-alarm behavior. Its contribution is therefore not a new segmentation architecture, but an evaluation framework for determining whether clean performance provides a sufficient account of model reliability under controlled image degradation [6].

3. Research Method

This study evaluated native corruption robustness under a common scene-independent protocol comprising dataset auditing, scene grouping and quality control, group-aware cross-validation, standardized preprocessing and training, test-only corruption, scene-level evaluation, and statistical analysis.

3.1. Dataset, Scene Grouping, and Quality Control

Experiments used the Magnetic Tile Surface Defects Dataset [14, 15], containing grayscale images with pixel-level masks. The initial audit identified 1,344 image-mask pairs: 952 defect-free images and 392 images from blowhole, break, crack, fray, and uneven categories. All defect categories were merged into one foreground class for binary semantic segmentation, and masks were binarized at an intensity threshold of 127. Because the dataset contains six nominal acquisitions (exp1–exp6) of underlying scenes, the scene group rather than the individual image was treated as the independent experimental and statistical unit. Scene grouping was completed before splitting, with exp4 predefined as the reference anchor. Non-reference acquisitions were matched to candidate exp4 anchors using normalized correlation of histogram-equalized 64×64 grayscale fingerprints within a ± 1000 identifier window, with neighboring exp4 anchors considered when no anchor existed in that interval. Deterministic quality control retained only groups containing exactly six valid acquisitions, consistent class labels, readable and spatially compatible image-mask pairs, and valid mask semantics. Of 224 candidate groups, 199 passed quality control, yielding 1,194 valid image-mask pairs from 150 defect-free and 49 defect-bearing scenes. The defect scenes comprised 16 blowhole, 13 uneven, 11 break, 5 crack, and 4 fray scenes. A frozen manifest was used for all subsequent experiments.

3.2. Group-Aware Cross-Validation and Preprocessing

The 199 retained scenes were assigned to five scene-level folds using seed 2026 while balancing the original classes. Folds contained 40, 40, 40, 40, and 39 scenes, respectively, with 30 defect-free scenes in each fold. For run (k), fold (k) was used for testing, fold $(k+1)$ mod 5 for validation, and the remaining three folds for training. Each scene therefore served as an out-of-fold test case exactly once.
All six acquisitions (exp1–exp6) from training scenes were eligible for model fitting, whereas validation and testing used only the predefined exp4 reference acquisition. This policy allowed training to use acquisition variability while standardizing evaluation to one image per independent scene. Across three architectures and five runs, the protocol produced 15 final training runs. The overall workflow is summarized in Figure 1.
Images were resized with preserved aspect ratio so that the longest side was 512 pixels and were zero-padded to 512×512; masks were resized with nearest-neighbor interpolation. Grayscale images were replicated to three channels and normalized using ImageNet statistics, while padded pixels were excluded from metric computation. Training-only geometric augmentation consisted of horizontal flipping, vertical flipping, and random 90°, 180°, or 270° rotation, each activated with probability 0.5. No photometric or evaluation-corruption augmentation was used during final training.

