Faithful but Not Plausible? A Comparative Evaluation of Explainable AI Methods for Bridge Deck Crack Detection

Reihaneh Samsami, Ph.D., P.E.*
Assistant Professor, Department of Civil Engineering
University of New Haven, 300 Boston Post Road, West Haven, CT 06516
Email: rsamsami@newhaven.edu

Mohamad Nassar, Ph.D.
Associate Professor
Department of Electrical & Computer Engineering and Computer Science
University of New Haven, West Haven, CT, 06516
Email: mnassar@newhaven.edu

Total Number of Pages: 14

*Corresponding Author

ABSTRACT

Objectives: Automated crack detection using deep learning is increasingly proposed for bridge deck inspection, but transportation agencies hesitate to adopt models whose decisions cannot be examined. Explainable AI (XAI) saliency methods are offered as a remedy, yet are rarely evaluated for both faithfulness to the model and plausibility to a human inspector. This study quantifies both, and their relationship, for crack detection.

Methods: Four post-hoc explanation methods (Grad-CAM^{++}, Eigen-CAM, Score-CAM, and SHAP) were compared across three lightweight classifiers (ResNet-18, MobileNetV3-Small, and ViT-Tiny) fine-tuned on the SDNET2018 bridge deck subset. Faithfulness was measured with deletion and insertion curves, localization against pixel-level DeepCrack masks, and plausibility through blinded expert ratings by a professional engineer.

Findings: All models exceeded 90 percent accuracy. Gradient- and activation-based explanations degraded sharply on the vision transformer (ViT), producing negative faithfulness gaps, while perturbation-based Score-CAM partially recovered. Most consequentially, SHAP was among the most faithful methods on convolutional models by machine metrics yet was rated least plausible by the expert, because its scattered pixel-level attributions did not align with the crack. Faithfulness and plausibility were only moderately correlated overall.

Novelty: This study is among the first infrastructure-inspection studies to jointly evaluate explanation faithfulness, localization, and expert plausibility, and to document a direct dissociation between machine-measured faithfulness and human-judged plausibility for crack detection.

Practical Applications: Agencies should not select XAI methods for inspection on machine faithfulness metrics alone; a lightweight expert-rating step during validation surfaces explanations that are technically faithful but practically unusable, and explanation method and model architecture must be chosen together.

INTRODUCTION

Bridge deck condition assessment remains one of the most resource-intensive responsibilities of state departments of transportation (DOTs) (Dorafshan and Maguire 2018). Cracking is among the earliest visible indicators of deck deterioration, and its timely detection supports maintenance prioritization and extends service life. Over the past decade, convolutional neural networks (CNNs) and more recently, vision transformers (ViTs) have been applied to automate crack detection from imagery collected by uncrewed aerial systems (UASs) and other platforms (Cha et al. 2017; Spencer et al. 2019; Bhupesh et al. 2026; Pokhrel et al. 2024) with reported accuracies that rival or exceed manual review under controlled conditions.

Despite this technical progress, their adoption within transportation agencies has lagged (Samsami 2024). A recurring obstacle is trust: an inspector or bridge engineer asked to act on a model's output reasonably wants to understand why a given region was flagged. A model that returns only a label or a bounding box, however accurate, offers no basis for the professional judgment that inspection ultimately requires. Explainable AI (XAI) methods, which produce visual saliency maps indicating the image regions most responsible for a prediction (Zhou et al. 2016; Selvaraju et al. 2017) are frequently proposed to close this gap.

However, the XAI literature in infrastructure inspection has a consistent weakness: explanations are typically displayed as qualitative heatmaps and declared satisfactory on visual inspection by the authors, without rigorous evaluation. Two distinct questions are routinely conflated. The first is faithfulness: does the explanation actually reflect the evidence the model used? The second is plausibility: does the explanation correspond to what a domain expert would consider the relevant visual feature? These are not the same property, and an explanation can satisfy one while failing the other. A method may faithfully trace a model's internal reasoning yet highlight scattered pixels that mean nothing to an inspector; conversely, a visually clean heatmap may reassure a human while poorly reflecting the model's true basis for decision.

