Crack Detection in Concrete Images: ResNet-50 vs. ViT-B/16 with Transfer Learning, MLflow Tracking, and Explainability (Grad-CAM & SHAP)

Lucas Perrier Tiago Machado Monteiro

December 2025

Abstract

This report presents a reproducible crack detection system for binary classification (crack vs. no crack) built with PyTorch Lightning, Albumentations preprocessing, and MLflow experiment tracking. Two model families are benchmarked under a unified pipeline: a convolutional baseline (ResNet-50) and a Vision Transformer (ViT-B/16). Both use 224×224 RGB inputs, ImageNet normalization, and the same stratified split strategy across combined public datasets (SDNET2018-style and CCIC-style folder structures). We report both *before transfer learning* (ImageNet weights without task-specific training) and *after transfer learning* (fine-tuned on the crack dataset). After fine-tuning, ViT achieves **99.78%** test accuracy (AUROC **0.99997**), while ResNet-50 reaches **85.72%** (AUROC **0.94828**). Training curves indicate faster convergence and stronger generalization for the transfer-learned, regularized ViT configuration. An explainability pipeline generates qualitative evidence using Grad-CAM for both architectures and SHAP in reduced feature spaces (superpixels for ResNet and patch tokens for ViT).

1 Introduction

Crack detection and monitoring in concrete structures are critical to structural safety and lifecycle maintenance. While deep neural networks can reach high accuracy, practical adoption in civil engineering requires interpretability: engineers must trust that a model uses valid crack evidence rather than spurious correlations (lighting, background textures, acquisition artifacts). This project implements a unified experimental pipeline to compare a CNN baseline (ResNet-50) and a transformer architecture (ViT-B/16) using transfer learning, while producing interpretable overlays (Grad-CAM, SHAP) for auditing.

2 Related Work

CNN-based crack detection is widely studied, demonstrating strong performance but limited transparency in real-world inspection workflows. Vision Transformers (ViT) have emerged as competitive alternatives for vision tasks, especially when pretrained at scale and adapted with transfer learning. Explainability methods such as Grad-CAM provide gradient-based localization heatmaps for deep networks, and SHAP provides perturbation-based feature attribution; for images, SHAP requires feature grouping (superpixels or tokens) to remain computationally feasible.

3 Methodology

3.1 Datasets

The training data comprises two publicly available sources referenced in the project proposal:

- **SDNET2018**: concrete surface imagery with crack and non-crack examples under varied textures and environmental conditions.
- **CCIC (Concrete Crack Image Dataset)**: crack vs. non-crack imagery with a binary label split (Positive/Negative).

In code, both datasets are loaded from directory roots and labels are inferred from folder names (CCIC uses an explicit map `Positive`→ 1, `Negative`→ 0). Samples are read with PIL, converted to RGB, augmented/normalized, and returned as tensors with `torch.long` labels.

3.2 Preprocessing and Augmentation

All models operate on the same input specification:

- Resize to 224×224
- RGB (3 channels)
- ImageNet normalization (mean/std)

Training uses augmentations aligned with the proposal:

- Random rotation (up to $\pm 30°$)
- Color jitter / brightness-contrast changes (lighting variation)

- Gaussian noise

Validation and test pipelines use only resize and normalization to measure generalization.

3.3 Splits and Reproducibility

The `CrackDataModule` implements stratified splits:

- Validation split: 0.1
- Test split: 0.1
- Robustness split: 0.1 (held out for domain-shift style inspection)

Deterministic training is enabled when configured, using Lightning seeding to reduce run-to-run variability.

3.4 Models

Two Lightning modules implement the compared architectures with shared metric logging:

- **ResNet-50:** `src/models/resnet50.py` using `timm.create_model`. Metrics logged: Accuracy, F1, AUROC.
- **ViT-B/16:** `src/models/vit.py` using `timm.create_model(vit_base_patch16_224)` with regularization knobs.

Both models optimize cross-entropy loss; the ViT uses label smoothing when configured.