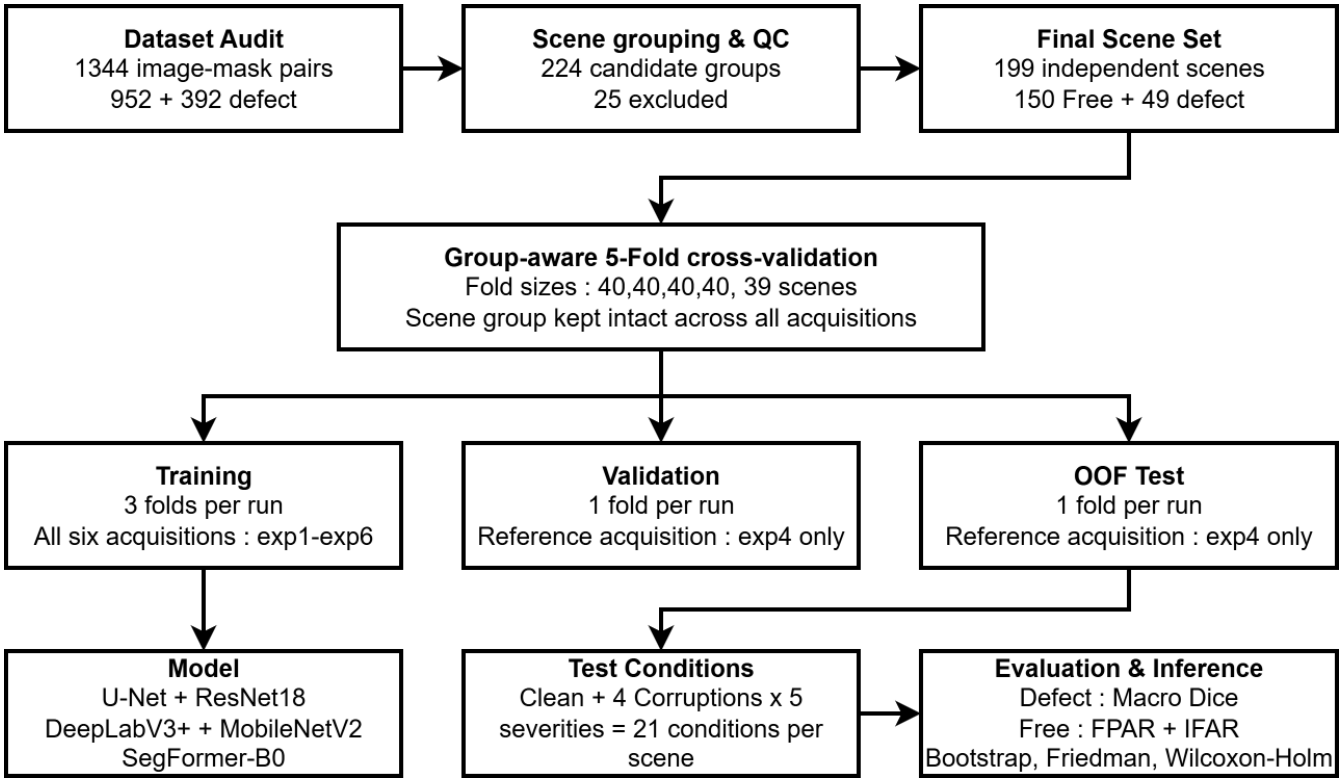

Figure 1: Overview of the scene-aware training, validation, and corruption-testing protocol.

3.3. Group-Aware Five-Fold Cross-Validation and Reference-Acquisition Protocol

Three ImageNet-pretrained configurations were compared: U-Net + ResNet18, DeepLabV3+ + MobileNetV2, and SegFormer-B0 with a MiT-B0 encoder. All models received $3 \times 512 \times 512$ inputs and produced a single-channel binary segmentation output. Binary cross-entropy with logits was calculated for all samples, while Soft Dice loss was calculated only for defect-bearing samples. For batches containing foreground, the objective was

$$L = 0.5L_{\text{BCE}} + 0.5L_{\text{Dice}} \tag{1}$$

whereas all-free batches used

$$L = L_{\text{BCE}}. \tag{2}$$

Training used AdamW with a learning rate of 1×10^{-4} and weight decay of 1×10^{-4}, CosineAnnealingLR with $T_{\max} = 100$ and a minimum learning rate of 1×10^{-6}, and a maximum of 100 epochs. Early stopping monitored clean validation Macro Dice with patience 15 and a minimum improvement of 1×10^{-4}. A global batch size of 4 and automatic mixed precision were used for all runs, with fold-specific seeds 2027–2031. Checkpoints were selected only from clean validation performance. Inference used a fixed threshold of 0.5 for every model and condition, and corrupted test performance was never used for model selection or tuning. Robustness testing was opened only after all 15 final checkpoints had been validated. Final experiments were executed on an NVIDIA Tesla T4 using PyTorch 2.11.0 and CUDA 12.8.

3.4. Test-Only Corruption Protocol

The four synthetic evaluation corruptions were applied only to out-of-fold exp4 test images after resizing and before padding; ground-truth masks were unchanged. Each corruption was evaluated at five predefined severity levels.

Table 1: Corruption parameters and severity levels.

