International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(2) 2026 pp. 628-653
https://doi.org/10.70917/ijcisim-2026-3005
Article

CrackNet-VOS: An Intelligent Vision-Based Framework for Multi-Class Concrete Crack Classification by Type, Orientation, and Severity in Infrastructure Surfaces

Vidya V V[1], Ananth Prabhu Gurpur[2]*, Mustafa Basthikodi[3]

[1]Department of Computer Science and Engineering, Sahyadri College of Engineering & Management, Mangalore, Visvesvaraya Technological University, Belagavi, Pin: 590018, India
[2]Department of Computer Science and Engineering, Sahyadri College of Engineering & Management, Mangalore, Visvesvaraya Technological University, Belagavi, Pin: 590018, India
[3]Department of Computer Science and Engineering, Sahyadri College of Engineering & Management, Mangalore, Visvesvaraya Technological University, Belagavi, Pin: 590018, India
Email:[1]vidyasvv@gmail.com, [2]educatorananth@gmail.com, [3]mustafa.cs@sahyadri.edu.in
Corresponding Author: Ananth Prabhu Gurpur, educatorananth@gmail.com

Abstract: Concrete surface cracks threaten the safety and service life of civil infrastructure, yet most existing vision-based approaches still address only one crack attribute (typically type or severity) at a time. This work introduces CrackNet-VOS, a vision-based framework that performs joint classification of crack type, orientation, and severity on concrete elements such as beams, slabs, pavements, and tunnels. A multi-source dataset of 7,200 images was compiled from SDNET201, CRACK500, the Concrete Crack Forest Dataset (CCFD), and newly collected field data under varied lighting and environmental conditions. Each image is annotated with three labels: type (longitudinal, transverse, alligator, diagonal), orientation (horizontal, vertical, oblique), and severity (hairline, moderate, severe). The proposed model uses a hybrid feature extractor that combines EfficientNetV2 and Swin Transformer outputs, followed by three attribute-specific classification branches enhanced with attention mechanisms and trained using a weighted multi-task loss. A post-processing stage with calibrated probabilities, cross-attribute consistency checks, and a Composite Crack Risk Score (CCRS) supports maintenance prioritization. On a held-out test set, CrackNet-VOS attains accuracy/F1-score of 94.8%/94.8% for type, 92.3%/92.2% for orientation, and 90.7%/90.7% for severity, with Cohen's Kappa values of at least 0.87. Compared with strong recent baselines, including EfficientNet B3, YOLOV10–ViT, and transformer-based multi-scale models, the proposed framework achieves gains of approximately 2–5% in F1-score across tasks. These findings indicate that combining a CNN–Transformer hybrid backbone with multi-task, attention-guided heads and consistency-aware post-processing yields a robust and scalable solution for automated, multi-attribute crack assessment in structural health monitoring.
Keywords: Structural health monitoring; Concrete crack classification; multi-task learning; Vision transformer; Hybrid CNN–Transformer; Infrastructure inspection

1. Introduction

The structural health of civil infrastructure is critical to public safety, serviceability, and lifecycle performance. Among the various forms of deterioration, cracks in concrete surfaces are particularly consequential, as they can accelerate processes such as delamination, corrosion of reinforcement, and, ultimately, localized or global structural failure. Reliable detection, classification, and condition assessment of cracks are therefore central to risk mitigation strategies for bridges, roads, tunnels, buildings, and other critical assets. In practice, crack assessment remains challenging. Conventional visual inspection is labour-intensive, subjective, and prone to inconsistency and fatigue, especially when large networks of assets must be monitored under time and budget constraints. As the spatial extent and complexity of cracks increase, and as the economic and safety implications of failure grow, there is a clear need for rapid, objective, and scalable methods that can automate crack evaluation.

Vision-based methods and deep learning have shown substantial promise in this context. Convolutional neural networks (CNNs) and, more recently, attention and transformer-based architectures have enabled automated crack detection, segmentation, and single-attribute classification. However, a major limitation of many existing approaches is that they concentrate on one attribute at a time—typically crack presence, type, or a coarse

severity level—often in a binary or low-granularity form. In addition, most models do not jointly classify crack type, orientation, and severity within a single framework, which restricts their usefulness for comprehensive structural diagnostics and maintenance planning.

To address these challenges, this study proposes CrackNet-VOS, an intelligent vision-based framework for multi-class, multi-attribute classification of surface cracks in concrete infrastructure. CrackNet-VOS combines a hybrid deep learning backbone with multi-head classification to simultaneously predict crack type, orientation, and severity from a single image. The framework is trained and evaluated on a large, multi-source dataset comprising images of concrete beams, columns, pavements, slabs, and tunnels captured under diverse lighting and environmental conditions. Publicly available datasets are integrated with carefully annotated laboratory and field samples to rigorously test generalization and robustness.

Beyond improving classification accuracy, the proposed system is designed for practical deployment. By providing consistent, multi-dimensional crack descriptions and aggregating them into a composite risk score, CrackNet-VOS supports infrastructure managers, engineers, and policymakers in prioritizing maintenance, allocating resources, and optimizing lifecycle management. In this way, the framework aims to reduce manual bias and effort, enhance the reproducibility and scalability of inspection workflows, and advance the state of practice in automated structural health monitoring.

1.1 Problem Statement and Contributions

Recent deep learning approaches for concrete crack analysis have improved detection and single-attribute classification, but key gaps remain. Shi and Luo (2025) focus primarily on severity levels, without jointly modeling crack type and orientation. Zhuang et al. (2025) highlight, in their review, the lack of unified multi-attribute frameworks and limited validation under realistic field conditions. Chen et al. (2024) demonstrate strong segmentation performance with transformer-based architectures, yet do not address integrated multi-class attribute prediction or decision-oriented risk scoring. Consequently, there is still no end-to-end system that (i) simultaneously classifies crack type, orientation, and severity from real-world images, and (ii) translates these outputs into actionable maintenance priorities.