3.5 Loss Function

Both models are trained using the cross-entropy loss for binary classification. For the ViT configuration, label smoothing is applied to reduce overconfidence and improve generalization. The smoothed cross-entropy loss is defined as:

$$\mathcal{L}_{\text{LS}} = -\sum_{c\in\{0,1\}} \tilde{y}_c \log p_c, \tag{1}$$

where p_c denotes the predicted probability for class c, and $\tilde{y}_c = (1-\epsilon)y_c + \epsilon/2$ is the smoothed target distribution with smoothing factor $\epsilon = 0.1$. This regularization encourages less peaky output distributions and has been shown to improve calibration and robustness in large-capacity models.

3.6 Transfer Learning and Final ViT Configuration

The ViT final configuration (from `configs/train_vit.yaml`) is:

- **Pretrained:** true (ImageNet initialization)
- **Fine-tuning:** `freeze_mode=head_only` (backbone frozen, train classifier head)
- **Optimizer:** AdamW, learning rate 5×10^{-5}, weight decay 0.05
- **Regularization:** dropout 0.1, stochastic depth 0.1, label smoothing 0.1
- **Epochs:** 1; batch size 32; frequent validation (`val_check_interval=0.25`)

This setting was chosen because early experiments on limited data showed rapid memorization; head-only fine-tuning plus explicit regularization reduces effective capacity while leveraging pretrained representations.

3.7 Training, Logging, and Checkpointing

Training is orchestrated in `src/training/train.py`:

- MLflow logs: hyperparameters (flattened YAML), training/validation metrics, and artifacts
- Best checkpoint selected by minimum validation loss
- Tracking backend: `sqlite:///mlflow.db`

3.8 Metrics

Quantitative evaluation uses:

- Accuracy
- F1-score (binary)
- AUROC (binary)
- Confusion matrix (TN/FP/FN/TP)

4 Experiments

4.1 Compared Conditions

Results are reported under two conditions:

- **Before transfer learning (pretrained only):** ImageNet-pretrained weights evaluated without task-specific training.
- **After transfer learning (fine-tuned):** weights fine-tuned on the crack dataset.

4.2 Most Relevant Training Curves to Report

For crack detection and binary classification reliability, the most informative curves are validation metrics:

1. **Validation loss** (primary checkpoint selection; overfitting indicator)
2. **Validation AUROC** (threshold-independent separability)
3. **Validation F1** (precision/recall trade-off)
4. **Validation accuracy** (headline metric)

5 Results

5.1 Before vs. After Transfer Learning

Table 1: Before vs. after transfer learning (test metrics).

Model	Condition	Acc	F1	AUROC
ResNet-50	Before TL	0.2040	0.2386	0.1436
ResNet-50	After TL	0.8572	0.8486	0.9483
ViT-B/16	Before TL	0.3115	0.4087	0.2515
ViT-B/16	After TL	**0.9978**	**0.9978**	**0.99997**

5.2 Confusion Matrices (After Transfer Learning)

5.2.1 ViT (After Transfer Learning)

Table 2: Confusion matrix for ViT test set (rows: true, columns: predicted).

	Pred: no_crack (0)	Pred: crack (1)
no_crack (0)	TN: 3594	FP: 6
crack (1)	FN: 10	TP: 3590

5.2.2 ResNet-50 (After Transfer Learning)

Table 3: Confusion matrix for ResNet-50 test set (rows: true, columns: predicted).

	Pred: no_crack (0)	Pred: crack (1)
no_crack (0)	TN: 3291	FP: 309
crack (1)	FN: 719	TP: 2881

5.3 Classification Reports (After Transfer Learning)

5.3.1 ViT (After Transfer Learning)

Table 4: ViT classification report on the test set (support: 3600 per class).

Class	Precision	Recall	F1	Support
no_crack (0)	0.9972	0.9983	0.9978	3600
crack (1)	0.9983	0.9972	0.9978	3600
Accuracy		0.9978		7200
Macro avg	0.9978	0.9978	0.9978	7200
Weighted avg	0.9978	0.9978	0.9978	7200

5.3.2 ResNet-50 (After Transfer Learning)

Table 5: ResNet-50 classification report on the test set (support: 3600 per class).