Corruption	Transformation / Parameter	S1–S5
Gaussian Noise	$I' = \text{clip}(I+\varepsilon, 0, 1)$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$	$\sigma = 0.08, 0.12, 0.18, 0.26, 0.38$
Gaussian Blur	Gaussian smoothing	$\sigma = 1, 2, 3, 4, 6$
Brightness Increase	$I' = \text{clip}(I+c, 0, 1)$	$c = 0.10, 0.20, 0.30, 0.40, 0.50$
Contrast Reduction	$I' = (I-\mu)c+\mu$	$c = 0.40, 0.30, 0.20, 0.10, 0.05$

Together with the clean reference, this produced 21 conditions per out-of-fold scene, or 4,179 scene-condition observations per model. These are repeated evaluations of 199 independent scenes and were not treated as 4,179 independent experimental units. The corruptions were designed as controlled image-quality stress tests rather than physical simulations of all production-domain shifts.

3.5. Evaluation Metrics and Robustness Measures

Metrics were computed for each independent scene over the valid non-padded region and then macro-averaged. For the 49 defect-bearing scenes, the primary endpoint was Dice,

$$\text{Dice} = \frac{2TP}{2TP+FP+FN}. \tag{3}$$

Macro IoU, precision, recall, and Boundary IoU were reported as secondary descriptive metrics; Boundary IoU used a boundary width equal to 2% of the valid-image diagonal. The 150 defect-free scenes were evaluated using complementary false-alarm measures. False Positive Area Ratio quantified the fraction of valid pixels predicted as defective,

$$\text{FPAR} = \frac{N_{\text{pred-positive}}}{N_{\text{valid}}}. \tag{4}$$

whereas Image False Alarm Rate quantified the fraction of defect-free scenes containing at least one positive pixel,

$$\text{IFAR} = \frac{N_{\text{alarm}}}{N_{\text{Free}}}. \tag{5}$$

No morphological filtering was applied before FPAR calculation. For corruption-level summaries on defect-free scenes, FPAR and IFAR were averaged across severity levels S1–S5.
Let D_0 denote clean Macro Dice and $D(c,s)$ Dice under corruption c at severity s. Absolute degradation was calculated as

$$\Delta D(c,s) = D_0 - D(c,s), \tag{6}$$

whereas relative degradation was calculated as

$$\text{RPD}(c,s) = \frac{D_0 - D(c,s)}{D_0} \times 100\%. \tag{7}$$

Aggregate robustness was summarized by

$$\text{mDice-C} = \frac{1}{20}\sum_{c}\sum_{s} D(c,s), \tag{8}$$

and

$$\text{mRPD} = \frac{D_0 - \text{mDice-C}}{D_0} \times 100\%. \tag{9}$$

Aggregate values were interpreted together with corruption- and severity-specific results rather than as universal architecture rankings.

3.6. Statistical Analysis

Statistical inference used the scene group as the independent unit. For model comparison, paired observations were the same 49 out-of-fold defect-bearing scenes evaluated by all three architectures. Two primary endpoints were predefined: clean scene-level Dice and scene-level mean corrupted Dice averaged across the 20 corrupted conditions. Corruption-specific, severity-specific, false-alarm, and defect-type comparisons were treated descriptively rather than as additional model-ranking endpoints.
Uncertainty was estimated using 95% percentile bootstrap confidence intervals with 5,000 scene-level resamples. For each primary endpoint, an omnibus Friedman test assessed differences among the three paired models at $\alpha = 0.05$. Pairwise Wilcoxon signed-rank tests were performed only when the corresponding Friedman test was significant, and the three pairwise p-values were adjusted using the Holm procedure. Paired mean Dice differences with 95% bootstrap confidence intervals were reported to indicate the direction and magnitude of model differences. Defect-type inference was not performed because several categories contained few independent scenes, particularly crack ($n = 5$) and fray ($n = 4$).

4. Results and Discussion

Results are presented from clean performance and aggregate corruption robustness to corruption-specific behavior and false-positive failure modes on defect-free scenes. Statistical inference was restricted to the two predefined primary endpoints, clean Dice and mean corrupted Dice; corruption-specific, severity-specific, false-alarm, and defect-type analyses were descriptive.

Table 2: Clean and Aggregate Corruption Performance.

