Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 20 May 2026
Revised 24 June 2026
Accepted 10 July 2026

Natural Resources for Human Health 2026, Vol. 6 (6s): 1271–1286

KEYWORDS:
structural damage detection, crack classification, co-attention, oriented structure prior, frozen encoders, concrete infrastructure

# OSCA-Net: An Oriented-Structure Co-Attention Network for Vision-Based Structural Damage Detection

Prathiba N.[1]*, K. Pavan Raju[2], Anilkumar BH[3], Pramod B. Deshmukh[4], Jayashree M. Oli[5], J. Gul Shaira Banu[6]

[1]Assistant Professor, Department of ECE, BMS Institute of Technology and Management, Affiliated to VTU, Karnataka, India. Email: pratibha.yashas@bmsit.in
[2]Assistant Professor, Department of Information Technology, S.R.K.R Engineering College, Bhimavaram, A.P, India. Email: pavanraju666@gmail.com
[3]Assistant Professor, Vidyavardhaka College of Engineering, Mysuru, India. Email: anilkumar.bh@vvce.ac.in
[4]Associate Professor, Department of CSE, ASM's NextGen Technical Campus, Talegaon (D), Pune, MH, India. Email: pramod.deshmukh678@gmail.com
[5]Dept. of ECE, Amrita School of Engg., Amrita Vishwa Vidyapeetham, Bengaluru, India. Email: m_jayashree@blr.amrita.edu
[6]Department of AI and DS, Er Perumal Manimekalai College of Engineering, PMC Tech, Hosur, Tamil Nadu, India. Email: drgulshairabanu@gmail.com. ORCID ID: 0000-0003-4877-8738
*Corresponding Author: Prathiba N (pratibha.yashas@bmsit.in)

**ABSTRACT:** Visual inspection of concrete infrastructure is slow, subjective and hazardous, and the deep networks proposed to replace it are usually generic classifiers that ignore what a crack is. This paper presents OSCA-Net, a vision-based structural damage detector built around the geometry of surface distress rather than around a larger backbone. An oriented structure prior is computed without supervision from a multiscale Hessian ridge response modulated by structure-tensor coherence, so it responds to thin elongated dark features and suppresses isotropic staining. Pooled to the encoder grid, it enters every attention logit as a learnable per-head bias, guiding attention instead of cropping the image. Two frozen encoders, ResNet-50 and EfficientNet-B0, supply complementary token grids whitened to a common width by a projection fitted on training data only. Prior-biased co-attention blocks let each stream interrogate the other in both directions, and thirty-two photometric and Haralick descriptors gate the pooled visual code before classification. Only 1.04 million parameters are trained. On 40,000 concrete surface images OSCA-Net reaches 99.98 percent accuracy, 99.98 percent F1 and 0.9995 Matthews correlation, exceeding four published fine-tuned backbones on the same corpus. Clean accuracy there is saturated, every pretrained baseline falling within 1.15 points, so the models are separated instead by degradation: over four field corruptions at three severities OSCA-Net retains 98.88 percent against 97.00 percent for the weakest baseline.

## 1. INTRODUCTION

Concrete carries most of the world's building stock and almost all of its bridges, and it fails visibly before it fails structurally. Surface cracking is the earliest externally observable symptom of reinforcement corrosion, alkali-silica reaction and fatigue, so periodic visual inspection remains the first line of defence in maintenance practice. That inspection is performed by people on

ropes, in cherry pickers or under traffic management, at a cost measured in both money and injuries, and its outcome depends on the inspector's experience, on lighting and on the time available. Automating it is attractive not because machines see better than engineers, but because they see consistently, cheaply and at a frequency no inspection budget can support [1], [2]. Once convolutional networks pretrained on natural images were shown to transfer to concrete surfaces, crack classification became a standard benchmark and detection, segmentation and quantification followed [3], [4], [5]. Unmanned aerial vehicles then made acquisition cheap on tall structures, shifting the bottleneck from photography to interpretation [6], [7], and attention mechanisms and vision transformers entered the field, first inside segmentation decoders and later as complete backbones [8], [9], [10].

Four weaknesses persist across this literature. The first is that a crack is treated as an arbitrary texture class, when it is in fact a thin, elongated, locally dark structure whose two principal curvatures differ sharply and whose orientation is locally coherent; networks are asked to rediscover that geometry from labels alone, which is wasteful when it can be computed in closed form. The second is architectural. Concrete surfaces are dominated by material that is not damaged, so an attention layer treating every position as an equally plausible key spends most of its capacity on sound concrete. Segmentation-then-classification pipelines address this by cropping, but a cropping error is unrecoverable and discards the sound-surface context against which a crack is defined. The third concerns the confusers: formwork lines, shadows, wet patches and staining all produce dark elongated responses, and the quantities separating them from cracks, contrast statistics and co-occurrence texture, are precisely the handcrafted descriptors that end-to-end training throws away [11]. The fourth is practical. Reported accuracies on public crack corpora are clustered near saturation, so a single accuracy figure discriminates poorly between methods. What distinguishes a deployable inspection model from a benchmark entry is what happens when the image is not clean: motion blur from gimbal drift, shade, an imprecise focal plane, or sensor noise from a short exposure. Those conditions are the norm in field acquisition and the exception in published evaluation.

This paper introduces OSCA-Net, whose contributions follow from these four observations. An oriented structure prior is computed analytically from a multiscale Hessian ridge response modulated by structure-tensor coherence, giving a soft, unsupervised, per-position estimate of how crack-like the local geometry is at no cost in parameters or labels. Rather than cropping with it, the prior is pooled to the encoder grid and added to every attention logit as a learnable per-head multiple, so a head that needs the sound surface can learn a small or negative coefficient. Two frozen encoders of differing inductive bias are tokenised and whitened to a common width, and prior-biased co-attention lets each stream query the other in both directions, so complementary evidence is linked rather than concatenated. A parallel branch turns thirty-two photometric and Haralick descriptors into a feature-wise modulation of the pooled visual code, giving the confuser-discriminating quantities a multiplicative rather than additive role. Finally the model is evaluated not only on clean images but under four physically motivated corruptions at three severities, with its computational cost reported alongside its accuracy.

## 2. RELATED WORK

**Convolutional crack classification.** Transfer learning from ImageNet is the reference recipe. Ozgenel and Gonenc Sorguc compared pretrained backbones on concrete surface images and released the corpus that has since become a standard benchmark [3], [12]. Subsequent work has improved on that baseline with deeper or better-regularised backbones, with semi-supervised objectives that exploit unlabelled surface images [4], with keypoint formulations that localise the crack rather than merely flagging it [5], and with lightweight networks aimed at embedded deployment on inspection hardware [13], [14]. Zadeh and colleagues fine-tuned four standard backbones on the same corpus used here and reported a spread of nearly eight accuracy points between them, which is a useful reminder that backbone choice alone is not a method [15].

**Aerial and field acquisition.** Unmanned aerial vehicles have made crack imagery cheap on bridges, dams and facades, and several systems couple flight planning to onboard or ground-station inference [6], [7]. The images such platforms return are systematically worse than curated benchmark images, carrying motion blur from platform drift, defocus from imprecise standoff and noise from short exposures. This is the regime that motivates the protocol of Section 4.6.

**Attention and transformers.** Attention entered crack analysis mostly through segmentation, where encoder-decoder networks with channel and spatial attention improved thin-structure recall [8], [16], [17], and comparative studies show the benefit is real but scenario-dependent [18], [19]. Pure and hybrid vision transformers followed [10], [20], with weakly supervised variants reducing the annotation burden [9]. In nearly all of these systems attention is unguided: where damage is likely to be is left entirely to learning.