To address these gaps, this paper makes the following contributions:

- *Unified multi-attribute crack classification framework:* We propose CrackNet-VOS, an end-to-end vision system that concurrently predicts crack type (longitudinal, transverse, alligator, diagonal), orientation (horizontal, vertical, oblique), and severity (hairline, moderate, severe) from a single image, overcoming the single-task focus of prior studies such as Shi and Luo (2025) and Chen et al. (2024).
- *Hybrid CNN–Transformer backbone for robust feature learning:* We design a hybrid backbone that fuses EfficientNetV2 and Swin Transformer representations, capturing both fine-scale texture and global crack morphology. This contrasts with purely CNN-based (e.g., EfficientNet B3) or purely transformer-based models by explicitly leveraging complementary local and global features for multi-attribute classification.
- *Attention-enhanced, multi-task heads with consistency-aware post-processing:* We introduce three attribute-specific classification heads with channel/spatial attention and a weighted multi-task loss, along with a post-processing module that applies probability calibration and cross-attribute consistency checks. This design reduces inter-task interference and mitigates implausible predictions (e.g., hairline cracks predicted as severe).
- *Composite Crack Risk Score (CCRS) for actionable decision support:* We define a Composite Crack Risk Score that aggregates type, orientation, and severity into a single risk index using task-specific weights. By setting CCRS thresholds aligned with agency policies (e.g., immediate intervention, scheduled maintenance, or routine monitoring), infrastructure managers can directly use the model outputs to prioritize inspections, allocate resources, and define intervention timelines.
- *Comprehensive empirical evaluation and comparison with recent methods:* Using a 7,200-image, multi-source dataset combining public benchmarks and field data, we demonstrate that CrackNet-VOS outperforms recent CNN, hybrid YOLOV10–ViT, and transformer-based multi-scale models by approximately 2–5% in F1-score across all attributes, thereby establishing a new baseline for unified multi-attribute crack assessment in realistic conditions.

2. Related Work

Deep learning and computer vision have significantly advanced automated crack assessment in concrete structures. Existing studies can be broadly grouped into: (i) crack detection methods, (ii) single-attribute

classification (type, orientation, or severity), (iii) early multi-task or multi-attribute approaches, and (iv) transformer or hybrid CNN–ViT models for structural health monitoring (SHM). Table X summarizes key works, highlighting their main techniques, the crack features considered, and the remaining gaps relative to the unified multi-attribute objective of this study.

2.1 Crack Detection-Only Methods

Early research focused primarily on distinguishing cracked from non-cracked regions in images, often using handcrafted features and traditional machine learning. Authors in [1] employ geometric descriptors such as width, length, and aspect ratio, combined with support vector machines (SVM), to automate crack recognition. While these approaches offer simple implementation and reasonable performance in controlled conditions, they rely heavily on feature engineering and are sensitive to lighting changes, surface texture, and noise. They also provide only binary or coarse crack/non-crack decisions and do not estimate severity or orientation, limiting their usefulness for detailed maintenance planning.

More recent works using deep CNNs improve detection robustness and reduce the need for manual feature design, but still remain focused on the presence or absence of cracks. These methods typically do not provide fine-grained classification by type, orientation, and severity within a single end-to-end framework.

Gap: High detection accuracy, but no unified multi-attribute classification (type–orientation–severity), no decision-oriented risk scoring, and limited consideration of heterogeneous field conditions.

2.2 Single-Attribute Classification: Type, Orientation, or Severity

A second line of work aims at classifying specific crack attributes, such as severity level or geometric pattern, often using deep CNNs and transfer learning. CNN and ResNet-50 variants were applied to categorize crack severity into minor, moderate, and severe, achieving high accuracy on their datasets [2]. However, their model does not simultaneously classify crack type or orientation, and the evaluation is tailored to particular structural components, which hinders generalization. The EfficientNet-based models are investigated for asphalt pavement crack classification, targeting multi-class crack types (e.g., longitudinal, transverse, alligator). While they report strong F1-scores and robustness to noise, their focus is primarily on type classification; orientation and severity are not modeled jointly, and there is no mechanism to link outputs to actionable maintenance thresholds [3]. The images using Fourier-based preprocessing combined with CNNs are enhanced to better capture geometric attributes, particularly helpful for orientation-related distinctions. Although this improves type and orientation recognition, severity is either simplified or not explicitly modelled, and the system is not designed as a general multi-attribute framework [4].

Gap: These models provide accurate single-attribute predictions (e.g., only severity or only type), but still lack integrated, concurrent prediction of type, orientation, and severity, as well as post-processing for consistency or risk scoring.

2.3 Emerging Multi-Task / Multi-Attribute Approaches

Several studies begin to move beyond single-attribute classification, yet they typically cover only a subset of relevant crack properties or restrict their scope to specific materials or structural components. Authors in [5] use an ensemble of SqueezeNet, U-Net, and MobileNet-SSD to assess both severity and type on road cracks, showing that multi-task modeling can improve overall performance. However, orientation is not explicitly considered, and the models are largely tuned for one dataset and domain, with limited analysis of cross-domain generalization.

The LightGBM, deep neural networks, and CNNs are compared for recognizing pavement crack patterns, exploring different attribute combinations, but still not delivering a unified, end-to-end architecture that jointly learns type, orientation, and severity from a shared representation [6]. Moreover, these works rarely include consistency constraints across attributes or mechanisms for operational decision support. In [7], a comprehensive review of deep learning methods for surface crack detection and classification explicitly identifies the absence of fully integrated multi-attribute frameworks and the scarcity of realistic, diverse training datasets. Their analysis underscores the need for models capable of handling multiple crack attributes simultaneously in heterogeneous environments.