This study evaluates both properties directly. Using lightweight classifiers fine-tuned on bridge deck imagery, it generates explanations from four widely used post-hoc methods and assess them along three axes of (1) quantitative faithfulness, (2) quantitative localization against ground-truth crack masks, and (3) expert plausibility through blinded ratings. The central contribution is empirical evidence that faithfulness and plausibility can diverge, and that the divergence has direct implications for how agencies should select and deploy explainable crack-detection systems.

The contributions of this paper are fourfold. First, it provides a head-to-head comparison of four explanation methods across three architectural families on a common bridge deck crack-detection task, holding the evaluation image set fixed so that methods are directly comparable. Second, it evaluates explanations on three complementary criteria, faithfulness, localization, and human plausibility, rather than relying on any single measure. Third, it documents a concrete dissociation in which the method judged best by a standard faithfulness metric is judged worst by a domain expert. Fourth, it translates these findings into actionable guidance for transportation agencies evaluating automated inspection tools.

BACKGROUND AND RELATED WORK

Automated Crack Detection for Bridge Decks

Deep learning approaches to concrete crack detection fall broadly into patch classification, object detection, and pixel-level segmentation (Cha et al. 2017; Zhang et al. 2016; Liu et al. 2019; Khatry et al. 2026), and have been shown to outperform traditional edge-detection techniques for concrete crack identification (Dorafshan et al. 2018b). Patch classification, adopted here, labels fixed-size image regions as cracked or non-cracked and is well suited to lightweight architectures deployable in resource-constrained inspection settings. The SDNET2018 dataset (Dorafshan et al. 2018), which includes a dedicated bridge deck subset with hairline cracks, has become a common benchmark; its fine cracks and visual obstructions make it a demanding one (Dorafshan et al. 2018).

Post-hoc Explanation Methods

Class Activation Mapping (Zhou et al. 2016) and its gradient-weighted successors, including Grad-CAM (Selvaraju et al. 2017) and Grad-CAM^{++} (Chattopadhay et al. 2018), attribute a prediction to spatial regions by combining feature-map activations with gradient information. Eigen-CAM (Muhammad and Yeasin 2020) instead uses the principal components of activations and requires no class-specific gradients. Score-CAM (Wang et al. 2020) replaces gradients with a perturbation-based procedure, masking the input with upsampled activation maps and measuring the effect on the prediction, at substantially higher computational cost. SHAP (Lundberg and Lee 2017), grounded in cooperative game theory, attributes the prediction to individual input features and is often regarded as theoretically principled but computationally demanding; related model-agnostic methods such as LIME (Ribeiro et al. 2016) share this attribution philosophy. A known complication is that gradient-based CAM variants were designed for convolutional networks and can behave unpredictably on transformer architectures (Dosovitskiy et al. 2021), where spatial structure is represented differently.

Evaluating Explanations

Faithfulness metrics such as deletion and insertion (Petsiuk et al. 2018) perturb an image in order of saliency and track the model's response: a faithful explanation causes a rapid probability drop as important pixels are removed (low deletion area under the curve) and a rapid recovery as they are introduced (high insertion area). Localization metrics compare saliency against ground-truth annotations. A broader literature cautions that saliency methods can fail basic sanity checks and that faithfulness is not guaranteed by visual appeal (Adebayo et al. 2018; Jacovi and Goldberg 2020). Plausibility, by contrast, is inherently human-centric and cannot be captured by machine metrics alone; it requires practitioner judgment (Doshi-Velez and Kim 2017). Prior inspection studies seldom combine all three, which is the gap this work addresses.

METHODOLOGY

Datasets

Two open-source datasets were used. Model training and faithfulness evaluation used the bridge deck subset of SDNET2018 (Dorafshan et al. 2018), comprising cracked and non-cracked deck image patches. Localization evaluation used the DeepCrack dataset (Liu et al. 2019), which provides pixel-level crack masks; applying the SDNET-trained models to DeepCrack imagery additionally serves as a cross-dataset generalization check. Crack-centered patches were extracted from DeepCrack images and retained only where crack pixels were present, and masks were dilated to tolerate minor localization offset.