Model	Clean Dice (95% CI)	mDice-C (95% CI)	mRPD
U-Net + ResNet18	0.5965 (0.5123–0.6767)	0.1912 (0.1520–0.2318)	67.95%
DeepLabV3+ + MobileNetV2	0.5710 (0.4798–0.6540)	0.2026 (0.1554–0.2499)	64.52%
SegFormer-B0	0.6247 (0.5417–0.7023)	0.2935 (0.2410–0.3439)	53.02%

4.1. Clean Performance and Aggregate Corruption Robustness

Across the 49 independent defect-bearing scenes, SegFormer-B0 achieved the highest numerical clean Macro Dice (0.6247), followed by U-Net + ResNet18 (0.5965) and DeepLabV3+ + MobileNetV2 (0.5710). SegFormer-B0 also obtained the highest numerical values for the secondary clean metrics, although formal model comparison was predefined only for Macro Dice. The Friedman test detected no significant model effect for clean Dice ($p = 0.1693$). The numerical clean ranking should therefore be interpreted descriptively rather than as evidence of statistical superiority.

A different pattern emerged under corruption. All architectures degraded substantially, but SegFormer-B0 retained the highest mean corrupted Dice (*mDice-C* $= 0.2935$) and the lowest mean Relative Performance Degradation (mRPD = 53.02%). U-Net and DeepLabV3+ achieved *mDice-C* values of 0.1912 and 0.2026, corresponding to mRPD values of 67.95% and 64.52%, respectively. Unlike the clean endpoint, the Friedman test for scene-level mean corrupted Dice indicated a significant model effect ($p = 0.001118$).

Holm-corrected Wilcoxon comparisons showed significantly higher mean corrupted Dice for SegFormer-B0 than for U-Net (paired difference U-Net − SegFormer $= -0.1022$; 95% CI: -0.1382 to -0.0664; adjusted $p = 0.0000326$) and DeepLabV3+ (DeepLabV3+ − SegFormer $= -0.0908$; 95% CI: -0.1457 to -0.0335; adjusted $p = 0.00122$). U-Net and DeepLabV3+ did not differ significantly (difference $= -0.0114$; 95% CI: -0.0574 to 0.0293; adjusted $p = 0.6933$).

The contrast between the two primary endpoints is central to the study. The models were not statistically distinguishable by clean Dice, whereas significant differences emerged when performance was averaged over the 20 corrupted conditions. Nevertheless, SegFormer-B0's *mDice-C* of 0.2935 and approximately 53% relative degradation indicate a comparative robustness advantage among the evaluated models rather than strong absolute resistance to corruption.

4.2. Corruption- and Severity-Specific Robustness

Aggregate robustness did not produce a uniform ranking across corruption types. Table 3 summarizes Macro Dice averaged across the five severity levels.

Table 3: Mean Macro Dice Across Five Severity Levels.

Model	Clean Dice	mDice-C	Absolute Degradation	mRPD (%)
U-Net + ResNet18	0.5965	0.1912	0.4053	67.95
DeepLabV3+ + MobileNetV2	0.5710	0.2026	0.3684	64.52
SegFormer-B0	0.6247	0.2935	0.3312	53.02

SegFormer-B0 retained the highest mean defect Dice under Brightness Increase, Contrast Reduction, and Gaussian Blur. Brightness Increase was the least damaging corruption for all three models, whereas Contrast Reduction and Gaussian Blur produced much stronger degradation. Gaussian Noise yielded a different ordering: U-Net obtained the highest mean Dice (0.1272), followed by DeepLabV3+ (0.1197) and SegFormer-B0 (0.1116). SegFormer-B0's aggregate advantage therefore did not imply superior defect localization under every corruption.

The severity trajectories in Figure 2 further show that degradation depended on both corruption type and intensity. Brightness Increase produced the most gradual decline; at S5, SegFormer-B0 retained a Dice of 0.3894, compared with 0.1411 for U-Net and 0.1506 for DeepLabV3+. Contrast Reduction and Gaussian Blur caused much steeper deterioration. At Contrast Reduction S5, U-Net and DeepLabV3+ reached zero Dice (0), while SegFormer-B0 retained only 0.0242. At Gaussian Blur S5, the corresponding values were 0, 0.0156, and 0.0436.

Gaussian Noise showed a distinct early-collapse pattern. At S1, U-Net achieved a Dice of 0.2900 and SegFormer-B0 0.2719, whereas DeepLabV3+ had already fallen to 0.1267. At subsequent severity levels, all models entered a low-performance regime; small non-monotonic fluctuations occurred after substantial performance loss and are more consistent with a floor effect than recovery. These trajectories demonstrate why corruption robustness should be interpreted by corruption type and severity rather than from a single aggregate score.