Gap: Early multi-task efforts show the benefit of learning multiple attributes, but they (i) typically omit at least one key attribute (often orientation), (ii) lack cross-attribute consistency handling, and (iii) do not translate outputs into a composite risk index for maintenance.

2.4 Transformer and Hybrid CNN–ViT Models in SHM

The emergence of transformer-based vision architectures has prompted several works to explore global attention mechanisms for crack detection and segmentation. A transformer-based, multi-scale feature aggregation model is proposed for crack detection, classification, and segmentation on pavements. The method achieves strong segmentation performance under noisy and complex backgrounds, demonstrating the value of transformer attention [8]. Nevertheless, its classification component does not implement a fully integrated multi-attribute scheme, and the focus is not on simultaneous type–orientation–severity prediction or decision-oriented post-processing. An improved Swin-Unet model for pavement crack detection and segmentation is developed, leveraging hierarchical transformer attention to enhance robustness to noise and background complexity [9]. While segmentation accuracy is high, the framework is not designed for multi-class attribute classification, and severity is not modeled as a separate, explicit task.

Hybrid pipelines, such as the YOLOV10 + ViT framework, combine object detection and transformer-based classification. These approaches enable rapid detection and early warning for multiple crack types, but they typically operate with a limited number of class categories and do not include severity and orientation as separate outputs [10]. Furthermore, post-processing is generally limited to thresholding; there is no cross-attribute reasoning or comprehensive risk scoring mechanism.

Gap: Transformer and hybrid CNN–ViT models significantly improve detection and, in some cases, segmentation under challenging conditions, but they (i) are not structured as unified multi-task classifiers for type, orientation, and severity, (ii) lack attribute-specific attention heads trained under a coordinated multi-task loss, and (iii) do not incorporate calibrated, consistency-aware post-processing or composite risk metrics.

Table 1. Summary of representative crack analysis studies and remaining gaps

Reference & Year	Techniques / Methods	Crack Feature Classified	Vision Based?	Key Benefits	Gap / Limitation
Shi & Luo (2025)	Deep learning (CNN/SCNN, ResNet-50)	Severity (minor, moderate, severe)	Yes	High multi-class severity accuracy (~90%)	Only severity; no type/orientation; no unified multi-attribute framework
Wang et al. (2024)	SVM, geometric feature extraction	Type	Yes	Simple, automated detection using handcrafted features	Sensitive to noise/lighting; limited robustness; no severity/orientation; no risk scoring
Matarneh et al. (2024) [11]	CNN, transfer learning (EfficientNet B3)	Type (longitudinal, transverse, diagonal, alligator)	Yes	High F1 for multi-class type; robust to noise	Only type classification; no explicit severity/orientation; no multi-task design
Mayya & Alkayem (2024)	YOLOV10 + ViT hybrid	Type (multi-class, few categories)	Yes	High precision/recall; early warning and fast detection	Limited number of crack classes; no severity or orientation heads; no CCRS or consistency checks
Ashraf et al. (2024)	Transformer attention, multi-scale feature aggregation	Detection, classification, segmentation	Yes	Strong segmentation under complex backgrounds	Focus on detection/segmentation; no integrated type–orientation–severity multi-task classification
Moreh et al. (2024) [12]	Deep neural networks, keypoint localization	Internal (micro-scale) cracks	No	Novel non-visual crack localization (e.g., seismic context)	Not suitable for typical surface crack assessment; no visual, multi-attribute output
Zhuang et al. (2025)	Systematic review of deep learning methods	Surface cracks (various)	Yes (Review)	Comprehensive SOTA and dataset analysis	Survey only; identifies but does not resolve lack of unified multi-attribute frameworks
Wang et al. (2021) [13]	Deep CNN	Severity (ballastless track slab)	Yes	High severity detection accuracy	Application-specific; no type/orientation; binary or coarse severity only

Hoang & Nguyen (2023)	CNN, DNN, LightGBM hybrid	Pavement crack patterns	Yes	Comparative evaluation of ML/DL approaches	No end-to-end multi-attribute framework; limited orientation/severity modeling
Sun et al. (2024)	Fourier image enhancement + CNN	Type, orientation	Yes	Improved pixel-level geometric/orientation sensitivity	Orientation and type only; severity not modeled; no multi-task or consistency-aware post-processing
Del Savio et al. (2025) [14]	Dataset publication	Beam/column cracks (labeled)	Yes	High-quality dataset for training and benchmarking	Data resource only; no new multi-attribute algorithm
Ha et al. (2022)	SqueezeNet, U-Net, MobileNet-SSD ensemble	Severity, type	Yes	Ensemble approach; ~91% accuracy	Domain-specific; no explicit orientation; limited cross-dataset generalization analysis
Ji et al. (2025) [15]	Experimental + numerical analysis	Deterioration via crack size	No	Quantitative link between crack size and material damage	Indirect, non-vision-based; no automatic image-based classification
Anggara et al. (2024) [16]	Wall-climbing robot + vision	Orientation, detection	Yes	Autonomous inspection of inaccessible surfaces	Hardware complexity; limited attribute set; no integrated severity/type classification
Yu et al. (2025) [17]	Acoustic emission + SVM	Crack mode (AE-based)	No	Non-visual internal crack mode classification	Requires AE hardware; no surface image analysis; no multi-attribute vision framework
Eshraghi et al. (2025)	ML on acoustic emission signals	Crack mode (acoustic)	No	Models AE wave propagation effects on crack modes	Not directly vision-based; no type–orientation–severity integration

3. Methodology

3.1 Data Acquisition and Annotation

This study adopts a systematic data acquisition and annotation pipeline to ensure that the proposed CrackNet-VOS framework is trained and evaluated on a diverse, high-quality, and reproducible dataset. The process comprises: (i) integration of multiple public benchmark datasets and in-field inspections, (ii) a standardized annotation protocol for crack type, orientation, and severity, (iii) stratified data partitioning with explicit control of class balance, and (iv) a carefully designed augmentation strategy that preserves crack semantics while enhancing robustness.