Models and Training

Three ImageNet-pretrained architectures (Deng et al. 2009) were fully fine-tuned for binary crack classification: ResNet-18 (He et al. 2016) and MobileNetV3-Small (Howard et al. 2019), both convolutional, and ViT-Tiny, a vision transformer (Dosovitskiy et al. 2021). To address class imbalance without discarding data, all non-cracked patches were retained and balanced batches were formed with a weighted sampler. Training used the AdamW optimizer with a one-cycle learning-rate schedule, label smoothing, and data augmentation. The decision threshold for each model was selected on the validation set. Because the classifier is not the contribution of this study, models were required only to be competent; all three exceeded 90 percent test accuracy, as reported below.

Explanation Generation

For each model, explanations were generated for the cracked class on test patches the model correctly identified as cracked, reflecting the inspector-facing use case of explaining a positive flag. To ensure a fair comparison, faithfulness was evaluated on the set of test images that all three models correctly classified as cracked, so every method was assessed on an identical image set. Grad-CAM^{++}, Eigen-CAM, and Score-CAM were applied to all three architectures; SHAP (via a gradient explainer) was applied to the two convolutional models only, as gradient-based SHAP is unreliable on the transformer.

Explanation Methods

This section defines the four post-hoc explanation methods evaluated in this study. All four produce a saliency map that assigns an importance value to each spatial location of the input image with respect to the predicted cracked class; they differ in how that importance is computed. Throughout, let the input image be x, let a trained network produce class score S_c for the cracked class c, and let A^k denote the k-th feature map (activation channel) of a chosen convolutional layer, with A_{ij}^k its value at spatial location (i, j).

Gradient-weighted Class Activation Mapping (Grad-CAM) forms a class-discriminative saliency map by weighing each feature map by the average gradient of the class score flowing into it. The importance weight of feature map k is the global-average-pooled gradient **(Equation 1)**:

$$\alpha_k^c = \frac{1}{Z}\sum_{i,j}\frac{\partial S_c}{\partial A_{ij}^k} \qquad (1)$$

Here Z is the number of spatial locations. The Grad-CAM saliency map is the ReLU-rectified weighted sum of feature maps, retaining only regions with a positive influence on the cracked class **(Equation 2)**:

$$L_{Grad-CAM} = ReLU(\sum_k \alpha_k^c A^k) \qquad (2)$$

Grad-CAM^{++} (Chattopadhay et al. 2018) generalizes Grad-CAM by replacing the global gradient average with a weighted combination of positive partial derivatives, which improves localization when a class appears at multiple locations, as with several cracks in one patch. The feature-map weights become **(Equation 3)**:

$$\alpha_k^c = \sum_{ij} a_{ij}^{kc} ReLU(\frac{\partial S_c}{\partial A_{ij}^k}) \qquad (3)$$

where the coefficients a_{ij}^{kc} are computed from second and third-order derivatives of the class score, giving pixels with stronger local response proportionally more influence. The final map is formed as in Grad-CAM by a ReLU-rectified weighted sum of feature maps.

Eigen-CAM (Muhammad and Yeasin 2020) requires no class-specific gradients. It projects the activations of the chosen layer onto their first principal component, so the saliency map captures the dominant direction of variation in the feature representation. Writing the layer activations as a matrix A (locations by channels), the map is the projection onto the leading right singular vector **(Equation 4)**:

$$L_{Eigen-CAM} = Av_1 \qquad (4)$$

where v_1 is the first eigenvector of A^TA (equivalently, the first right singular vector of A from its singular value decomposition). Because it ignores the class score, Eigen-CAM is fast but not class-discriminative, a property whose consequences appear in the results.

Score-CAM (Wang et al. 2020) removes the dependence on gradients entirely and instead measures the contribution of each feature map by its effect on the model output. Each feature map A^k is upsampled to the input size and normalized to form a mask, which is multiplied with the input; the network is then run on the masked image, and the resulting increase in the cracked-class score becomes that map's weight **(Equation 5)**:

$$w_k^c = softmax(f(x \odot Up(A^k)))_c \qquad (5)$$

Here Up(·) denotes upsampling to the input resolution, $\odot$ is element-wise multiplication, and f is the network. The saliency map is the ReLU-rectified weighted sum of feature maps using these forward-pass weights. Because Score-CAM needs one network evaluation per feature map, it is far more computationally expensive than the gradient methods, a cost quantified in the results **(Equation 6)**.