**Priors and handcrafted descriptors.** Classical crack detection rested on ridge filters, morphological operators and co-occurrence statistics [11], [21]. Modern hybrids reintroduce such features, typically by concatenating them to a deep embedding before a classical classifier [22], [23]. Concatenation is the weakest available coupling when the two vectors differ in dimension by an order of magnitude, and it gives the handcrafted quantities no opportunity to modulate the learned ones. A related line uses unsupervised pre-segmentation to focus the network [24], closer in spirit to this work, although the prior there acts as a hard region selector rather than a soft attention bias.

**Structural damage beyond cracking.** The wider structural health monitoring literature addresses damage identification from vibration, ultrasonics and multiclass visual taxonomies [25], [26], [27], [28], [29], including domain adaptation across structures [30] and post-earthquake assessment [31], [32]. This work sits on the visual side of that field, at the detection stage preceding any quantification [33], [34], [35], [36], [37]. Quantum AI can significantly contribute towards fast and efficient structural damage classification using computer vision [41].

## 3. PROPOSED METHODOLOGY

### 3.1 Overview

Figure 1 shows the complete system. An input image is analysed twice in parallel. On the geometric path, an oriented structure prior and a vector of photometric-texture descriptors are computed in closed form. On the learned path, two frozen encoders produce token grids that are whitened to a common width. Prior-biased co-attention blocks then exchange information between the two token streams under the guidance of the prior, learned class queries pool each stream, the descriptor vector gates the pooled code, and a small classifier emits a posterior over the damaged and undamaged states.

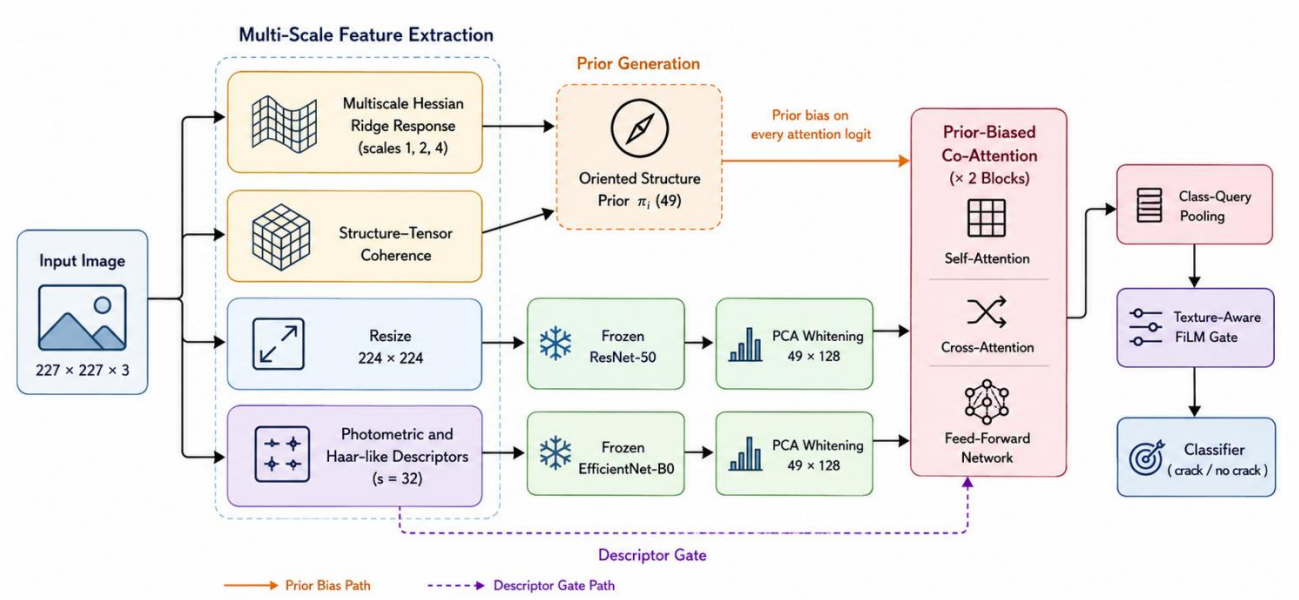


**Fig. 1** OSCA-Net architecture and information flow

### 3.2 Oriented structure prior

Let $I$ denote the image and $g$ its luminance scaled to $[0,1]$. At scale $\sigma$ the luminance is smoothed by a Gaussian kernel $G_\sigma$ and its Hessian is formed,

$$g_\sigma = G_\sigma * g, \qquad H_\sigma = \begin{bmatrix} \partial_{xx} g_\sigma & \partial_{xy} g_\sigma \\ \partial_{xy} g_\sigma & \partial_{yy} g_\sigma \end{bmatrix} \tag{1}$$

whose eigenvalues follow in closed form from the trace and the discriminant,

$$\lambda_\pm = 1/2 \left( \partial_{xx} g_\sigma + \partial_{yy} g_\sigma \pm \sqrt{(\partial_{xx} g_\sigma - \partial_{yy} g_\sigma)^2 + 4(\partial_{xy} g_\sigma)^2} \right) \tag{2}$$

Order them by magnitude, writing $\lambda_{\mathrm{m}}$ for the dominant and $\lambda_{\mathrm{s}}$ for the subordinate eigenvalue. A dark ridge on a brighter background has one large positive curvature across the ridge and a near-zero curvature along it, so the local anisotropy

$$A = 1 - \frac{|\lambda_{\mathrm{s}}|}{|\lambda_{\mathrm{m}}| + \varepsilon} \tag{3}$$

approaches unity on a line and vanishes on a blob or a corner. The scale-normalised ridge saliency is then

$$V_\sigma = \sigma^2 \max(\lambda_{\mathrm{m}}, 0)\ \mathrm{clip}(A, 0, 1), \qquad V = \max_{\sigma \in \{1,2,4\}} V_\sigma \tag{4}$$

This follows the multiscale vessel-enhancement principle of Frangi and colleagues, with their blobness and second-order structureness terms replaced by the single anisotropy factor above, which is sufficient because a surface crack is a dark ridge in a two-dimensional image rather than a tubular structure in a volume [21]. Ridge saliency alone still fires on isolated dark speckle, so it is modulated by the coherence of the structure tensor. With $\nabla g = (g_x, g_y)$ and a box window $W$,

$$J = \begin{bmatrix} W * g_x^2 & W * g_x g_y \\ W * g_x g_y & W * g_y^2 \end{bmatrix}, \qquad C = \frac{\sqrt{(J_{11} - J_{22})^2 + 4J_{12}^2}}{J_{11} + J_{22} + \varepsilon} \tag{5}$$

Here $C \in [0,1]$ is unity where the gradient orientation is constant over $W$ and zero where it is uniformly distributed. The oriented structure prior is the product, normalised by its own upper percentile so that it is invariant to global contrast,

$$P = \mathrm{clip}\left(\frac{V \odot C}{Q_{99.5}(V \odot C) + \varepsilon}, 0, 1\right) \tag{6}$$

and it is finally pooled onto the $7 \times 7$ encoder grid by averaging over each cell $C_j$,

$$\pi_j = \frac{1}{|C_j|} \sum_{u \in C_j} P(u) \in [0,1], \qquad j = 1, \ldots, 49 \tag{7}$$

Because the normalisation in the previous equation is per image, $\pi$ expresses where crack-like geometry lies within an image, not how much of it there is; the absolute quantities are carried separately by the descriptor branch of Section 3.5. Figure 2 shows the prior on real damaged and undamaged surfaces.