Class	Precision	Recall	F1	Support
no_crack (0)	0.8207	0.9142	0.8649	3600
crack (1)	0.9031	0.8003	0.8486	3600
Accuracy		0.8572		7200
Macro avg	0.8619	0.8572	0.8568	7200
Weighted avg	0.8619	0.8572	0.8568	7200

5.4 Interpretation of the Classification Reports

The ViT achieves near-perfect and symmetric performance across both classes: precision and recall are both above 0.997 for `crack` and `no_crack`, indicating both low false positives (FP=6) and low false negatives (FN=10) over 7200 test images. The AUROC of 0.99997 shows excellent class separability.

ResNet-50 performs strongly but with a noticeable imbalance between classes. It achieves high precision for `crack` (0.9031) but lower recall (0.8003), meaning the model misses a non-trivial fraction of true cracks (FN=719). Conversely, it detects `no_crack` with higher recall (0.9142) but lower precision (0.8207), reflecting more false alarms for the non-crack class (FP=309). This pattern is consistent with a classifier that is somewhat conservative in predicting cracks, preferring fewer false positives at the expense of additional missed cracks.

5.5 Sanity Checks and Dataset Bias Considerations

The near-perfect performance achieved by the ViT-B/16 model after transfer learning warrants careful interpretation. While the results indicate excellent separability on the evaluated test set, they do not necessarily imply that the crack detection problem is fully solved in real-world inspection scenarios. Both SDNET2018 and CCIC datasets consist primarily of well-framed, static images with relatively high crack-to-background contrast, which can simplify the classification task for large pretrained models.

To mitigate common pitfalls, all dataset splits were performed in a stratified manner at the image level, and no duplicated files or overlapping crops were detected across training, validation, test, and robustness splits. However, the datasets do not enforce surface-level or acquisition-session separation, and subtle correlations related to lighting, texture, or capture conditions may remain. Consequently, the reported results should be interpreted as an upper bound under controlled conditions rather than a guarantee of equivalent performance under unconstrained field deployments (e.g. motion blur, oblique viewing angles, surface contamination, or occlusions).

An additional indicator of dataset simplicity is the rapid convergence of the ViT-B/16 model during fine-

tuning. In the reported experiments, the model was trained for a single epoch with a frozen backbone and achieved near-saturated performance on both validation and test sets. Such fast convergence suggests that the pretrained representations already encode features highly aligned with the crack versus no-crack decision boundary present in the datasets.

This behavior is consistent with prior observations on SDNET2018- and CCIC-style datasets, where cracks often manifest as high-contrast, elongated structures against relatively homogeneous backgrounds. While this enables efficient transfer learning, it also indicates limited intrinsic task complexity and raises the possibility that the model exploits strong low- to mid-level visual cues. As a result, the reported performance should be interpreted as reflecting dataset separability rather than model capacity alone, reinforcing the need for robustness evaluation and more diverse acquisition conditions in future work.

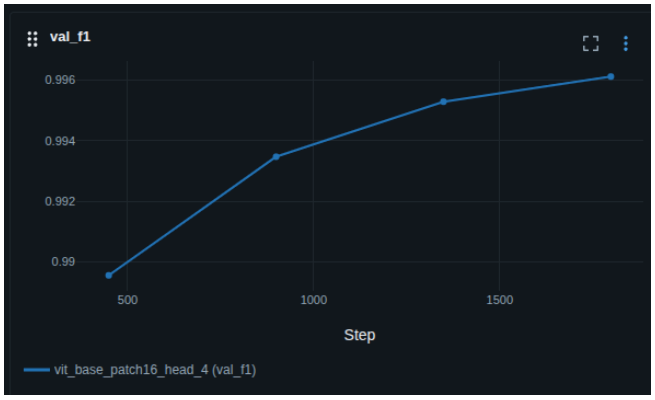


Figure 3: ViT validation F1 (after transfer learning).