$$L_{Score-CAM} = ReLU(\sum_k w_k^c A^k) \qquad (6)$$

SHAP (Lundberg and Lee 2017) takes a different route grounded in cooperative game theory. It treats the model prediction as a payout to be distributed among input features, and assigns each feature the Shapley value, its average marginal contribution across all possible orderings in which features are added to the model. For a feature i and model f evaluated on feature subsets S drawn from the full set F, the attribution is **(Equation 7)**:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (7)$$

In this study SHAP is applied at the pixel (input-feature) level to the two convolutional models via a gradient-based approximation, and the per-pixel Shapley magnitudes are aggregated into a saliency map comparable to the CAM outputs. Shapley values are the only attribution here with axiomatic guarantees (local accuracy, consistency), but as the results show, this theoretical faithfulness does not translate into visual plausibility for crack inspection.

Evaluation Protocol

Faithfulness was quantified using deletion and insertion area-under-the-curve (AUC) computed over 30 perturbation steps, with a blurred image as the baseline. This study additionally report the faithfulness gap (insertion AUC minus deletion AUC) which is positive for a faithful explanation. Localization was measured with the pointing game (whether the single most-salient pixel falls within the dilated crack mask) and the fraction of saliency energy contained within the mask. Plausibility was assessed by a civil engineer who rated blinded explanation overlays, without knowledge of which model or method produced each, on a three-point scale: 3 (the highlighted region fully captures the crack), 2 (partial capture), and 1 (does not capture the crack). Runtime per explanation was recorded as a measure of deployment practicality.

RESULTS

Model Performance

Table 1 reports test-set performance at the validation-selected threshold. All three models exceeded 90 percent accuracy. The models were precision-favoring, reflecting the selected operating point; in a bridge-safety context the relative cost of missed versus false detections would guide threshold selection, a point returned to in the discussion.

TABLE 1 Classification Performance on the SDNET2018 Bridge Deck Test Set

Model	Accuracy	Precision	Recall	F1-score
ResNet-18	0.920	0.784	0.647	0.709
MobileNetV3-Small	0.921	0.827	0.602	0.697
ViT-Tiny	0.906	0.750	0.563	0.643

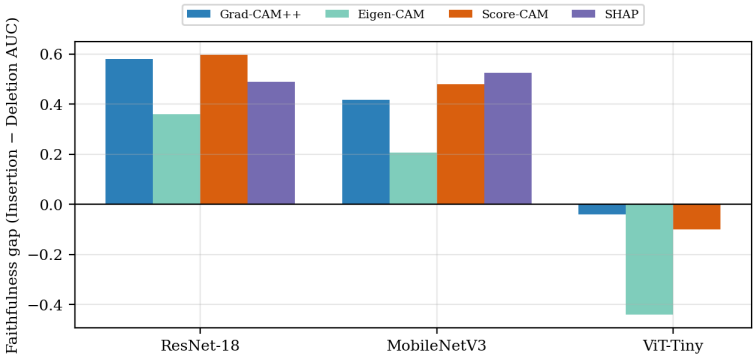
Faithfulness

Table 2 and **Figure 1** report faithfulness. On the convolutional models, all methods achieved positive faithfulness gaps, with Score-CAM and SHAP generally strongest. The transformer told a different story: on ViT-Tiny, Eigen-CAM produced a strongly negative gap and the other gradient-based methods hovered near zero, indicating that these explanations were not faithful to the transformer's predictions. Score-CAM, which relies on perturbation rather than gradients, was the only method to remain informative on the transformer.

TABLE 2 Faithfulness: Deletion AUC (lower is better), Insertion AUC (higher is better), and Gap

Model	Method	Deletion ↓	Insertion ↑	Gap ↑
ResNet-18	Grad-CAM^{++}	0.240	0.820	0.580
	Eigen-CAM	0.349	0.708	0.359
	Score-CAM	0.248	0.846	0.597
	SHAP	0.428	0.918	0.490
MobileNetV3-Small	Grad-CAM^{++}	0.471	0.888	0.417
	Eigen-CAM	0.570	0.776	0.206
	Score-CAM	0.425	0.905	0.480
	SHAP	0.355	0.881	0.526
ViT-Tiny	Grad-CAM^{++}	0.671	0.631	-0.040
	Eigen-CAM	0.835	0.395	-0.439
	Score-CAM	0.687	0.587	-0.101

Figure 1 Faithfulness gap (insertion minus deletion AUC) by model and method. Negative values on ViT-Tiny indicate unfaithful explanations.