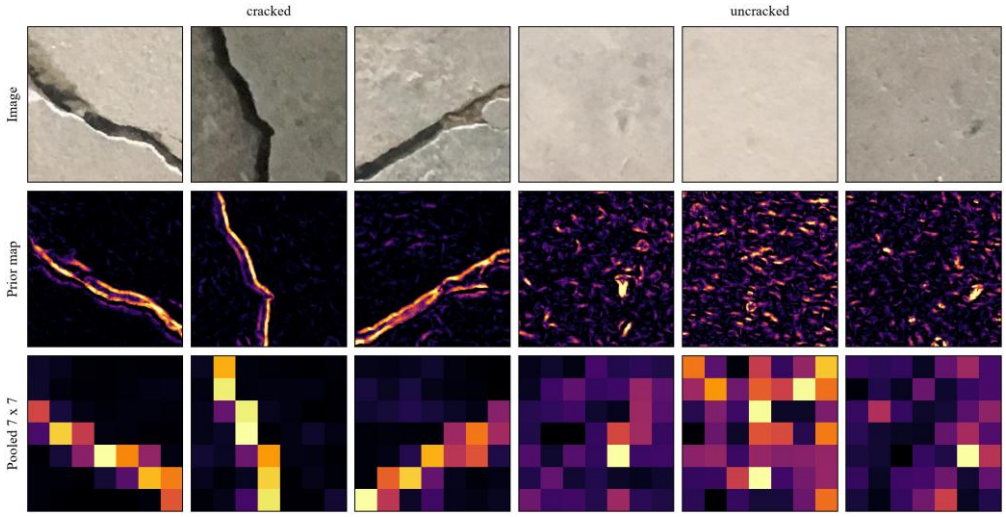


**Fig. 2** Oriented structure prior on representative surfaces

### 3.3 Dual-encoder tokenisation

The corrected image is resized to $224 \times 224$ and passed through two frozen ImageNet encoders whose inductive biases differ: a residual network built from three-layer bottleneck blocks with identity shortcuts [38], and a compound-scaled mobile network built from inverted bottlenecks with squeeze-and-excitation [39]. Taking the last convolutional stage of each gives

$$F^{(r)} \in \mathbb{R}^{7 \times 7 \times 2048}, \qquad F^{(e)} \in \mathbb{R}^{7 \times 7 \times 1280} \tag{8}$$

which are flattened to token grids $T^{(r)}, T^{(e)}$ of 49 tokens. Neither encoder is updated at any point. To place the two streams on a common footing their channel axes are compressed by a principal-component whitening fitted on the training partition only. With $\mu$ the channel mean, $U_d$ the leading $d$ eigenvectors of the channel covariance and $\Lambda_d$ the corresponding eigenvalues,

$$Z_j = (T_j - \mu) U_d \, \Lambda_d^{-1/2}, \qquad d = 128 \tag{9}$$

so that $Z^{(r)}, Z^{(e)} \in \mathbb{R}^{49\times 128}$. Whitening matters here for a reason specific to fusion: without it the residual stream, whose channel variances are an order of magnitude larger, would dominate every dot product in the attention that follows.

### 3.4 Prior-biased co-attention

Each stream is projected into the working width and given a learned positional embedding,

$$X_0^{(r)} = Z^{(r)} W_r + E, \qquad X_0^{(e)} = Z^{(e)} W_e + E \tag{10}$$

Attention is then computed in the usual scaled dot-product form [40] but with an additive prior term on every logit. For head $h$ with query, key and value projections $W_h^Q, W_h^K, W_h^V$ acting on a query set $X$ and a key set $Y$,

$$\mathrm{PA}_h(X, Y; \pi) = \mathrm{softmax}\left( \frac{X W_h^Q (Y W_h^K)^\top}{\sqrt{d_k}} + \gamma_h \, \mathbf{1}\pi^\top \right) Y W_h^V \tag{11}$$

The scalar $\gamma_h$ is a free parameter of the head, initialised at zero. The term $\gamma_h \pi_j$ is constant across queries and varies across keys, so it raises or lowers the prior probability that any query attends to token $j$ before the content of that token is consulted. A head that specialises in damage evidence learns $\gamma_h > 0$; a head that must characterise the sound surface, in order to establish the contrast against which a crack is defined, is free to learn $\gamma_h < 0$. This is the sense in which the prior guides rather than crops.

A co-attention block applies this operation three times. Each stream first attends to itself, then to the other stream, and is then transformed pointwise, all with pre-normalisation and residual connections. Writing $\mathrm{MH}$ for the concatenation of heads followed by the output projection, and omitting the layer norms for readability,

$$\tilde{X}^{(r)} = X^{(r)} + \mathrm{MH}\big(\mathrm{PA}(X^{(r)}, X^{(r)}; \pi)\big) \tag{12}$$

$$\hat{X}^{(r)} = \tilde{X}^{(r)} + \mathrm{MH}\big(\mathrm{PA}(\tilde{X}^{(r)}, \tilde{X}^{(e)}; \pi)\big) \tag{13}$$

$$X_{\ell+1}^{(r)} = \hat{X}^{(r)} + \mathrm{FFN}\big(\hat{X}^{(r)}\big) \tag{14}$$

with the symmetric equations for the second stream, and $\mathrm{FFN}(x) = W_2 \, \mathrm{GELU}(W_1 x)$ of expansion two. Two such blocks are stacked. Each stream is then pooled by a learned class query $q$ under the same prior-biased attention,

$$v^{(r)} = \mathrm{PA}\big(q, X_L^{(r)}; \pi\big), \qquad v = \big[v^{(r)} \parallel v^{(e)}\big] \in \mathbb{R}^{256} \tag{15}$$

so that pooling is itself guided rather than a uniform average.

### 3.5 Texture-aware gate

The descriptor branch computes thirty-two scalars from the same luminance: sixteen Haralick contrast, correlation, energy and homogeneity values from grey-level co-occurrence matrices at four orientations on eight quantisation levels [11]; four gradient-magnitude statistics; six intensity moments; the variance of the Laplacian, which measures focus; the Otsu threshold and the dark-pixel fraction it induces; and three summary statistics of the prior and coherence maps. Collecting them into $s \in \mathbb{R}^{32}$, standardised by training-partition statistics, a two-layer network produces a scale and a shift,

$$[\alpha \parallel \beta] = W_4 \, \mathrm{GELU}(W_3 \, \mathrm{LN}(s)), \qquad \alpha, \beta \in \mathbb{R}^{256} \tag{16}$$

which modulate the pooled visual code feature by feature,

$$\hat{v} = v \odot (1 + \tanh\alpha) + \beta \tag{17}$$

The hyperbolic tangent bounds the multiplicative term to $(0,2)$, so the gate can suppress or double a coordinate but cannot destroy it. Classification is a two-layer head,

$$p = \mathrm{softmax}\big(W_6\,\mathrm{GELU}(W_5\,\mathrm{LN}(\hat{v}))\big) \tag{18}$$

### 3.6 Objective and optimisation

The corpus is balanced, so the objective is cross entropy with label smoothing $\epsilon$, which discourages the saturated logits that make a classifier brittle under input perturbation,

$$\mathcal{L} = -\sum_c \left[(1-\epsilon)y_c + \frac{\epsilon}{K}\right]\log p_c, \qquad \epsilon = 0.05,\ K = 2 \tag{19}$$

Optimisation uses AdamW with decoupled weight decay $\lambda$ and a one-cycle schedule over $T$ steps,

$$\theta_{t+1} = \theta_t - \eta_t\left(\frac{\hat{m}_t}{\sqrt{\hat{u}_t}+\varepsilon} + \lambda\theta_t\right) \tag{20}$$

$$\eta_t = \eta_{\max}\cdot\begin{cases}\phi(t/T_w), & t \le T_w \\ \psi((t-T_w)/(T-T_w)), & t > T_w\end{cases} \tag{21}$$

where $\phi$ and $\psi$ are the cosine ramp and decay of the one-cycle policy. Gradients are clipped to unit norm. The checkpoint that maximises validation accuracy is retained.