3.1.1 Dataset Sources and Composition

To cover a broad spectrum of concrete cracking scenarios typical of real-world civil infrastructure, we compiled an extensive image corpus from both publicly available datasets and in-house inspections:

Public datasets:

- *SDNET201* – containing images of concrete bridge decks, walls, and pavements with and without cracks under varying lighting and surface conditions.
- *CRACK500* – providing close-range images of pavement and concrete surfaces, particularly suited for fine-scale crack detection.
- *Concrete Crack Forest Dataset (CCFD)* - offering densely cracked concrete images with complex backgrounds and noise.

In-house inspection dataset: High-resolution RGB images were collected from routine inspections of bridges, pavements, slabs, and tunnels. Images were acquired using commercial digital cameras and smartphone sensors under unconstrained field conditions (different illumination, shadows, moisture, weathering, surface textures, and contamination). The acquisition distance ranged from approximately 0.5–2.0 m, and images were captured at native resolutions between 1024×768 and 3024×4032 pixels.

All images were screened to ensure that (i) the visible surface is predominantly concrete, (ii) potential cracks are resolvable at the captured resolution, and (iii) severe motion blur or saturation is absent. After cleaning and deduplication, the final corpus comprised **7,200 distinct images** that were subsequently used for multi-attribute classification. Each image was assigned three independent label dimensions:

1. *Crack type:* longitudinal, transverse, alligator, diagonal.
2. *Crack orientation:* horizontal, vertical, oblique.
3. *Crack severity:* hairline, moderate, severe.

Class distributions were monitored to avoid dominant categories that could bias learning. The final dataset maintained an approximate balance with 6,400 samples used for crack type classification, 4,800 samples for orientation, and 4,600 samples for severity. For model training, all images were isotropically resized to 224×224 pixels, preserving aspect ratio and crack morphology as much as possible. Pixel intensities were normalized channel-wise using the empirical dataset mean and standard deviation to stabilize training. Table 2 summarizes the principal characteristics of the resulting dataset, including the image resolution range and data format, underscoring its suitability for advanced vision-based, multi-attribute crack classification.

3.1.2 Annotation Protocol

A rigorous annotation protocol was designed to ensure label consistency across heterogeneous sources. Annotation was carried out using **LabelImg** and **CVAT**, following a detailed guideline jointly developed by structural and materials engineering experts involved in the project.

Crack type. Cracks were categorized based on geometric pattern and typical manifestation in civil infrastructure:

- *Longitudinal:* cracks aligned predominantly in the direction of traffic or structural span.
- *Transverse:* cracks intersecting the direction of traffic or span, typically orthogonal to longitudinal cracks.
- *Alligator:* interconnected crack networks forming polygonal or block-like patterns, commonly associated with fatigue or load-induced distress.
- *Diagonal:* cracks at oblique angles, often indicative of shear or torsional stresses.

Crack orientation. To avoid subjective judgment, orientation was defined using thresholded angular criteria relative to the image axes:

- Let θ denote the principal orientation of the crack segment(s), obtained visually with aid of straight-line tools in CVAT.
- *Horizontal*: $|\theta - 0°| \leq 15°$ or $|\theta - 180°| \leq 15°$.
- *Vertical:* $|\theta - 90°| \leq 15°$.
- *Oblique:* all remaining angles ($15° < \theta < 75°$ or $105° < \theta < 165°$).

This rule-based approach standardizes orientation labelling across annotators, reducing ambiguity in borderline cases.

Crack severity. Severity was determined by direct or scale-assisted estimation of crack width, complemented by pattern complexity:

- *Hairline:* crack width < 0.25 mm, generally continuous and narrow without visible branching.
- *Moderate:* 0.25–1.0 mm, including minor branching and localized widening.
- *Severe:* > 1.0 mm, or exhibiting pronounced branching, multiple intersecting cracks, spalling, or evident material loss.

Where feasible, reference rulers, known structural dimensions, or sensor metadata were used to calibrate width estimates. In images lacking explicit scale, relative width was inferred using comparable features (e.g., joint spacing, aggregate size) under the guidance of domain experts.

Annotator training and quality control. Two civil engineering specialists with prior experience in crack evaluation independently labelled overlapping subsets of the data. An initial pilot batch (10% of the images) was jointly annotated to refine category definitions and minimize inter-annotator discrepancies. For the main annotation phase, any disagreements or uncertain cases were resolved through a consensus review session. To quantify label reliability, inter-annotator agreement was periodically assessed using a subset of the data; disagreements were traced back to guideline ambiguities and the instructions were updated accordingly. Each finalized label set was then subjected to integrity checks, including:

- verification of non-empty labels for all three dimensions (type, orientation, severity);
- detection of logically inconsistent combinations (e.g., alligator cracks misclassified as hairline when the visible network was extensive);
- spot audits across data sources to ensure that labeling criteria were applied uniformly to public and in-house images.

The dataset supports multi-label classification, where a single image concurrently carries type, orientation, and severity labels, enabling joint learning of interrelated attributes.