Localization

Table 3 reports localization against DeepCrack ground-truth masks. On the convolutional models, explanations localized moderately well, with the most-salient pixel falling inside the crack mask in roughly half to two-thirds of cases. On the transformer, gradient-based localization was poor, consistent with the faithfulness results, while Score-CAM again performed comparatively better. Because these images are drawn from a different dataset than training, the results also indicate reasonable cross-dataset transfer of the learned crack features.

TABLE 3 Localization Against DeepCrack Ground-Truth Masks (higher is better)

Model	Method	Pointing Game	Energy in Mask
ResNet-18	Grad-CAM^{++}	0.533	0.294
	Eigen-CAM	0.513	0.342
	Score-CAM	0.500	0.259
MobileNetV3-Small	Grad-CAM^{++}	0.673	0.402
	Eigen-CAM	0.607	0.437
	Score-CAM	0.540	0.321
ViT-Tiny	Grad-CAM^{++}	0.353	0.307
	Eigen-CAM	0.113	0.078
	Score-CAM	0.620	0.487

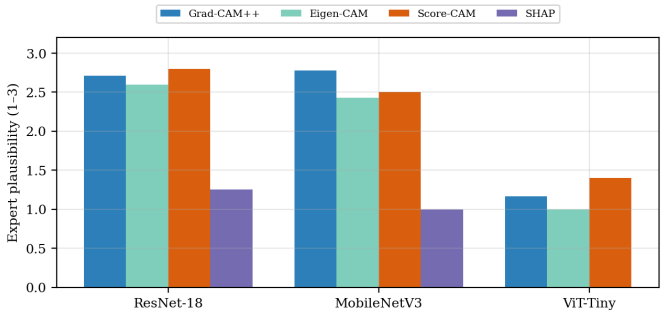
Expert Plausibility

A professional civil engineer rated 71 blinded explanation overlays. **Table 4** reports mean plausibility by method and model on the three-point scale. Two patterns are notable. First, on the transformer, explanations were rated close to the floor of the scale, corroborating the machine metrics with human judgment. Second, and most striking, SHAP received the lowest plausibility of any method (mean 1.13) despite being among the most faithful on the machine metrics. The expert consistently judged that SHAP's scattered pixel-level attributions did not capture the crack, even when those attributions were faithful to the model. No overlays were flagged as actively misleading.

TABLE 4 Expert Plausibility Ratings (1–3 scale; higher is better)

Method	ResNet-18	MobileNetV3	ViT-Tiny	Method Mean
Grad-CAM^{++}	2.71 (n=7)	2.78 (n=9)	1.17 (n=6)	2.32
Eigen-CAM	2.60 (n=5)	2.43 (n=7)	1.00 (n=4)	2.12
Score-CAM	2.80 (n=5)	2.50 (n=8)	1.40 (n=5)	2.28
SHAP	1.25 (n=8)	1.00 (n=7)	—	1.13

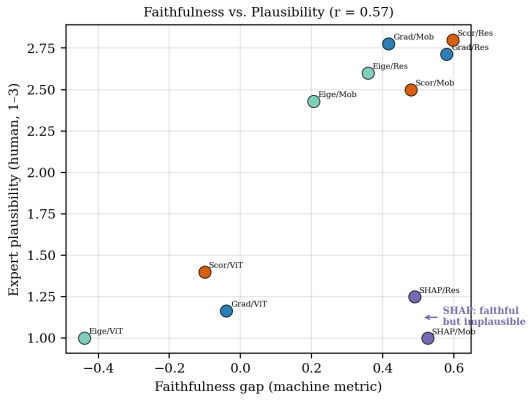
Figure 2 Expert plausibility by model and method.