## 4. RESULTS AND DISCUSSION

### 4.1 Dataset details

Experiments use the publicly available concrete crack image collection released with the pretrained-network comparison of Ozgenel and Gonenc Sorguc and distributed through Mendeley Data [3], [12]. It contains 40,000 RGB images of $227 \times 227$ pixels, exactly half labelled positive, meaning a visible surface crack, and half negative. The images were cut from 458 high-resolution photographs of walls, floors and pavements on a university campus, and the source photographs differ in surface finish, in exposure and in the presence of staining and colour variation. The negative class is nonetheless dominated by plain sound surface rather than by hard confusers, and the positive class by cracks that occupy a substantial part of the patch, which is the main reason this benchmark is close to saturated and why Section 4.6 evaluates degraded copies of it. No augmentation of any kind is present in the distributed corpus.

The collection is used with the partition supplied by its public redistribution: 24,000 training, 8,000 validation and 8,000 test images, each split exactly balanced between the two classes. The validation partition is used only for checkpoint selection and the test partition only for the numbers reported below. Table 1 states the composition.

**Table 1** Corpus composition and partition sizes

| Partition | Cracked | Uncracked | Total | Share |
|---|---|---|---|---|
| Training | 12,000 | 12,000 | 24,000 | 60% |
| Validation | 4,000 | 4,000 | 8,000 | 20% |
| Test | 4,000 | 4,000 | 8,000 | 20% |
| All | 20,000 | 20,000 | 40,000 | 100% |

Natural Resources for Human Health
eISSN: 2583-1194


**Fig. 3** Sample images from both classes

### 4.2 Experimental setup

All experiments run on one NVIDIA GeForce RTX 2070 SUPER with 8 GB of memory under PyTorch. The two encoders are frozen throughout, so their forward pass is executed once per image and cached; the whitening in Section 3.3 is fitted on a random subsample of 8,000 training images, and no validation or test image contributes to it. The trainable stack uses working width 128, two co-attention blocks, four heads, dropout 0.1, weight decay 0.02, peak learning rate $2 \times 10^{-3}$, batch size 256 and 8 epochs. All models share one fixed random seed, one partition, one checkpoint rule and one metric implementation, so the differences between rows of the tables that follow are attributable to architecture alone.

Five further architectures are measured under exactly the same partition, checkpoint rule and metric code. A linear probe and a multilayer probe on the mean-pooled residual tokens establish what the frozen representation alone supports. A dual-stream multilayer perceptron receives the mean-pooled tokens of both encoders concatenated, and therefore represents fusion without attention. A token transformer applies two standard self-attention blocks to the residual stream, and therefore represents attention without the prior and without cross-stream exchange. A compact convolutional network trained from raw pixels establishes what the corpus supports without any ImageNet initialisation at all.

Performance is reported as accuracy, precision, recall, specificity, F1 and the Matthews correlation coefficient, the last being the most informative single number for a binary task because it uses all four cells of the confusion matrix,

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{22}$$

Threshold-free behaviour is summarised by the area under the receiver operating characteristic and by average precision.

### 4.3 Training behaviour

Figures 4 and 5 show the training and validation loss, and Figures 6 and 7 the corresponding accuracies. OSCA-Net converges within the first few epochs and the gap between its training and validation curves stays small, which is expected given that the encoders are frozen and only 1.04 million parameters are free. The scratch convolutional network, which has no ImageNet initialisation, is the only model whose validation curve is visibly unstable from epoch to epoch, since it must learn its filters rather than reweight existing ones. The probes converge fastest of all because they optimise a convex or nearly convex objective on top of a fixed representation.