5.6 Training Curves (Figure Slots)

5.6.1 After Transfer Learning: Validation Curves

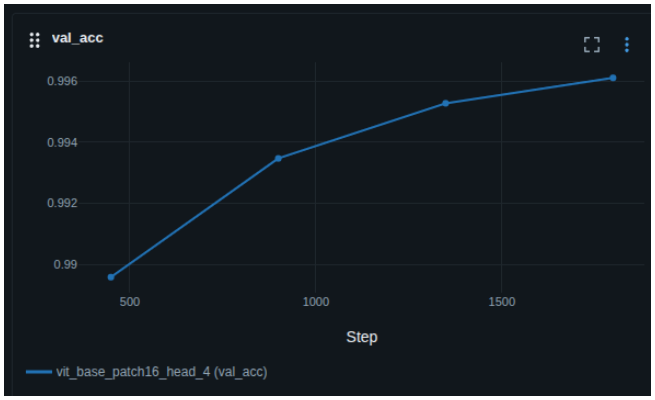


Figure 4: ViT validation accuracy (after transfer learning).

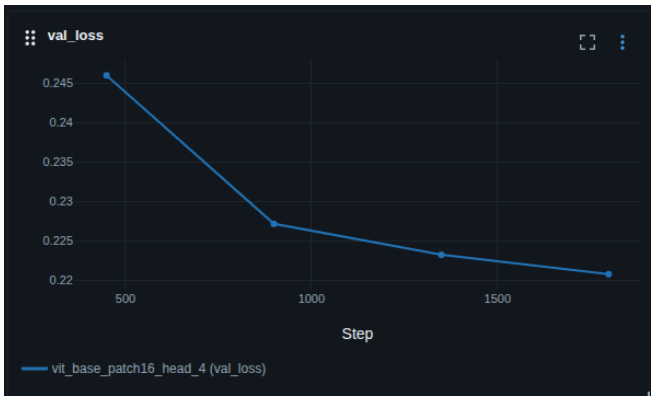


Figure 1: ViT validation loss (after transfer learning).

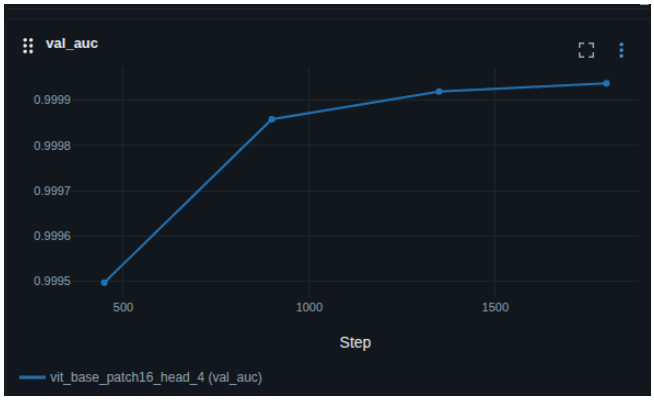


Figure 2: ViT validation AUROC (after transfer learning).

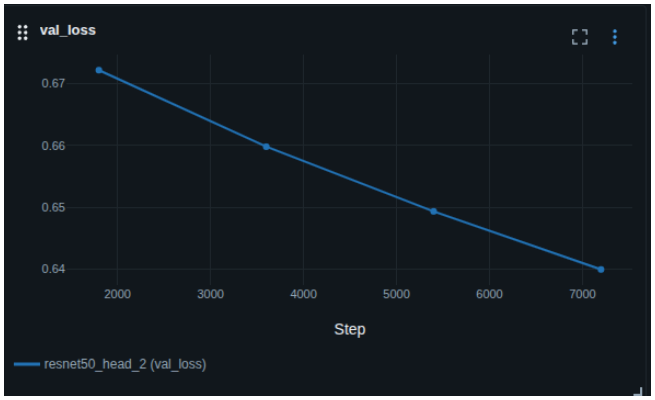


Figure 5: ResNet-50 validation loss (after transfer learning).