Faithfulness Versus Plausibility

Figure 3 plots expert plausibility against the machine faithfulness gap for every model–method combination. Across all points the two are positively but only moderately correlated, indicating substantial agreement in aggregate but meaningful exceptions. The clearest exception is SHAP on the convolutional models, which occupies the high-faithfulness, low-plausibility region of the plot: faithful to the model, yet not recognizable to the expert as capturing the crack. This dissociation is the central empirical finding of the study.

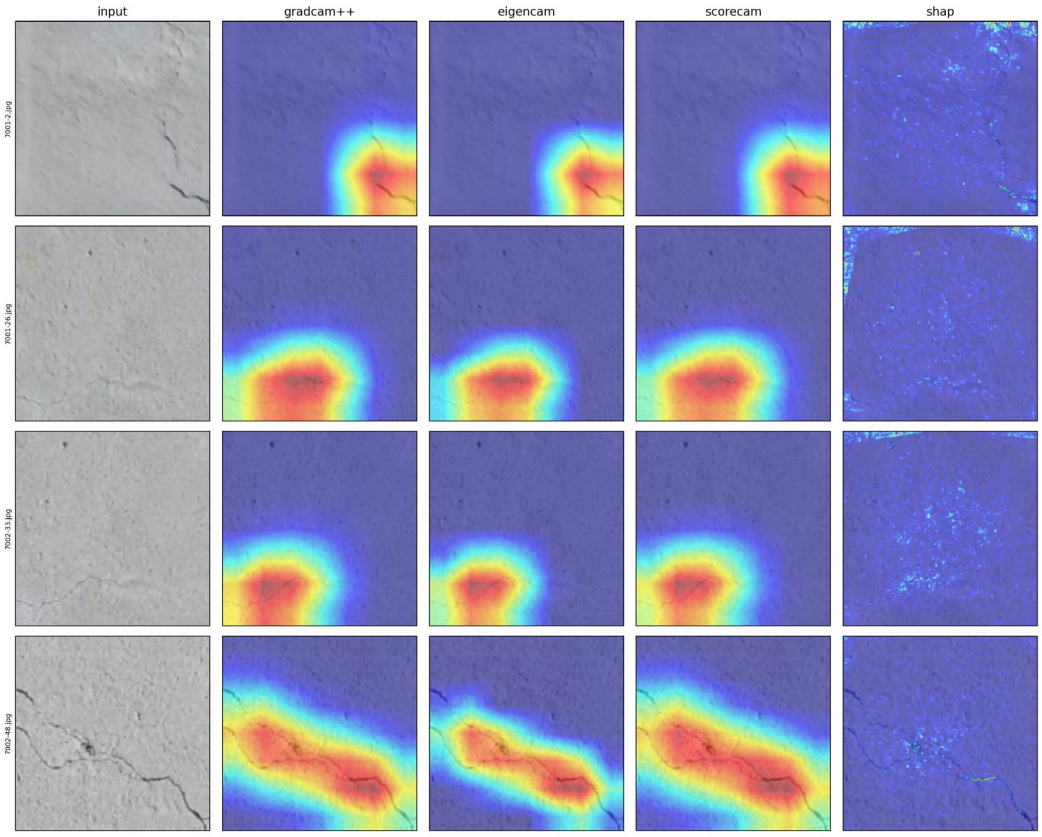
Figure 3 Expert plausibility versus machine faithfulness gap. SHAP is faithful yet implausible.



Qualitative Comparison

Figure 4 shows representative overlays for the convolutional models across methods. Grad-CAM^{++}, Eigen-CAM, and Score-CAM produce spatially coherent heat concentrated along the crack path, which the expert generally rated as capturing the defect. In contrast, SHAP attributions are dispersed across many small pixel clusters; although these can be faithful to the model, they lack the contiguous spatial structure a human associates with a crack, which explains the low plausibility ratings despite strong faithfulness scores.

Figure 4 Representative explanation overlays for ResNet-18 across methods, illustrating the contrast between spatially coherent CAM heatmaps and dispersed SHAP attributions.



Computational Cost

Table 5 reports mean runtime per explanation. Grad-CAM^{++} and Eigen-CAM were effectively instantaneous, while Score-CAM was roughly one to two orders of magnitude slower because it requires many forward passes per image. This cost is directly relevant to deployment: the method that proved most robust on the transformer is also the most expensive, though still well under one second per image and therefore viable for batch report generation if not real-time overlay.