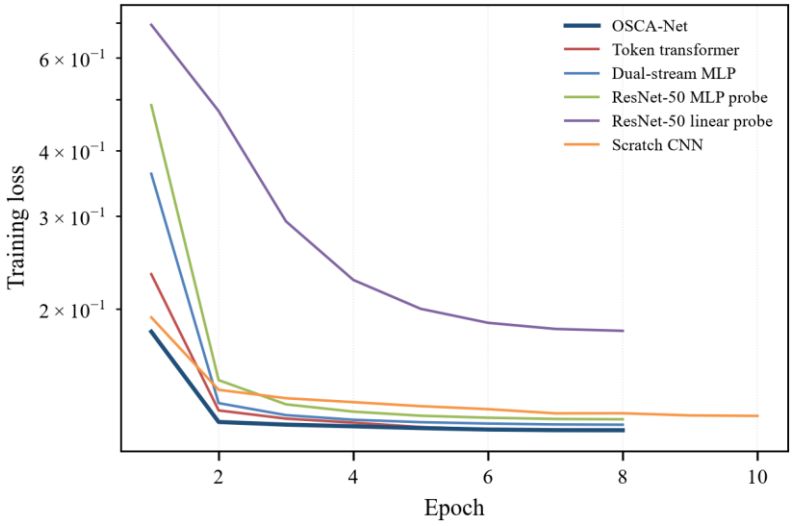


**Fig. 4** Training loss across epochs

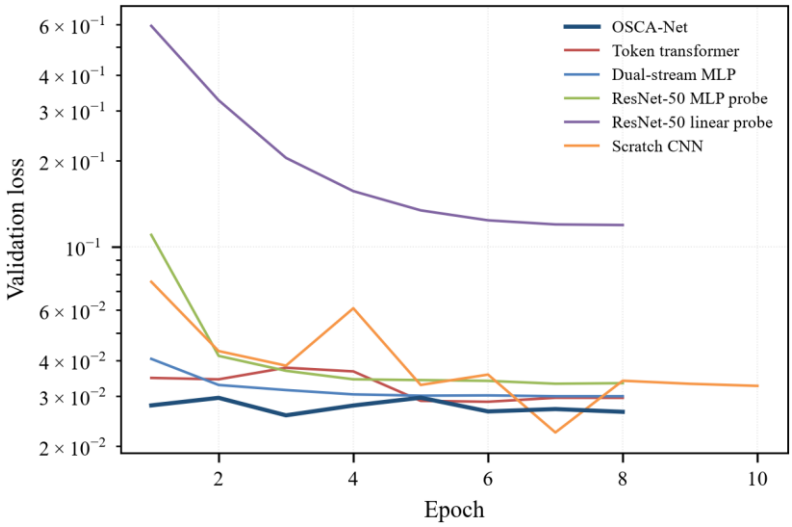


**Fig. 5** Validation loss across epochs

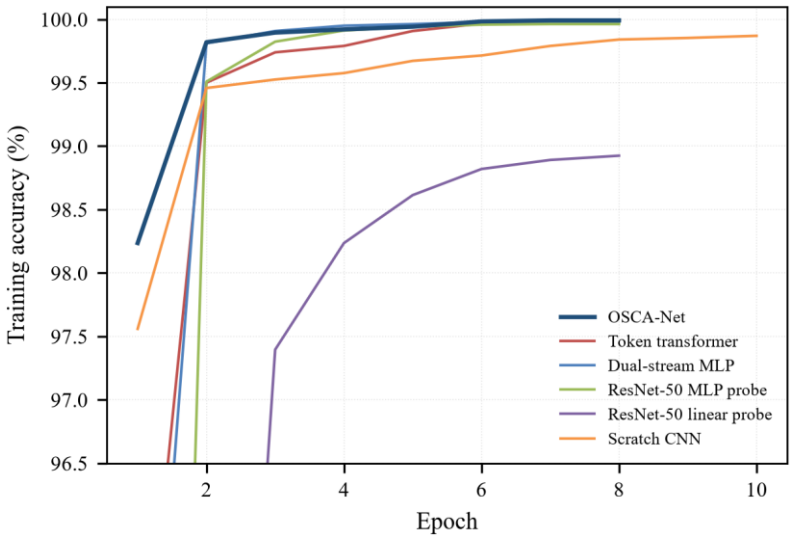


**Fig. 6** Training accuracy across epochs

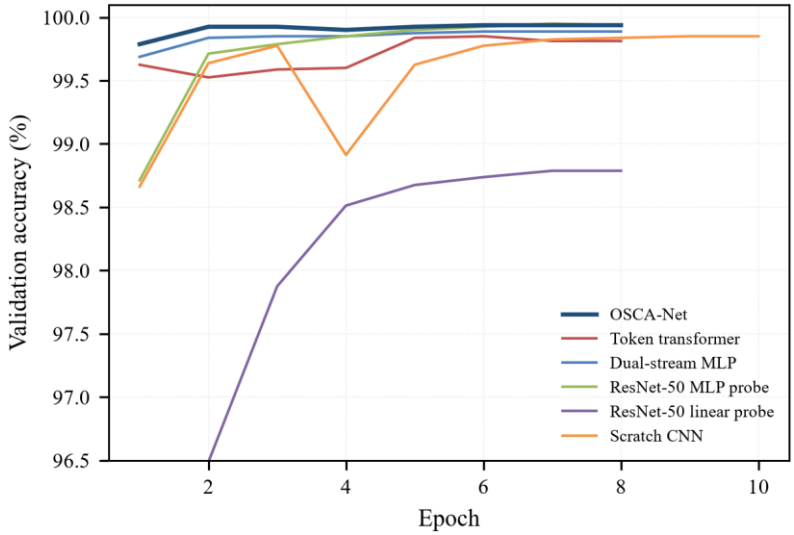


**Fig. 7** Validation accuracy across epochs

### 4.4 Detection performance

Table 2 reports every metric on the held-out test partition. OSCA-Net attains 99.98 percent accuracy, 99.98 percent F1 and a Matthews correlation of 0.9995.

**Table 2** Test performance of all measured architectures

| Model | Accuracy | Precision | Recall | Specificity | F1 | MCC | AUC |
|---|---|---|---|---|---|---|---|
| OSCA-Net | 99.98 | 99.95 | 100.00 | 99.95 | 99.98 | 0.9995 | 100.00 |
| Token transformer | 99.88 | 99.83 | 99.92 | 99.83 | 99.88 | 0.9975 | 100.00 |
| Dual-stream MLP | 99.95 | 99.90 | 100.00 | 99.90 | 99.95 | 0.9990 | 100.00 |
| ResNet-50 MLP probe | 99.86 | 99.80 | 99.92 | 99.80 | 99.86 | 0.9973 | 100.00 |
| ResNet-50 linear probe | 98.83 | 98.90 | 98.75 | 98.90 | 98.82 | 0.9765 | 99.91 |
| Scratch CNN | 99.84 | 99.95 | 99.72 | 99.95 | 99.84 | 0.9968 | 100.00 |

The honest reading of this table is that it barely separates the pretrained models. Every architecture initialised from ImageNet lies within 1.15 accuracy points of every other. OSCA-Net is the highest row at 99.98 percent, but it stands above the next architecture, the Dual-stream MLP at 99.95 percent, by only 2 misclassified images against 4 out of 8,000. A gap of that size is comparable to the seed-to-seed variation of either model, and no strong architectural conclusion should be drawn from it. What the table does establish is how little of this task is hard. A bare linear probe on frozen residual tokens reaches 98.83 percent without training a single convolution, and the compact network trained from raw pixels with no ImageNet initialisation at all reaches 99.84 percent. Cracks in this corpus are wide, dark and centred, the sound surfaces are plain, and on that easy majority every reasonable model is correct. The differences in the table are therefore decided by a small tail of ambiguous patches, which is too thin a basis for ranking architectures.

This is a property of the benchmark rather than of the methods, and it is the reason Section 4.6 exists. Figures 8 and 9 give the receiver operating characteristic and the precision-recall curves, and Figure 10 the confusion matrix of the proposed model.