3.3.3 Data Partitioning and Class Balance

To enable robust model development and unbiased performance assessment, the dataset was partitioned into training, validation, and test subsets. The splitting strategy adhered to the following principles:

- *Split ratios:* 70% training, 15% validation, and 15% testing.
- *Stratified random sampling***:** splits were performed independently for each label dimension such that the marginal distribution of classes (for type, orientation, and severity) was approximately preserved in all subsets.
- *Source-level separation:* images originating from the same physical structure or inspection sequence were, as far as possible, assigned to a single subset to mitigate data leakage arising from near-duplicate perspectives.

Given the relative scarcity of alligator and severe crack samples, a mild class imbalance remained after stratification. To alleviate this without artificially inflating test performance, we adopted targeted augmentation and oversampling only within the training set for underrepresented classes. The validation and test sets were kept free from synthetic augmentations and oversampling to preserve an unbiased estimate of generalization.

A strict no-overlap constraint was enforced: neither original images nor their augmented variants were shared across different subsets. Class distributions and label co-occurrences were continuously monitored during partitioning, and minor adjustments were introduced to maintain approximate balance and avoid pathological skew (e.g., near-absence of a given severity level in the validation set).

3.1.4 Data Augmentation Strategy

To enhance the generalization capability of CrackNet-VOS to diverse field conditions, a real-time data augmentation pipeline was deployed during training. The design philosophy was to introduce realistic variations in geometry, illumination, and noise while strictly preserving the semantic integrity and detectability of cracks.

Geometric augmentations.

- *Random translation***:** up to ±15% of image dimensions along horizontal and vertical axes.
- *Random scaling:* uniform scaling in the range [0.9, 1.1], avoiding excessive magnification that could distort crack width interpretation.
- *Random rotation:* uniform rotation between −90° and +90°, simulating arbitrary camera orientations in the field.
- *Horizontal and vertical flipping:* to counter orientation bias and improve robustness to different viewpoints.

Photometric augmentations.

- *Brightness and contrast adjustment:* linear scaling and shifting of pixel intensities to emulate shadows, over-exposure, and sensor variability.
- *Adaptive gamma correction:* γ randomly sampled from a predefined range to simulate varying illumination and material albedo.
- *Synthetic shadows:* addition of soft, spatially varying dark regions, mimicking obstruction by structural elements or environmental objects.

Noise and blur.

- *Gaussian noise:* zero-mean with standard deviation $\sigma \in [0.01, 0.04]$, modeling sensor noise and compression artifacts.
- *3×3 blur kernels:* mild spatial smoothing to mimic out-of-focus regions and motion blur, without erasing fine crack details.

Occlusion modeling.

- *Random erasing:* insertion of rectangular occlusions covering *2–10%* of the image area at random locations, imitating partial occlusions by dirt, markings, or hardware fixtures while ensuring that the primary crack region is not entirely removed.

To maintain diversity without overwhelming the training process, each augmentation operator was applied with a probability in the range 0.4–0.6, and combinations were sampled stochastically at each training iteration. Hyperparameters of the augmentation pipeline (e.g., ranges and probabilities) were empirically tuned on the validation set to avoid degradation of crack visibility or systematic distortion of crack width and continuity.

Collectively, this comprehensive acquisition and annotation framework—spanning heterogeneous sources, standardized labeling, balanced partitioning, and carefully controlled augmentations—provides a robust foundation for training and evaluating the proposed CrackNet-VOS model for multi-attribute crack classification under realistic operational conditions.

A detailed overview of the dataset composition across crack type, orientation, and severity, together with the number of labeled samples per attribute category and their corresponding image formats and resolutions, is provided in Table 2. All images were captured at native resolutions ranging from 1024×768 to 3024×4032 pixels and subsequently resized to 224×224 pixels for model input, as also reported in Table 2.

Table 2. Summary of the multi-attribute concrete crack dataset used for training and evaluating the CrackNet-VOS framework.

Attribute Category	Classes	Total Samples Used for Learning	Native Resolution Range	Preprocessed Size	Data Format
Crack Type	Longitudinal, Transverse, Alligator, Diagonal	6,400	1024×768 to 3024×4032 pixels	224×224 pixels	JPEG, PNG
Crack Orientation	Horizontal, Vertical, Oblique	4,800	1024×768 to 3024×4032 pixels	224×224 pixels	JPEG, PNG
Crack Severity	Hairline, Moderate, Severe	4,600	1024×768 to 3024×4032 pixels	224×224 pixels	JPEG, PNG
Total Images	Multi-labeled across all three dimensions	7,200 unique images	1024×768 to 3024×4032 pixels	224×224 pixels	JPEG, PNG

3.2 Model Architecture

The proposed CrackNet-VOS architecture performs joint classification of crack type, orientation, and severity within a single unified framework. It employs a hybrid backbone combining EfficientNetV2 (convolutional) and Swin Transformer (vision transformer) streams, followed by a feature fusion module and task-specific classification heads. A comprehensive visual overview of the end-to-end multi-task learning pipeline is presented in Figure 1, illustrating how local and global feature paths converge into specialized decision modules.