Descriptive defect-type results were also heterogeneous. SegFormer-B0 achieved the highest *mDice-C* for blowhole, break, and crack, whereas DeepLabV3+ was numerically highest for fray and uneven. Because crack and fray contained only five and four independent scenes, respectively, these patterns were not used for inferential architecture ranking.

4.3. False-Alarm Failure Modes on Defect-Free Scenes

Evaluation of the 150 defect-free scenes revealed behavior not captured by defect Dice. Even under clean conditions, FPAR and IFAR produced different rankings. U-Net achieved the lowest clean IFAR (0.280), compared with 0.493 for SegFormer-B0 and 0.700 for DeepLabV3+. In contrast, SegFormer-B0 achieved the lowest clean mean FPAR (0.00167), followed by DeepLabV3+ (0.00314) and U-Net (0.00676). False-alarm frequency and false-positive spatial extent therefore represent distinct properties.

Among the four evaluated corruptions, Gaussian Noise produced the clearest false-positive failure mode. Contrast Reduction and Gaussian Blur could substantially reduce defect Dice without producing comparably extensive false-positive regions, indicating failures dominated more by loss or suppression of defect prediction. Gaussian Noise instead combined deterioration of defect localization with increasing false-positive segmentation on normal surfaces.

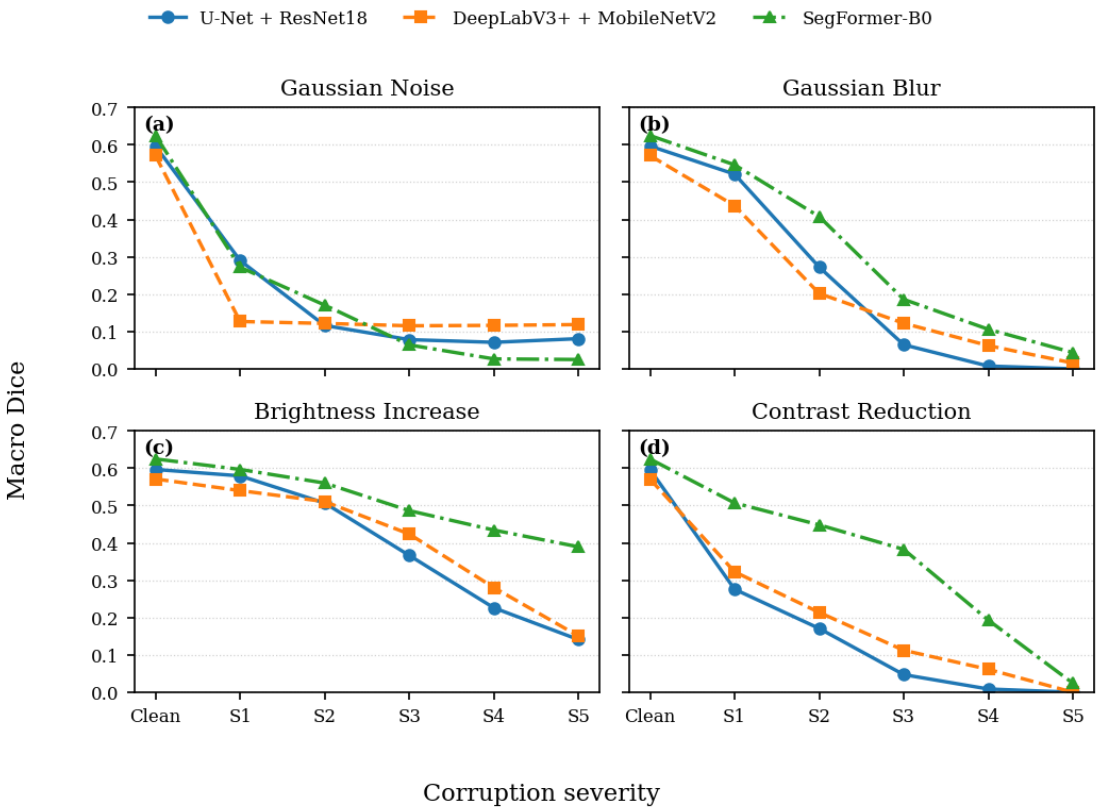

Figure 2: Severity-dependent Macro Dice under (a) Gaussian Noise, (b) Gaussian Blur, (c) Brightness Increase, and (d) Contrast Reduction. The clean condition is included as the reference baseline before severity levels S1–S5.