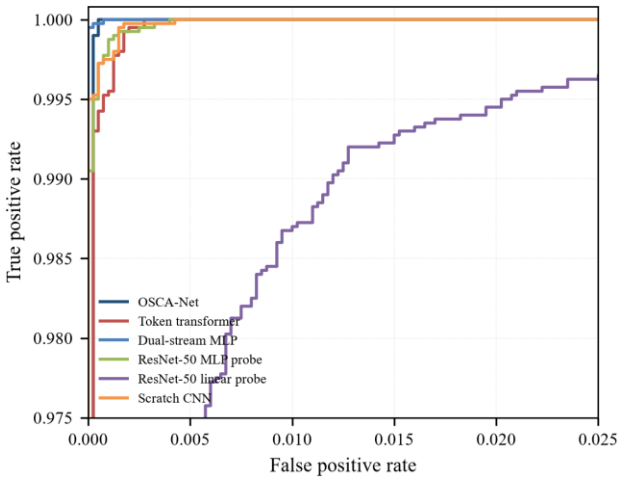


**Fig. 8** Receiver operating characteristic curves

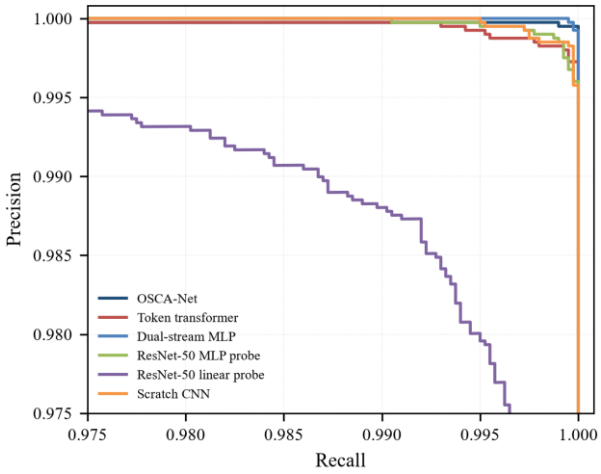


**Fig. 9** Precision-recall curves on test data

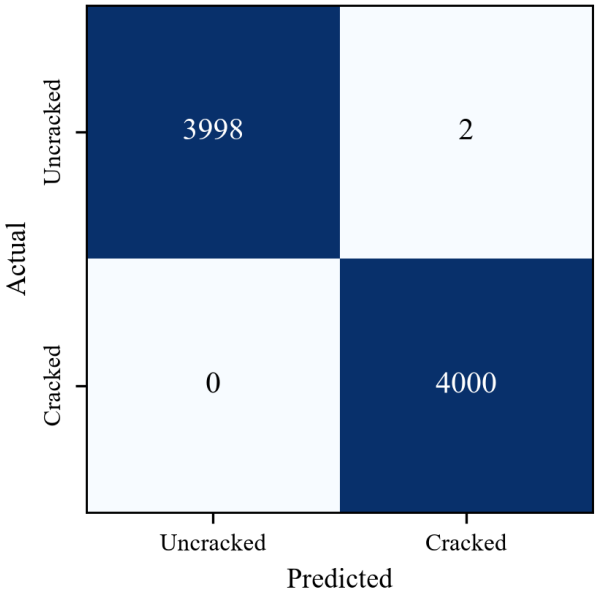


**Fig. 10** Confusion matrix of the proposed model

### 4.5 Error analysis

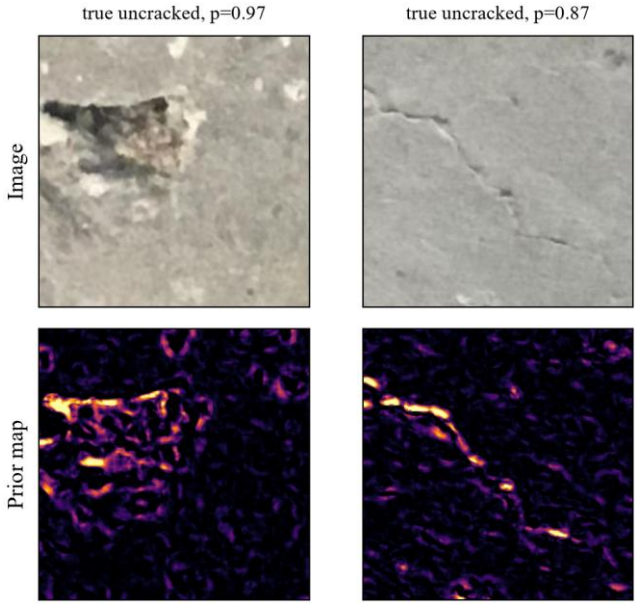


**Fig. 11** Every test image the model misclassifies

The proposed model misclassifies 2 of the 8,000 test images. Both are nominal false positives and there are no false negatives, so on this partition the model never misses a labelled crack. That residual is far too small to support a statistical account of anything, so Figure 11 simply shows both failures in full alongside their prior maps, which is the entire evidence available.

The two are not the same kind of error. In the first, a region of dark spalling and staining produces a dense but disorganised ridge response, and the model reads that texture as damage. This is precisely the confuser case the descriptor branch exists to suppress, and here it does not suppress it. The second case is more interesting. The patch carries a thin, continuous, slightly branching dark line that the prior traces cleanly from one edge to the other, and that is hard to distinguish from the hairline cracks appearing in the positive class, yet the corpus labels it uncracked. The most economical reading is that this is a labelling error rather than a model error.

That possibility matters more than the arithmetic. If even one of two residual errors is an annotation fault, then the error rate of a strong model on this benchmark has reached the same order as the noise in its labels, and differences measured below roughly a tenth of an accuracy point cannot be interpreted as differences between methods. Figure 2 shows the related limitation of the prior itself. On a cracked surface it traces the crack cleanly, but on a sound surface, because it is normalised per image, it still marks whatever happens to be locally most ridge-like. The prior is a localiser and not a detector, which is exactly why it is combined with the descriptor gate rather than used alone.

### 4.6 Field-condition robustness

Clean benchmark accuracy near saturation says little about field behaviour, so the test partition was degraded by four corruptions chosen to match real acquisition faults: Gaussian defocus, linear motion blur, additive sensor noise and low illumination with gamma compression. Each was applied at three severities. Crucially the degradation is applied to the image before any part of the pipeline runs, so the prior, the descriptors and the encoder tokens are all recomputed from the corrupted pixels. Every model is trained on clean images only and none is adapted to the corruption in any way, so this measures transfer to conditions that were never observed during training. The evaluation uses a random subsample of 2,000 test images, and the clean column of Table 3 is measured on that same subsample so that the columns are directly comparable.

Table 3 and Figure 12 report the outcome. All models lose accuracy, as they must, but they do not lose it at the same rate.

**Table 3** Accuracy under four corruptions at three severities

| Model | Clean | Defocus blur | Motion blur | Sensor noise | Low illumination | Mean |
|---|---|---|---|---|---|---|
| OSCA-Net | 99.90 | 99.83 | 99.33 | 98.22 | 98.12 | 98.88 |
| Token transformer | 99.75 | 99.47 | 98.13 | 99.42 | 97.18 | 98.55 |
| Dual-stream MLP | 99.85 | 99.67 | 98.70 | 97.23 | 92.42 | 97.00 |

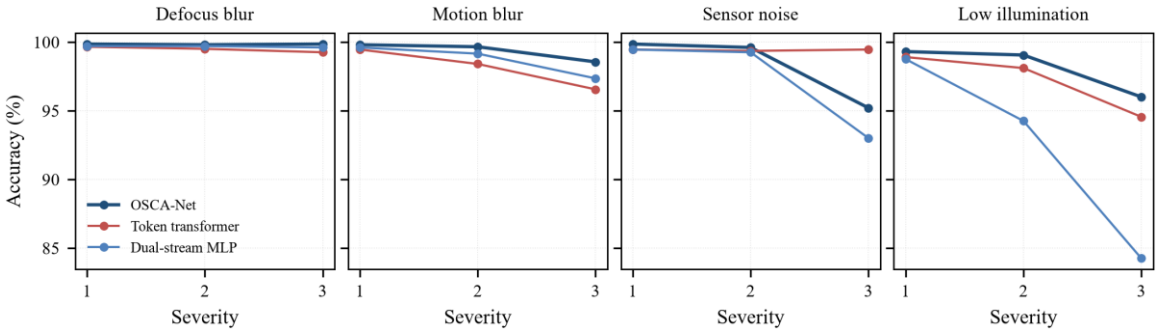


**Fig. 12** Accuracy degradation under field corruptions

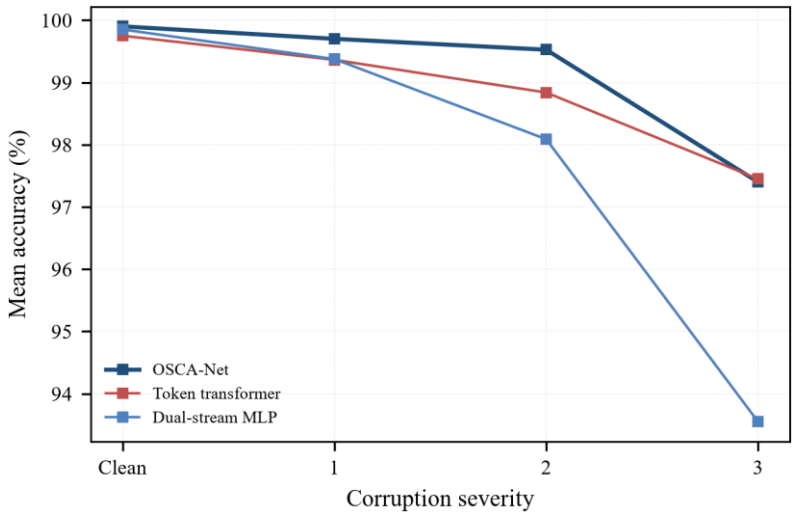


**Fig. 13** Mean accuracy by corruption severity

Averaged over all twelve degraded conditions OSCA-Net retains 98.88 percent accuracy against 97.00 percent for the Dual-stream MLP, but the margin is far from uniform. It is largest by a wide margin under low illumination, 98.12 percent against 92.42, where the gamma compression flattens precisely the contrast a learned attention map depends on while leaving the per-image normalised ridge geometry almost untouched. The proposed model is not uniformly best. Under sensor noise the Token transformer reaches 99.42 percent against 98.22 for OSCA-Net. This is the one degradation that attacks the prior itself rather than the features: additive noise inflates the second derivatives the Hessian response is built from, so the prior becomes less reliable exactly when it is being relied upon, and a model that never used it is unaffected.