TABLE 5 Mean Runtime per Explanation (seconds per image)

Model	Grad-CAM^{++}	Eigen-CAM	Score-CAM
ResNet-18	0.0055	0.0151	0.5284
MobileNetV3-Small	0.0047	0.0337	0.2810
ViT-Tiny	0.0048	0.0138	0.2627

DISCUSSION

Faithfulness and Plausibility Are Not Interchangeable

The most important practical implication is that an agency cannot select an explanation method for crack detection on the basis of a single machine metric. SHAP would be the recommended method if faithfulness alone were the criterion, yet the expert found its explanations largely uninterpretable for the inspection task. Selecting XAI purely by machine faithfulness could therefore field a system whose explanations are technically sound but practically unusable, undermining the very trust the explanations were meant to build.

Architecture Matters for Explanation Choice

The results also show that explanation method and model architecture cannot be chosen independently. For the transformer, the common gradient-based CAM methods were both unfaithful and implausible; only the perturbation-based Score-CAM produced usable explanations, at meaningfully higher computational cost. Agencies considering transformer-based inspection models should budget for this cost or reconsider the architecture if lightweight explanation is a requirement.

Implications for Agency Adoption

For transportation agencies weighing automated inspection, these findings argue for evaluating explanations the way they will be used: reviewed by inspectors. A brief expert-rating exercise, of the kind conducted here, is inexpensive relative to model development and surfaces failure modes that machine metrics miss. Explanation quality should be a procurement and validation criterion, not an afterthought.

LIMITATIONS

Several limitations temper these findings. Plausibility was assessed by a single expert rater on a subset of the generated overlays, and coverage across some model–method combinations is limited; the plausibility results should therefore be read as preliminary rather than definitive, and inter-rater reliability could not be assessed. A planned extension will recruit multiple certified bridge inspectors, enabling agreement analysis and more robust plausibility estimates. The study also uses patch-level classification rather than segmentation or detection, evaluates on two specific datasets, and considers four explanation methods; other tasks, datasets, and methods may behave differently. Finally, deletion and insertion metrics are known to favor certain saliency distributions, which may partly explain SHAP's strong machine faithfulness, and this interaction warrants further study.

CONCLUSIONS

This study evaluated four explainable AI methods across three lightweight architectures for bridge deck crack detection, assessing faithfulness, localization, and expert plausibility together rather than in isolation. The results show that these properties can diverge: SHAP was among the most faithful methods by machine metrics yet was judged least plausible by a domain expert, and gradient-based explanations that were adequate on convolutional models failed on the transformer. For transportation agencies, the practical message is that explanation quality for inspection should be judged partly by practitioners, not by machine faithfulness metrics alone, and that explanation method and model architecture must be chosen together. Incorporating even a lightweight expert-rating step into model validation can prevent the deployment of explanations that are technically faithful but practically unhelpful.

ACKNOWLEDGEMENTS

The authors used Claude, an AI assistant developed by Anthropic, to support drafting of this paper. All study design decisions, interpretations, and conclusions are the responsibility of the author.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: R. Samsami, M. Nassar; data collection and analysis: R. Samsami; interpretation of results: R. Samsami, M. Nassar;

manuscript preparation: R. Samsami. Both authors reviewed the results and approved the final version of the manuscript.

REFERENCES

Adebayo, Julius, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. "Sanity checks for saliency maps." Advances in neural information processing systems 31 (2018).

Bhupesh, Chand, Ayele Frezer, Pineiro-Dakers Ian, Samsami Reihaneh, and Chang Byungik. "Uncrewed aerial system (UAS) applications in bridge inspection: a comprehensive review of platforms, sensors, and operational effectiveness." Drones 10, no. 2 (2026): 144. https://doi.org/10.3390/drones10020144.

Cha, Young-Jin, Wooram Choi, and Oral Büyüköztürk. "Deep learning-based crack damage detection using convolutional neural networks." Computer-Aided Civil and Infrastructure Engineering 32, no. 5 (2017): 361-378. https://doi.org/10.1111/mice.12263.