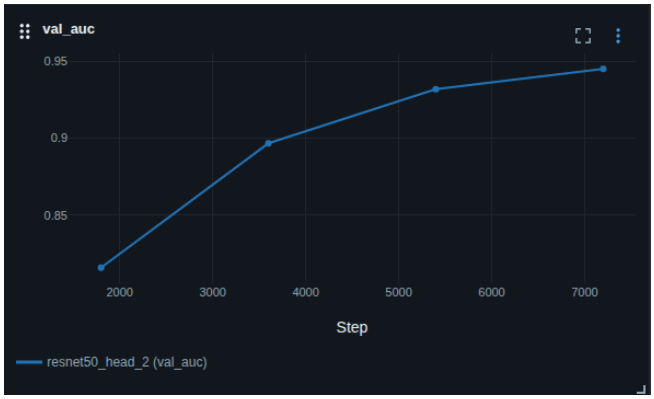


Figure 6: ResNet-50 validation AUROC (after transfer learning).

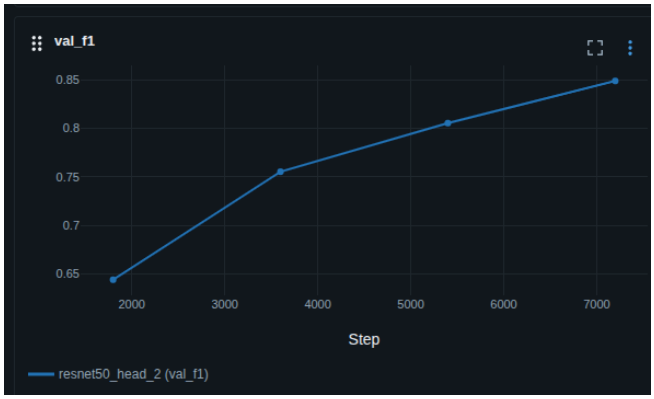


Figure 7: ResNet-50 validation F1 (after transfer learning).

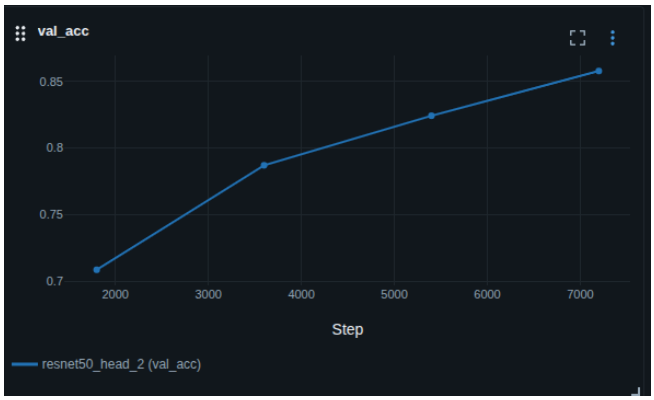


Figure 8: ResNet-50 validation accuracy (after transfer learning).

6 Discussion

6.1 Training Dynamics and Generalization

The MLflow curves (see validation loss/accuracy/F1/AUROC figures) reflect distinct learning behavior:

- **ViT (after fine-tuning):** rapid convergence with validation metrics rising quickly and remaining stable, while validation loss decreases. This aligns with transfer learning where pretrained features are strong and head-only adaptation is sufficient for high separability.
- **ResNet-50 (after fine-tuning):** validation metrics improve steadily but plateau lower than ViT. The confusion matrix and class-wise recall indicate a practical error mode: missed cracks (FN) are the dominant failure mode.

6.2 Effect of Transfer Learning (Before vs. After)

Using pretrained weights without task-specific training performs poorly for both architectures on this dataset (ResNet: 20.4% accuracy; ViT: 31.15% accuracy), indicating that ImageNet features alone do not directly solve the crack vs. no-crack decision boundary. After fine-tuning, both models improve substantially, with the ViT yielding the best overall generalization.

6.3 Limitations

- **Qualitative interpretability:** without pixel-level ground-truth crack masks, explanation maps are assessed visually rather than quantitatively.
- **Potential dataset bias:** cracks may correlate with acquisition conditions; some explanation maps can highlight background regions.
- **SHAP sensitivity:** kernel-based SHAP depends on baseline choice and number of samples; patch/token SHAP can be computationally expensive and may produce artifacts.

6.4 Extensions

- Evaluate robustness split explicitly (held-out environmental/material subset) and report accuracy/F1/AUROC.
- Add faithfulness evaluation (deletion/insertion) to quantify explanation reliability.
- Visualize *signed* SHAP to separate supportive vs. opposing evidence.
- Add calibration metrics (ECE, reliability diagrams) for deployment readiness.