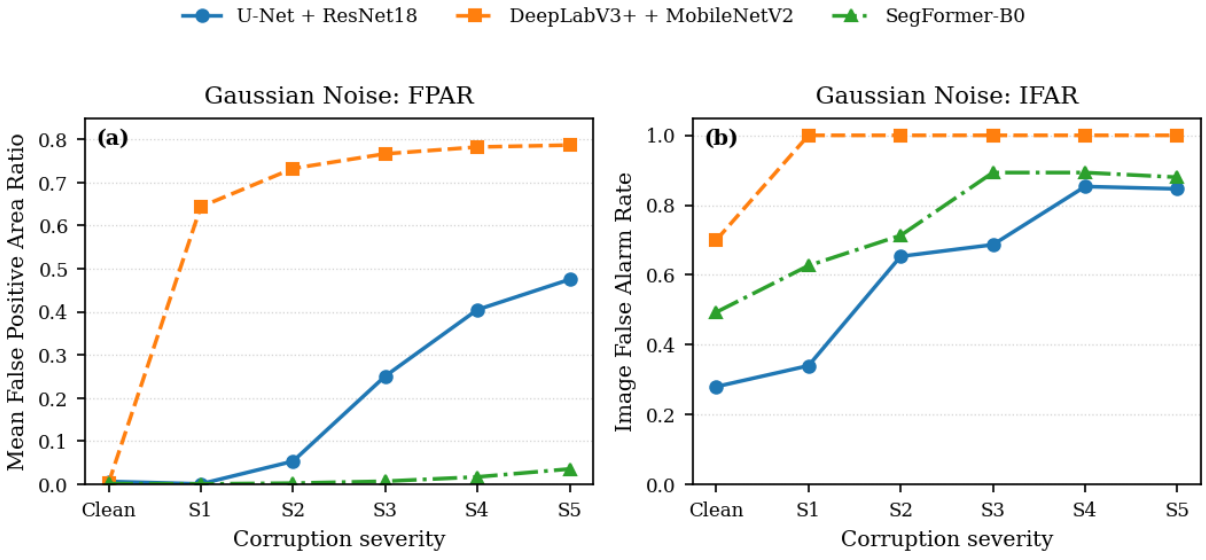

Figure 3: False-positive behavior on defect-free scenes under Gaussian Noise: (a) mean False Positive Area Ratio (FPAR) and (b) Image.

As shown in Figure 3, DeepLabV3+ exhibited the most abrupt response. Mean FPAR increased from 0.00314 under clean conditions to 0.6438 at Gaussian Noise S1, while IFAR reached 1.000 and remained there through S5. Mean FPAR increased further to 0.7869 at S5. Thus, even the mildest evaluated noise induced extensive over-segmentation rather than isolated false-positive pixels.
U-Net exhibited a more progressive transition. Mean FPAR was only 0.00104 at S1 but increased to 0.0537, 0.2510, 0.4052, and 0.4753 from S2 through S5; IFAR reached 0.847 at S5. SegFormer-B0 showed a markedly smaller spatial response: mean FPAR increased from 0.00118 at S1 to only 0.03571 at S5. However, its IFAR still reached 0.880 at S5. SegFormer-B0 therefore did not eliminate false alarms; rather, its false-positive regions remained substantially more spatially constrained.
The contrast at Gaussian Noise S5 is particularly informative. DeepLabV3+ incorrectly labeled approximately 78.7% of the valid normal-tile area as defective on average, compared with approximately 47.5% for U-Net and 3.6% for SegFormer-B0. At the same time, all three architectures exhibited poor defect Dice, and SegFormer-B0 did not achieve the highest mean defect Dice under Gaussian Noise. Defect-overlap performance alone would therefore provide an incomplete account of the observed failure. Joint interpretation of Dice, FPAR, and IFAR distinguishes loss of defect localization from frequent image-level alarms and extensive false-positive segmentation.