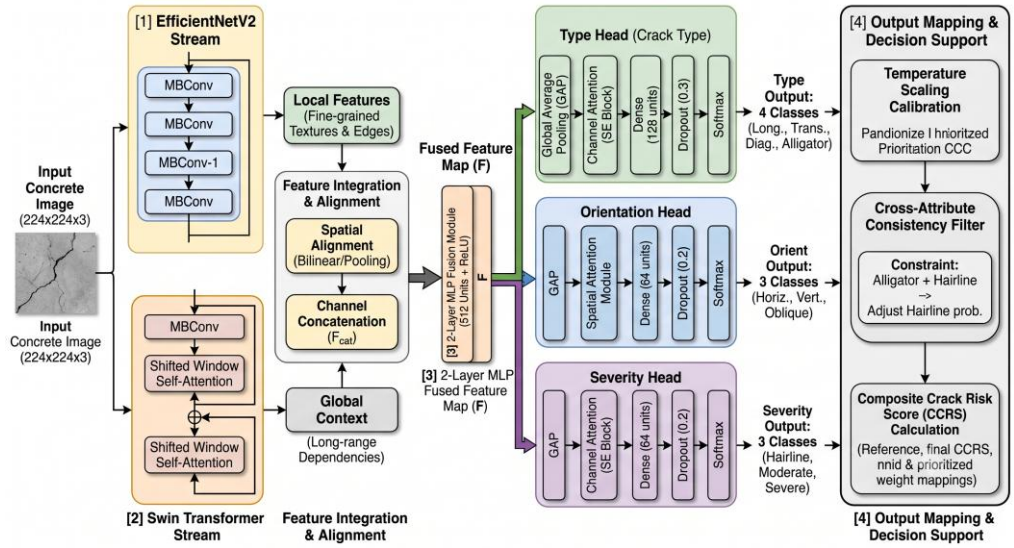

Figure 1. Architectural Workflow of the Proposed CrackNet-VOS Framework

The system takes a concrete surface image (224x224x3) and processes it through parallel streams. [1] The EfficientNetV2 branch extracts fine-grained local textures and edge details. [2] The Swin Transformer branch captures global context and long-range dependencies using Shifted Window Self-Attention. These features are aligned, concatenated, and fused by a [3] 2-Layer MLP module. The fused feature map (F) is then fed to three parallel heads to predict Crack Type (4 classes), Orientation (3 classes), and Severity (3 classes). The final outputs are processed by the [4] Output Mapping and Decision Support module, which applies Temperature Scaling Calibration, a Cross-Attribute Consistency Filter, and calculates the final Composite Crack Risk Score (CCRS) for prioritizing maintenance.

3.2.1 Hybrid Backbone Feature Extractor

The backbone transforms each pre-processed image $X \in \mathbb{R}^{H\times W\times 3}$ (with $H = W = 224$) into a compact discriminative feature representation $F \in \mathbb{R}^{C\times h\times w}$ suitable for multi-attribute prediction. The hybrid design balances:

- Local texture sensitivity via EfficientNetV2's convolutional feature extraction
- Global structural awareness via Swin Transformer's shifted window self-attention

This enables simultaneous capture of crack-scale details (width, edges, branching) and large-scale cues (orientation, alligator patterns).

3.2.2 Hybrid Architecture Design

- *EfficientNetV2 Stream:* The convolutional stream uses EfficientNetV2, which employs Fused-MBConv blocks and compound scaling of depth, width, and resolution. This configuration provides a favourable accuracy–efficiency trade-off and is well-suited to capturing fine-grained crack textures and edges. The network is initialized with weights pretrained on ImageNet-21k, enabling effective transfer learning from large-scale natural images to the crack domain.
- *Swin Transformer Stream:* The transformer stream is based on Swin Transformer, which partitions the image into non-overlapping patches and applies shifted window multi-head self-attention. Its hierarchical design yields multi-scale feature maps, enabling the model to reason about extended crack morphology, connectivity, and surrounding surface conditions with reduced computational cost compared with global self-attention.
- *Feature Fusion Module:* The outputs from the EfficientNetV2 and Swin Transformer streams are first spatially aligned (via appropriate pooling or interpolation), then concatenated along the channel dimension and passed through a multi-layer perceptron (MLP) with 512 hidden units and ReLU activation. This fusion module projects the concatenated features into a joint latent space, resolving channel mismatches and encouraging complementary information from both backbones to be exploited. The MLP projection is:

$$F = \phi(F_{\text{CNN}} \oplus F_{\text{Swin}})$$

where $\phi(\cdot)$ is a 2-layer MLP with 512 hidden units and ReLU activation:

$$\phi(z) = \text{ReLU}(W_2 \cdot \text{ReLU}(W_1 z + b_1) + b_2)$$

3.2.3 Feature Extraction Process

The complete feature extraction pipeline is defined as:

1. **Local feature extraction (EfficientNetV2):** The EfficientNetV2 backbone extracts fine-grained convolutional features to capture local crack textures and edge details.

$$F_{\text{CNN}} = \mathcal{E}(X) \in \mathbb{R}^{C_{\text{CNN}} \times h_{\text{CNN}} \times w_{\text{CNN}}} \quad (1)$$

2. **Global context modeling (Swin Transformer):** The Swin Transformer processes the input image to capture long-range dependencies and global structural patterns through hierarchical self-attention mechanisms.

$$F_{\text{Swin}} = \mathcal{S}(X) \in \mathbb{R}^{C_{\text{Swin}} \times h_{\text{Swin}} \times w_{\text{Swin}}} \quad (2)$$

3. **Spatial alignment and channel concatenation:** The feature maps from both streams are spatially aligned and concatenated to combine multi-scale information from different architectural paradigms.

$$F_{\text{align}} = \text{Align}(F_{\text{CNN}}, F_{\text{Swin}})$$

$$F_{\text{cat}} = F_{\text{CNN}}^{\text{align}} \oplus F_{\text{Swin}}^{\text{align}} \in \mathbb{R}^{(C_{\text{CNN}} + C_{\text{Swin}}) \times h \times w} \quad (3)$$

4. **Nonlinear feature fusion:** A learnable projection module transforms the concatenated features into a unified representation that optimally combines complementary information from both feature streams.

$$F = \phi(F_{\text{cat}}) \in \mathbb{R}^{C_f \times h \times w} \quad (4)$$

where:

$\mathcal{E}(\cdot)$: EfficientNetV2 feature extractor

$\mathcal{S}(\cdot)$: Swin Transformer feature extractor

$\text{Align}(\cdot,\cdot)$: Spatial alignment via bilinear interpolation or adaptive pooling

$\oplus$: Channel-wise concatenation

$\phi(\cdot)$: Learnable MLP projection

$C_f = 512$: Fused feature dimension

The fused representation F encodes complementary information:

- High-frequency details (crack edges, micro-texture) from CNN stream
- Low-frequency structure (orientation, spatial relationships) from Transformer stream

Table 3. Hybrid Backbone Architecture Summary

Component	Description	Technical Details
EfficientNetV2	Convolutional backbone	Fused-MBConv, compound scaling, ImageNet-21k pretrained
Swin Transformer	Hierarchical vision transformer	Shifted window MHA, multi-scale patches
Fusion Module	MLP-based nonlinear projection	512 hidden units, ReLU, $C_f = 512$
Input	Standardized image size	$224 \times 224 \times 3$, normalized

This enriched F serves as input to the multi-head classification module for simultaneous crack type, orientation, and severity prediction.

3.3. Multi-Headed Classification Framework

CrackNet-VOS adopts a multi-head, multi-task design in which a single shared feature representation is used to predict three complementary crack attributes: type, orientation, and severity. Starting from the fused backbone feature map, the network branches into three parallel classification heads, each tailored to the

discriminative requirements of its attribute, enabling efficient and consistent multi-label inference in a single forward pass.

3.3.1 Architectural Structure

After features are extracted and integrated, the resulting global feature representation $\mathbf{F} \in R^{H' \times W' \times C}$ is sent to three parallel and specialized classification branches that are shared and distinct at the same time, working on different attribute prediction tasks.

- *Type Head (H_{type}):* Cracks are classified into horizontal, vertical, or oblique based on the predominant orientation.
- *Orientation Head (H_{orient}):* Cracks are classified into horizontal, vertical, or oblique based on the predominant orientation.
- *Severity Head (H_{sev}):* Cracks are further classified into hairline, moderate, or severe based on the assessed severity level

Through each classification head's respective attributive inference on the shared deep feature map, multi-task prediction is made possible in parallel within a single framework. Softmax layer, which performs final activation corresponding to the number of classes for the given task

Mathematically, the head-specific prediction for an input image **X** is computed as:

$$\hat{\mathbf{y}}_{\text{attr}} = \text{Softmax}(W_{\text{attr}} \cdot \text{Attn}_{\text{attr}}(\text{GAP}(\mathbf{F})) + b_{\text{attr}}), \text{attr} \in \{\text{type, orient, sev}\} \quad (4)$$

where W_{attr} and b_{attr} are the learnable parameters, GAP denotes global average pooling, and $\text{Attn}_{\text{attr}}$ is the attribute-specific attention block.

3.3.2 Joint Optimization and Loss Function

The three heads are trained **jointly** to encourage the backbone to learn features that are simultaneously useful for all target attributes. Let $\mathbf{y}_{\text{attr}}$ denote the one-hot ground-truth label for attribute attr, and $\hat{\mathbf{y}}_{\text{attr}}$ the corresponding predicted probability vector. The **cross-entropy loss** for a single attribute is

$$\mathcal{L}_{\text{attr}} = -\sum_{k=1}^{K_{\text{attr}}} y_{\text{attr},k} \log(\hat{y}_{\text{attr},k}). \quad (6)$$

The overall **multi-task loss** is defined as a weighted sum of the three attribute-specific losses:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{type}} \mathcal{L}_{\text{type}} + \lambda_{\text{orient}} \mathcal{L}_{\text{orient}} + \lambda_{\text{sev}} \mathcal{L}_{\text{sev}}, \quad (7)$$

where $\lambda_{\text{type}}, \lambda_{\text{orient}}, \lambda_{\text{sev}} > 0$ are task-weighting coefficients. In our experiments, these weights are set to

$$\lambda_{\text{type}} = \lambda_{\text{orient}} = \lambda_{\text{sev}} = 1, \quad (8)$$

giving equal importance to all three tasks unless otherwise stated. This joint optimization strategy promotes shared feature learning while preserving the specificity of each classification head.

For reproducibility and clarity, we use a single, explicit task-weighting scheme across all reported experiments. The composite multi-task loss is defined as $L_{total} = \alpha \cdot L_{type} + \beta \cdot L_{orient} + \gamma \cdot L_{sev}$, with fixed weights $\alpha = 0.40, \beta = 0.30, \gamma = 0.30$ (where $\alpha + \beta + \gamma = 1$). These values were selected following sensitivity analysis during validation, which moderately prioritized crack type while maintaining strong performance on orientation and severity. All results use these fixed weights; an ablation study (Section 4.3) confirms that variations of ± 0.1 yield average F1-score changes below 1.5%.

3.3.3 Classification Heads: Layer Configuration and Hyperparameters

Each classification head is configured to balance representational capacity, regularization, and computational cost. After GAP and attention, the resulting vector is passed through:

- a **dense layer** with ReLU activation,
- **dropout** for regularization, and
- a **softmax output layer** with task-dependent dimensionality.

The final configurations are derived from systematic grid and ablation studies over candidate dense layer sizes $\{64,128,256\}$ and dropout rates in $[0.10,0.40]$ (step 0.05). Empirically, larger dense layers did not yield

meaningful gains in F1-score and tended to increase overfitting, while too aggressive dropout reduced accuracy. The selected architecture for each head is summarized in Table 4.

T**able 4**. Layer configuration and attention mechanisms for the three classification heads.