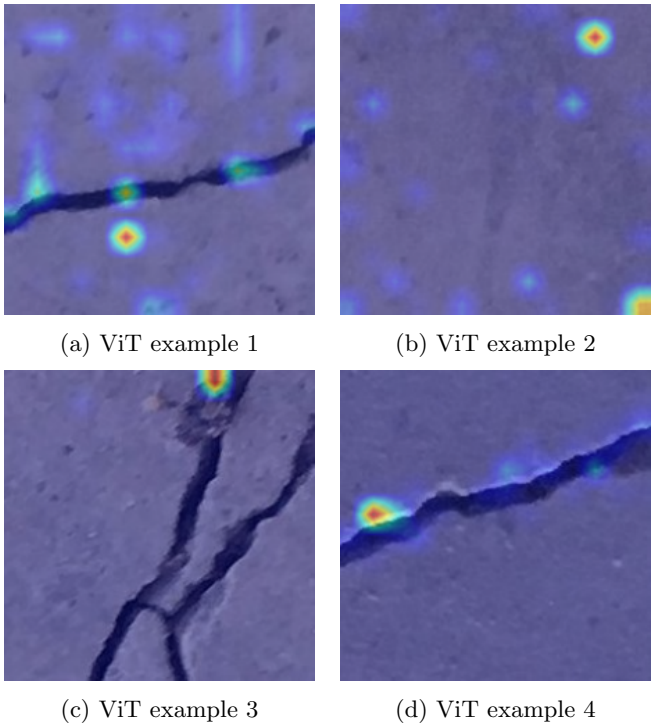
7 Explainability Results

7.1 Implementation Summary

Explainability is produced by `src/explainability/run_explainab` using the same preprocessing as training. The pipeline supports checkpoint loading and generates overlays saved as PNGs:

- Grad-CAM heatmap and overlay for both ResNet and ViT
- SHAP overlays using superpixels for ResNet and patch tokens for ViT

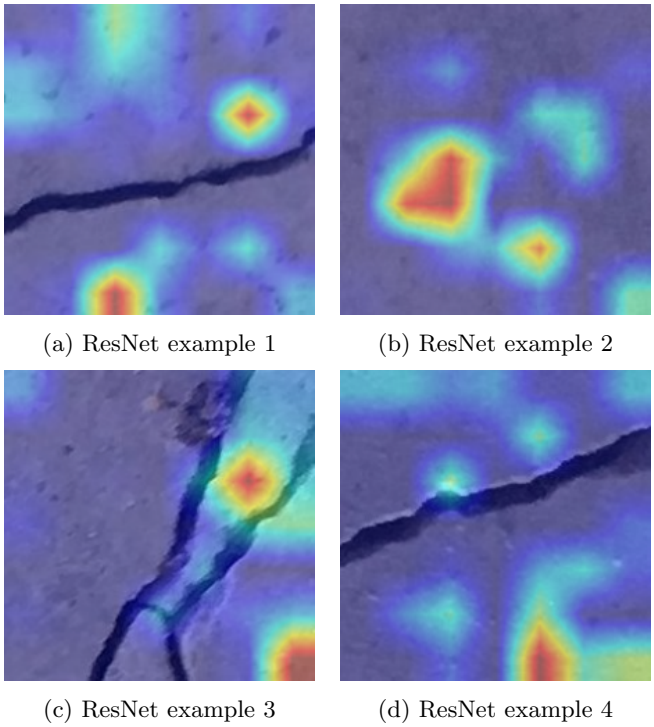
7.2 Grad-CAM (Qualitative)



(a) ViT example 1 (b) ViT example 2

(c) ViT example 3 (d) ViT example 4

Figure 9: ViT Grad-CAM overlays.