Chattopadhay, Aditya, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." In 2018 IEEE winter conference on applications of computer vision (WACV), pp. 839-847. IEEE, 2018. https://doi.org/10.1109/WACV.2018.00097.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database." In 2009 IEEE conference on computer vision and pattern recognition, pp. 248-255. IEEE, 2009. https://doi.org/10.1109/CVPR.2009.5206848.

Dorafshan, Sattar, and Marc Maguire. "Bridge inspection: Human performance, unmanned aerial systems and automation." Journal of Civil Structural Health Monitoring 8, no. 3 (2018): 443-476.

Dorafshan, Sattar, Robert J. Thomas, and Marc Maguire. "Comparison of deep convolutional neural networks and edge detectors for image-based crack detection in concrete." Construction and Building Materials 186 (2018b): 1031-1045. https://doi.org/10.1016/j.conbuildmat.2018.08.011.

Dorafshan, Sattar, Robert J. Thomas, and Marc Maguire. "SDNET2018: An annotated image dataset for non-contact concrete crack detection using deep convolutional neural networks." Data in brief 21 (2018a): 1664-1668. https://doi.org/10.1016/j.dib.2018.11.015.

Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017). https://doi.org/10.48550/arXiv.1702.08608.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

He, Kaiming, 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.

Howard, Andrew, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang et al. "Searching for mobilenetv3." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 1314-1324. 2019.

Jacovi, Alon, and Yoav Goldberg. "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?." In Proceedings of the 58th annual meeting of the association for computational linguistics, pp. 4198-4205. 2020. https://doi.org/10.18653/v1/2020.acl-main.386.

Khatry, Kalyan, Sayda Elmi, and Reihaneh Samsami. "Automated Concrete Bridge Deck Inspection Using Unmanned Aerial System–Collected Data: A Deep Learning Approach." ASCE OPEN: Multidisciplinary Journal of Civil Engineering 4, no. 1 (2026): 04026004. https://doi.org/10.1061/AOMJAH.AOENG-0091.

Liu, Yahui, Jian Yao, Xiaohu Lu, Renping Xie, and Li Li. "DeepCrack: A deep hierarchical feature learning architecture for crack segmentation." Neurocomputing 338 (2019): 139-153. https://doi.org/10.1016/j.neucom.2019.01.036.

Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems* 30 (2017).

Muhammad, Mohammed Bany, and Mohammed Yeasin. "Eigen-cam: Class activation map using principal components." In 2020 international joint conference on neural networks (IJCNN), pp. 1-7. IEEE, 2020. https://doi.org/10.1109/IJCNN48605.2020.9206626.

Petsiuk, Vitali, Abir Das, and Kate Saenko. "Rise: Randomized input sampling for explanation of black-box models." arXiv preprint arXiv:1806.07421 (2018). https://doi.org/10.48550/arXiv.1806.07421.

Pokhrel, R., Samsami, R., Elmi, S., & Brooks, C. N. (2024). Automated concrete bridge deck inspection using unmanned aerial system (UAS)-collected data: A machine learning (ML) approach. Eng, 5(3), 1937-1960. https://doi.org/10.3390/eng5030103.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135-1144. 2016. https://doi.org/10.1145/2939672.2939778.

Samsami, Reihaneh. "A systematic review of automated construction inspection and progress monitoring (ACIPM): applications, challenges, and future directions." CivilEng 5, no. 1 (2024): 265-287. https://doi.org/10.3390/civileng5010014.

Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization." In Proceedings of the IEEE international conference on computer vision, pp. 618-626. 2017.

Spencer Jr, Billie F., Vedhus Hoskere, and Yasutaka Narazaki. "Advances in computer vision-based civil infrastructure inspection and monitoring." Engineering 5, no. 2 (2019): 199-222. https://doi.org/10.1016/j.eng.2018.11.030.

Wang, Haofan, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. "Score-CAM: Score-weighted visual explanations for convolutional neural networks." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 24-25. 2020.

Zhang, Lei, Fan Yang, Yimin Daniel Zhang, and Ying Julie Zhu. "Road crack detection using deep convolutional neural network." In 2016 IEEE international conference on image processing (ICIP), pp. 3708-3712. IEEE, 2016. https://doi.org/10.1109/ICIP.2016.7533052.

Zhou, Bolei, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning deep features for discriminative localization." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921-2929. 2016.