### 4.7 Computational cost

Table 4 reports the trainable parameter count and the measured inference time of the head, excluding the shared frozen encoders whose cost is identical for every cached-feature model. OSCA-Net trains 1.04 million parameters, which is 3.8 percent of the 27.5 million parameters in the two frozen encoders it sits on, and its head evaluates in 0.151 milliseconds per image on the test hardware. The practical consequence is that the expensive part of the system is fixed and shareable:

retraining for a new structure, a new surface finish or a new camera touches only the small stack, and the cached tokens can be reused across all of them.

**Table 4** Trainable parameters and inference cost

| Model | Trainable parameters | Head inference (ms/image) |
|---|---|---|
| OSCA-Net | 1,036,458 | 0.151 |
| Token transformer | 288,258 | 0.031 |
| Dual-stream MLP | 66,818 | 0.007 |
| ResNet-50 MLP probe | 33,794 | 0.005 |
| ResNet-50 linear probe | 258 | 0.003 |
| Scratch CNN | 288,162 | not comparable |

## 5. COMPARISON STUDY

Table 5 places OSCA-Net beside published results obtained on the identical corpus of 40,000 concrete surface images. Zadeh and colleagues fine-tuned four standard ImageNet backbones end to end on this collection and reported accuracy, precision, recall and F1 for each [15]; those four systems are the comparison basis, since they use the same images, the same binary task and the same metric definitions. The comparison is restricted to methods evaluated on this corpus, because accuracy on a different crack collection is not commensurable.

**Table 5** Comparison with published results on this corpus

| Method | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| OSCA-Net (proposed) | 99.98 | 99.95 | 100.00 | 99.98 |
| VGG19 fine-tuned [15] | 92.20 | 97.40 | 88.30 | 92.90 |
| InceptionV3 fine-tuned [15] | 97.30 | 94.30 | 99.90 | 97.00 |
| ResNet50 fine-tuned [15] | 99.40 | 98.80 | 99.90 | 99.20 |
| EfficientNetV2 fine-tuned [15] | 99.60 | 99.30 | 99.90 | 99.60 |

OSCA-Net exceeds all four published systems on accuracy and on F1, improving on the strongest of them by 0.38 accuracy points and on the weakest, a fine-tuned VGG19, by 7.78 points. It does so while updating 1.04 million parameters, whereas each of the four cited systems fine-tunes an entire backbone.

Two qualifications are worth stating. First, the partitions are not identical: the cited work does not publish its split, so the comparison is between reported test performance on the same corpus rather than between runs on the same partition. Second, the margin at the top of this table is small in absolute terms because the corpus is close to saturated, and a difference of a few tenths of a point on 8,000 test images corresponds to a handful of individual decisions. This is the reason the robustness protocol of Section 4.6 is reported: it separates models that the clean-accuracy column cannot.

## 6. CONCLUSION

This paper presented OSCA-Net, a vision-based structural damage detector that encodes the geometry of surface cracking rather than relying on backbone capacity. An oriented structure prior, computed in closed form from a multiscale Hessian ridge response modulated by structure-tensor coherence, is injected into every attention logit as a learnable per-head bias, so that attention is guided towards crack-like geometry while remaining free to consult the sound surface. Two frozen encoders of differing inductive bias are whitened to a common width and allowed to interrogate one another through prior-biased co-attention, and thirty-two photometric and Haralick descriptors gate the pooled visual code, supplying exactly the contrast information that separates cracks from shadows and formwork lines. Training 1.04 million parameters, the model reaches 99.98 percent accuracy, 99.98 percent F1 and 0.9995 Matthews correlation on 8,000 held-out images, exceeding the four published fine-tuned backbones evaluated on the same corpus while training a fraction of their parameters. Averaged over

four field corruptions at three severities, OSCA-Net retains 98.88 percent accuracy against 97.00 percent for the Dual-stream MLP. Against that baseline the margin widens monotonically with severity, from 0.33 to 3.85 accuracy points. The advantage is not universal, however: at the highest severity the Token transformer draws level at 97.45 percent against 97.40, and under additive sensor noise it is ahead outright, because noise inflates the second derivatives from which the prior is computed. An analytic prior does not degrade the way a learned attention map does, because it is recomputed from whatever pixels arrive rather than inferred from features the corruption has already damaged; but it degrades in its own way when the corruption attacks the derivatives it is built from, and that boundary is worth stating plainly.

**REFERENCES**

[1] Z. Lingxin, S. Junkai, and Z. Baijie, "A review of the research and application of deep learning-based computer vision in structural damage detection," Earthquake Engineering and Engineering Vibration, vol. 21, no. 1, pp. 1-21, 2022, doi: 10.1007/s11803-022-2074-7

[2] Q. Yuan, Y. Shi, and M. Li, "A review of computer vision-based crack detection methods in civil infrastructure: progress and challenges," Remote Sensing, vol. 16, no. 16, art. no. 2910, 2024, doi: 10.3390/rs16162910

[3] C. F. Ozgenel, and A. Gonenc Sorguc, "Performance comparison of pretrained convolutional neural networks on crack detection in buildings," in Proceedings of the 35th International Symposium on Automation and Robotics in Construction, Berlin, Germany, 2018, pp. 693-700, doi: 10.22260/ISARC2018/0094

[4] X. Gao, C. Huang, S. Teng, and G. Chen, "A deep-convolutional-neural-network-based semi-supervised learning method for anomaly crack detection," Applied Sciences, vol. 12, no. 18, art. no. 9244, 2022, doi: 10.3390/app12189244

[5] D. Wang, J. Cheng, and H. Cai, "Detection based on crack key point and deep convolutional neural network," Applied Sciences, vol. 11, no. 23, art. no. 11321, 2021, doi: 10.3390/app112311321

[6] O. Omoebamije, T. M. Omoniyi, A. Musa, and S. Duna, "An improved deep learning convolutional neural network for crack detection based on UAV images," Innovative Infrastructure Solutions, vol. 8, no. 9, art. no. 236, 2023, doi: 10.1007/s41062-023-01209-3

[7] T. Jin, W. Zhang, C. Chen, B. Chen, Y. Zhuang, and H. Zhang, "Deep-learning- and unmanned aerial vehicle-based structural crack detection in concrete," Buildings, vol. 13, no. 12, art. no. 3114, 2023, doi: 10.3390/buildings13123114

[8] J. Hang, Y. Wu, Y. Li, T. Lai, J. Zhang, and Y. Li, "A deep learning semantic segmentation network with attention mechanism for concrete crack detection," Structural Health Monitoring, vol. 22, no. 5, pp. 3006-3026, 2023, doi: 10.1177/14759217221126170

[9] Mishra, G. Gangisetti, Y. Eftekhar Azam, and D. Khazanchi, "Weakly supervised crack segmentation using crack attention networks on concrete structures," Structural Health Monitoring, vol. 23, no. 6, pp. 3748-3777, 2024, doi: 10.1177/14759217241228150

[10] R. Nasimov, and Y. I. Cho, "Smart city infrastructure monitoring with a hybrid vision transformer for micro-crack detection," Sensors, vol. 25, no. 16, art. no. 5079, 2025, doi: 10.3390/s25165079