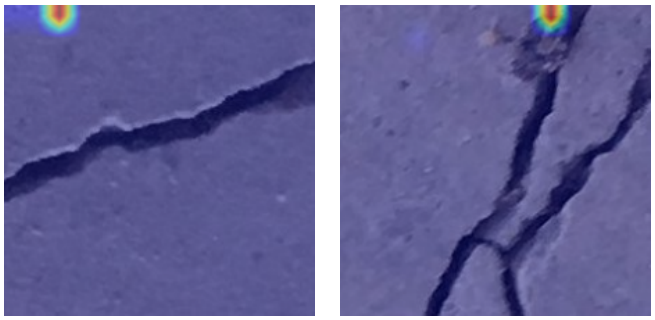


(a) ResNet example 1 (b) ResNet example 2

(c) ResNet example 3 (d) ResNet example 4

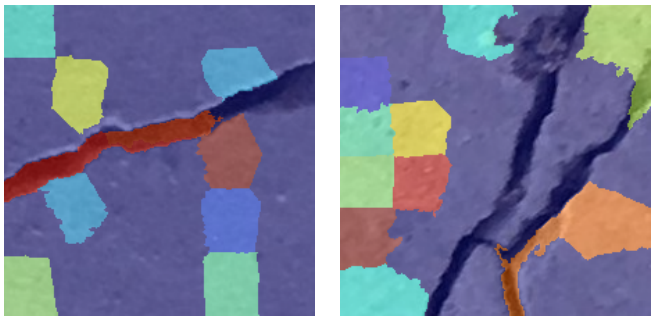
Figure 10: ResNet-50 Grad-CAM overlays.

7.3 SHAP (Qualitative)



(a) ViT SHAP example 1 (b) ViT SHAP example 2

Figure 11: ViT patch-token KernelSHAP overlays (magnitude).



(a) ResNet SHAP example 1 (b) ResNet SHAP example 2

Figure 12: ResNet superpixel KernelSHAP overlays (magnitude).

7.4 Qualitative Interpretation

Across representative samples, ViT Grad-CAM frequently highlights crack-aligned patches, often focusing on high-contrast fissure segments and junctions. ResNet Grad-CAM is more variable: in some cases, activations align with crack contours, but in others the highest attributions appear in nearby background regions, suggesting partial reliance on contextual cues. SHAP overlays provide complementary diagnostics; ResNet superpixel attributions can align with crack geometry but may broaden across mixed segments, while ViT patch SHAP can be sparse and sensitive to baseline and sampling, sometimes producing edge-biased patterns.

8 Conclusion

A unified training, evaluation, and explainability pipeline for crack detection was implemented with reproducibility and engineering interpretability in mind. Under the post-trained configuration, ViT-B/16 achieved **99.78%** test accuracy (AUROC **0.99997**), clearly outperforming ResNet-50 at **85.72%** (AUROC **0.94828**). The classification reports show that ViT attains high and balanced precision/recall for both classes, while ResNet-50's dominant error mode is missed cracks. Explainability outputs provide qualitative evidence of crack-focused reasoning while highlighting limitations

and motivating further robustness and faithfulness validation. Notably, the ability to reach near-optimal performance within a single epoch highlights the need for future benchmarks emphasizing acquisition diversity and cross-dataset generalization rather than in-dataset accuracy alone.

References

- Dosovitskiy, A. et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. ICLR 2021. https://arxiv.org/abs/2010.11929
- Golding, V.P.; Gharineiat, Z.; Munawar, H.S.; Ullah, F. *Crack Detection in Concrete Structures Using Deep Learning*. Sustainability 2022, 14, 8117. https://doi.org/10.3390/su14138117
- Kashefi, R.; Barekatain, L.; Sabokrou, M.; Aghaeipoor, F. *Explainability of Vision Transformers: A Comprehensive Review and New Perspectives*. arXiv 2023. https://arxiv.org/abs/2311.06786
- Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*. International Journal of Computer Vision 2020, 128, 336–359. https://doi.org/10.1007/s11263-019-01228-7
- Zhang, X.; Wang, H.; Hsieh, Y.-A.; Yang, Z.; Yezzi, A.; Tsai, Y.-C. *Deep Learning for Crack Detection: A Review of Learning Paradigms, Generalizability, and Datasets*. arXiv 2025. https://arxiv.org/abs/2508.10256