4.4. Interpretation, Practical Implications, and Limitations

The results indicate that architecture-associated differences became more apparent under degraded inputs than under clean evaluation. SegFormer-B0's aggregate advantage was driven primarily by better preservation of defect Dice under Brightness Increase, Contrast Reduction, and Gaussian Blur. Its hierarchical multiscale representation provides one plausible interpretation, consistent with prior reports of favorable robustness for some Transformer-based vision models. However, the present experiment cannot identify self-attention, hierarchical representation, overlapping patch embedding, or another component as the causal mechanism. The models differ simultaneously in encoder and decoder design, feature aggregation, parameterization, and pretrained representation. The findings should therefore be interpreted as architecture-associated robustness patterns rather than evidence of general Transformer superiority.
For industrial inspection, the results support combining three complementary evaluation components: clean segmentation performance, corruption-induced degradation, and false-positive behavior on defect-free scenes. Clean Dice establishes reference predictive capability, corruption testing quantifies how that capability changes under controlled degradation, and FPAR and IFAR reveal failure behavior not captured by foreground-overlap metrics. None of the evaluated models can be considered deployment-ready from these experiments alone; even SegFormer-B0 lost approximately 53% of its clean Dice on average across the corrupted conditions.
Several limitations constrain generalization. First, the experiments used a single public dataset containing 199 independent scenes and only 49 defect-bearing scenes, with substantially imbalanced defect categories. Second, the four isolated synthetic corruptions do not represent the full range of real production-domain shifts or compound degradations. Third, only three architecture configurations were evaluated using a fixed threshold of 0.5 without corruption-aware training, robustness-specific adaptation, or calibration analysis. Finally, the experiment neither isolates causal architectural mechanisms nor evaluates operational factors such as uncertainty, latency, throughput, and production-specific error costs. Future work should therefore examine documented real acquisition shifts, larger and more balanced datasets, corruption-aware training and calibration, and controlled architectural ablations while preserving scene-independent evaluation and separation between model development and final robustness testing.
Overall, the findings support the central proposition that clean segmentation performance alone provides an incomplete characterization of reliability under the evaluated conditions. No significant model effect was detected for clean Macro Dice, whereas significant differences emerged for mean corrupted Dice. Moreover, Gaussian Noise showed that models with similarly poor defect overlap can exhibit qualitatively different failure modes on normal surfaces. Corruption-aware evaluation therefore provides reliability information that is not apparent from clean benchmark performance alone.

5. Conclusion

This study shows that clean segmentation performance alone provides an incomplete characterization of reliability for magnetic-tile defect segmentation. No significant model effect was detected for the predefined clean Macro Dice endpoint ($p = 0.1693$), whereas a significant difference emerged for mean corrupted Dice ($p = 0.001118$). SegFormer-B0 achieved the strongest aggregate corruption performance, with an mDice-C of 0.2935 and the lowest mRPD of 53.02%, and its paired mean corrupted Dice was significantly higher than those of U-Net + ResNet18 and DeepLabV3+ + MobileNetV2. However, this aggregate advantage was not universal across corruption types; in particular, SegFormer-B0 did not achieve the highest defect Dice under Gaussian Noise.
The defect-free analysis further showed that models with similarly poor defect-overlap performance can exhibit substantially different failure modes. At Gaussian Noise S5, mean FPAR reached approximately 0.7869 for DeepLabV3+ and 0.4753 for U-Net, compared with only 0.0357 for SegFormer-B0. Nevertheless, SegFormer-B0 still exhibited an IFAR of 0.880, indicating that its distinguishing behavior was not the elimination of false alarms but the substantially smaller spatial extent of false-positive predictions. These findings demonstrate that defect Dice, false-alarm frequency, and false-positive spatial extent provide complementary information and should be interpreted jointly when assessing segmentation reliability.
Overall, corruption-aware evaluation revealed reliability differences and failure behavior that were not apparent from clean benchmark performance alone. None of the evaluated models can be considered deployment-ready from these experiments, because substantial degradation remained and the evaluated corruptions were controlled synthetic stress tests rather than complete representations of production-domain shifts. Future work should evaluate documented real acquisition changes, larger and more balanced collections of independent scenes, and robustness-enhancing strategies such as corruption-aware training, calibration, and uncertainty-aware prediction while preserving scene-independent evaluation and separation between model development and final robustness testing.