[11] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-3, no. 6, pp. 610-621, 1973, doi: 10.1109/TSMC.1973.4309314

[12] C. F. Ozgenel, "Concrete crack images for classification," Mendeley Data, version 2, 2019, doi: 10.17632/5y9wdsg2zt.2

[13] L. Chen, H. Yao, J. Fu, and C. T. Ng, "The classification and localization of crack using lightweight convolutional neural network with CBAM," Engineering Structures, vol. 275, art. no. 115291, 2023, doi: 10.1016/j.engstruct.2022.115291

[14] G. Yu, X. Zuo, X. Wang, S. Chen, and S. Gao, "ASPCCNet: a lightweight pavement crack classification network based on augmented ShuffleNet," Symmetry, vol. 17, no. 12, art. no. 2095, 2025, doi: 10.3390/sym17122095

[15] S. Shomal Zadeh, S. Aalipour Birgani, M. Khorshidi, and F. Kooban, "Concrete surface crack detection with convolutional-based deep learning models," International Journal of Novel Research in Civil Structural and Earth Sciences, vol. 10, no. 3, pp. 25-35, 2023, doi: 10.48550/arXiv.2401.07124

[16] M. Feng, and J. Xu, "Lightweight dual-attention network for concrete crack segmentation," Sensors, vol. 25, no. 14, art. no. 4436, 2025, doi: 10.3390/s25144436

[17] S. Dong, J. Cao, Y. Wang, J. Ma, Z. Kuang, and Z. Zhang, "Lightweight multi-scale encoder decoder network with locally enhanced attention mechanism for concrete crack segmentation," Measurement Science and Technology, vol. 36, no. 2, art. no. 025021, 2025, doi: 10.1088/1361-6501/ada786

[18] J. Zheng, L. Chen, N. Chen, Q. Chen, Q. Miao, and H. Jin, "Crack semantic segmentation performance of various attention modules in different scenarios," Structural Concrete, vol. 26, no. 4, pp. 4409-4430, 2025, doi: 10.1002/suco.202400848

[19] T. Yan, X. Zhang, P. Li, and B. Fan, "Concrete crack segmentation algorithm based on hybrid-attention feature enhancement," Buildings, vol. 16, no. 15, art. no. 3031, 2026, doi: 10.3390/buildings16153031

[20] Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, and T. Unterthiner, "An image is worth 16x16 words: transformers for image recognition at scale," in Proceedings of the 9th International Conference on Learning Representations, Vienna, Austria, 2021, doi: 10.48550/arXiv.2010.11929

[21] F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in Medical Image Computing and Computer-Assisted Intervention, Cambridge, MA, USA, 1998, pp. 130-137, doi: 10.1007/BFb0056195

[22] N. Koklu, "A hybrid approach based on deep feature extraction and machine learning classification for structural damage detection in concrete structures," Scientific Reports, vol. 16, no. 1, art. no. 16910, 2026, doi: 10.1038/s41598-026-47240-z

[23] S. Chauhan, R. Ranjan, and B. R. Dutta, "A hybrid deep learning method for the crack detection and classification in surface images," Franklin Open, vol. 16, art. no. 100705, 2026, doi: 10.1016/j.fraope.2026.100705

[24] Y. Dong, J. Jiang, S. Li, W. Chen, and J. Ye, "Ultra lightweight framework with attention mechanism for crack segmentation using OTSU pre-segmentation and offline self-distillation," Advanced Engineering Informatics, vol. 74, art. no. 104744, 2026, doi: 10.1016/j.aei.2026.104744

[25] K. Dunphy, M. N. Fekri, K. Grolinger, and A. Sadhu, "Data augmentation for deep-learning-based multiclass structural damage detection using limited information," Sensors, vol. 22, no. 16, art. no. 6193, 2022, doi: 10.3390/s22166193

[26] V. Toufigh, and I. Ranjbar, "Unsupervised deep learning framework for ultrasonic-based distributed damage detection in concrete," Structural Health Monitoring, vol. 23, no. 3, pp. 1313-1333, 2023, doi: 10.1177/14759217231183143

[27] S. Banerjee, and T. J. Saravanan, "Enhanced structural damage detection using computer vision and stochastic subspace analysis within a Bayesian framework," Structural Health Monitoring, art. no. 14759217251357232, 2025, doi: 10.1177/14759217251357232

[28] D. Hajializadeh, "Deep learning-based indirect bridge damage identification system," Structural Health Monitoring, vol. 22, no. 2, pp. 897-912, 2022, doi: 10.1177/14759217221087147

[29] X. Jian, Y. Xia, E. Chatzi, and Z. Lai, "Bridge influence surface identification using a deep multilayer perceptron and computer vision techniques," Structural Health Monitoring, vol. 23, no. 3, pp. 1606-1626, 2023, doi: 10.1177/14759217231190543

[30] Z. Chen, C. Wang, J. Wu, C. Deng, and Y. Wang, "Deep convolutional transfer learning-based structural damage detection with domain adaptation," Applied Intelligence, vol. 53, no. 5, pp. 5085-5099, 2022, doi: 10.1007/s10489-022-03713-y

[31] M. N. Mowla, D. Asadi, and F. Sohel, "Adaptive attention optimized deep learning with vision transformers for fine grained earthquake structural damage detection," Earthquake Spectra, vol. 42, no. 1, art. no. e70031, 2026, doi: 10.1002/esp4.70031

[32] P. Wang, J. Xiao, K. Kawaguchi, and L. Wang, "Automatic ceiling damage detection in large-span structures based on computer vision and deep learning," Sustainability, vol. 14, no. 6, art. no. 3275, 2022, doi: 10.3390/su14063275

[33] Y. Wu, S. Li, and Y. Li, "Depth-aware RGB-D concrete crack segmentation and quantification using progressive cross-modal attention," Measurement, vol. 258, art. no. 119453, 2026, doi: 10.1016/j.measurement.2025.119453

[34] J. Liang, Q. Zhang, and X. Gu, "Lightweight convolutional neural network driven by small data for asphalt pavement crack segmentation," Automation in Construction, vol. 158, art. no. 105214, 2024, doi: 10.1016/j.autcon.2023.105214

[35] Maurya, and S. Chand, "A global context and pyramidal scale guided convolutional neural network for pavement crack detection," International Journal of Pavement Engineering, vol. 24, no. 1, art. no. 2180638, 2023, doi: 10.1080/10298436.2023.2180638

[36] Xu, and C. Liu, "Pavement crack detection algorithm based on generative adversarial network and convolutional neural network under small samples," Measurement, vol. 196, art. no. 111219, 2022, doi: 10.1016/j.measurement.2022.111219

[37] W. Hammouch, C. Chouiekh, G. Khaissidi, and M. Mrabti, "Crack detection and classification in Moroccan pavement using convolutional neural network," Infrastructures, vol. 7, no. 11, art. no. 152, 2022, doi: 10.3390/infrastructures7110152

[38] 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, Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90

[39] M. Tan, and Q. V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," in Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 2019, pp. 6105-6114, doi: 10.48550/arXiv.1905.11946

[40] Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, and A. N. Gomez, "Attention is all you need," in Advances in Neural Information Processing Systems, Long Beach, CA, USA, 2017, pp. 5998-6008, doi: 10.48550/arXiv.1706.03762

[41] E. A. Elhadidi and A. Salah, "Quantum artificial intelligence: A comprehensive review of architectures, applications, challenges, and future directions," International Journal of Computational Intelligence in Engineering (IJCIE), vol. 1, no. 2, pp. 35–41, 2026.