Article Information

Author's Contributions: Andi Shafira Dyah Kurniasari conceived the research idea, designed the study, developed the models, conducted the experiments and testing, and prepared the original manuscript. Karisa Ardelia Hanifah and Khalis Sofi contributed to manuscript preparation, review, and revision. All authors reviewed and approved the final version of the manuscript.

Acknowledgments: The authors would like to thank the School of Computing, Telkom University, for providing academic support for this research.

Conflict of Interest: The authors declare that there is no conflict of interest regarding the publication of this paper.

Support Section: This research received no specific grant from any funding agency, institution, or organization.

Ethical Approval: Not applicable. This study did not involve human participants, animals, or personally identifiable data. The experiments were conducted using a publicly available image dataset.

Availability: The dataset used in this study is publicly available through the sources cited in the manuscript. Additional experimental materials may be made available by the corresponding author upon reasonable request.

Artificial Intelligence Statement: Generative artificial intelligence tools were used only to assist with language refinement and manuscript formatting. The research design, experiments, data analysis, interpretation of results, and scientific conclusions were performed and verified by the authors. The authors reviewed the AI-assisted content and take full responsibility for the final manuscript.

Plagiarism Statement: This article has been checked for plagiarism using Turnitin.

References

[1] Y. Huang, C. Qiu, and K. Yuan, "Surface defect saliency of magnetic tile," *Vis. Comput.*, vol. 36, no. 1, pp. 85–96, 2020, doi: 10.1007/s00371-018-1588-5.
[2] F. Luo, Y. Cui, X. Wang, Z. Zhang, and Y. Liao, "Adaptive rotation attention network for accurate defect detection on magnetic tile surface," *Mathematical Biosciences and Engineering*, vol. 20, no. 9, pp. 17554–17568, 2023, doi: 10.3934/mbe.2023779.
[3] C. Jiang *et al.*, "MT-U2Net: Lightweight detection network for high-precision magnetic tile surface defect localization," *Mater. Today Commun.*, vol. 41, p. 110480, 2024, doi: 10.1016/j.mtcomm.2024.110480.
[4] C. Kamann and C. Rother, "Benchmarking the Robustness of Semantic Segmentation Models with Respect to Common Corruptions," *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 462–483, 2021, doi: 10.1007/s11263-020-01383-2.
[5] L. Xie, X. Xiang, H. Xu, L. Wang, L. Lin, and G. Yin, "FFCNN: A Deep Neural Network for Surface Defect Detection of Magnetic Tile," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 4, pp. 3506–3516, 2021, doi: 10.1109/TIE.2020.2982115.
[6] P. de Jorge, R. Volpi, P. H. S. Torr, and G. Rogez, "Reliability in Semantic Segmentation: Are We on the Right Track?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7173–7182. doi: 10.1109/CVPR52729.2023.00693.
[7] H. Liu, G. Wang, Q. Li, and N. Wang, "Surface defect segmentation of magnetic tiles based on cross self-attention module," *Journal of Intelligent & Fuzzy Systems*, vol. 45, no. 6, pp. 9523–9532, 2023, doi: 10.3233/JIFS-232366.
[8] X. Jiang *et al.*, "CINFormer: Transformer network with multi-stage CNN feature injection for surface defect segmentation," arXiv, 2023. doi: 10.48550/arXiv.2309.12639.
[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.
[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, 2018, pp. 801–818. doi: 10.1007/978-3-030-01234-2_49.
[12] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.
[13] D. Hendrycks and T. Dietterich, "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations," in *International Conference on Learning Representations (ICLR)*, 2019. [Online]. Available: https://arxiv.org/abs/1903.12261
[14] "Magnetic Tile Surface Defects." Accessed: Aug. 15, 2026. [Online]. Available: https://www.kaggle.com/datasets/alex000kim/magnetic-tile-surface-defects
[15] Y. Huang, C. Qiu, Y. Guo, X. Wang, and K. Yuan, "Surface Defect Saliency of Magnetic Tile," *IEEE International Conference on Automation Science and Engineering*, vol. 2018-August, pp. 612–617, Dec. 2018, doi: 10.1109/COASE.2018.8560423.