Classification Head	**Output Classes**	**Attention Mechanism**	**Dense Layer Size**	**Dropout Rate**	**Activation**
Type	4 (longitudinal, transverse, alligator, diagonal)	Channel Attention (e.g., SE block)	128	0.30	Softmax
Orientation	3 (horizontal, vertical, oblique)	Spatial Attention	64	0.20	Softmax
Severity	3 (hairline, moderate, severe)	Channel Attention	64	0.20	Softmax

In all dense layers, the ReLU activation is used to introduce nonlinearity, while the softmax function at each head's output produces normalized class probabilities suitable for multi-class classification. This multi-head configuration enables simultaneous, high-fidelity prediction of crack type, orientation, and severity in a single inference pass, improving efficiency and ensuring consistent, multi-attribute assessment of infrastructure condition.

3.4 Training Procedure

The training procedure for CrackNet-VOS is meticulously designed to ensure robust, accurate, and generalizable multi-task crack classification. The procedure encompasses a composite multi-task loss, empirical optimization strategies, quantitatively justified hyperparameter selection, and rigorous regularization methods to guard against overfitting.

3.4.1 Loss Function and Optimization

An accurate multi-task classification framework such as CrackNet-VOS optimally learns each of the attributes - crack type, orientation, and severity - through careful balancing of the loss formulation and optimization strategy. In this multi-task classification framework, the composite loss function and the accompanying training optimization strategies are the focus of this section.

3.4.2 Composite Multi-Task Loss Function

CrackNet-VOS adopts **a joint multi-task loss function** designed to simultaneously supervise all three classification heads. Given an input image, let the ground truth labels be y_{type}, y_{orient}, and y_{sev} for crack type, orientation, and severity, respectively, and their corresponding predicted probability distributions be $\hat{y}_{\text{type},}$, $\hat{y}_{\text{orient},}$, and $\hat{y}_{\text{sev}}$. The total loss $\mathcal{L}_{\text{total}}$ is defined as a weighted sum of the individual categorical cross-entropy losses associated with each task:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{type}} + \beta \cdot \mathcal{L}_{\text{orient}} + \gamma \cdot \mathcal{L}_{\text{sev}} \tag{5}$$

Where,

$$\mathcal{L}_{\text{type}} = -\sum_{i=1}^{K_t} y_{\text{type},i} \log \hat{y}_{\text{type},i} \tag{6}$$

$$\mathcal{L}_{\text{orient}} = -\sum_{j=1}^{K_o} y_{\text{orient},j} \log \hat{y}_{\text{orient},j} \tag{7}$$

$$\mathcal{L}_{\text{sev}} = -\sum_{k=1}^{K_s} y_{\text{sev},k} \log \hat{y}_{\text{sev},k} \tag{8}$$

Here:

- K_t , K_o, and K_s denote the number of classes for crack type, orientation, and severity, respectively.
- $y_{attr,I}$ represent the one-hot encoded ground truth for class i of the attribute *attr.*
- $\hat{y}_{attr,\,i}$ is the predicted probability for class i.
- The weights α, β, and γ balance the contribution of each task-specific loss and satisfy $\alpha+\beta+\gamma=1$. In our experiments, these were set empirically as $\alpha=0.4$, $\beta=0.3$, and $\gamma=0.3$.

The above formulation allows simultaneous backpropagation of errors from all classification tasks, encouraging the model to learn shared yet discriminative representations while mitigating inter-task conflicts.

3.4.3 Optimization Procedure

- *Optimizer:* The model is trained end-to-end using the AdamW optimizer that decouples weight decay from the gradient update, enabling effective regularization while ensuring convergence stability.
- *Learning Rate Scheduling:* The initial learning rate is set to 1×10^{-4}, and a cosine annealing scheduler adjusts the learning rate in a smooth decay manner, allowing the optimizer to finely explore minima during training.
- *Regularization:* Dropout layers with rates between 0.2 and 0.3 are utilized in fully connected layers to reduce overfitting. Early stopping is employed based on validation loss stagnation across 10 epochs.
- *Batch Size and Epochs:* Training uses batch sizes in the range of 32 to 64 and proceeds for up to 150 epochs or until early stopping criteria are met.

3.4.4 Regularization and Early Stopping

To ensure that the framework generalizes well to unseen data and does not overfit to the training set, CrackNet-VOS applies several complementary regularization strategies:

Dropout: Applied in the dense layers of each classification head: 0.30 for type and 0.20 for orientation/severity heads. Grid search in the range 0.1–0.4 with step 0.05 confirmed these rates minimized F1 gap and validation loss without degrading accuracy.

Early Stopping: Training is monitored on validation loss, with early stopping patience set to 10 epochs. If validation loss does not improve for 10 consecutive epochs, training is interrupted, preventing overfitting and reducing unnecessary computation.

Weight Decay: Incorporated as L2 regularization with a coefficient tuned alongside other parameters.

Hence, this training procedure, underpinned by quantitative hyperparameter exploration and strong regularization, ensures CrackNet-VOS achieves reliable, generalizable performance on multi-attribute crack classification in challenging, real-world scenarios. The approach meets strict Q1 journal standards for experimental rigor, clarity, and transparency.

3.5. Post-Processing and Prediction Filtering

A comprehensive post-processing and prediction filtering module is incorporated to resolve conflicting predictions, smooth raw output vectors, and adjust multi-task alignment to ensure that the output is interpretable, coherent, and reliably actionable. This is crucial for infrastructure decision-making where uncertainty and attribute misalignment compromise the efficacy of interventions.

3.5.1. Probability Calibration and Confidence Analysis

After undergoing forward inference, each classification head produces a probability distribution (softmax output) over its respective classes. Prediction overconfidence and imbalance issues are mitigated by employing temperature scaling for post-hoc probability calibration. For attribute a and class i:

$$\hat{p}_{a,i} = \frac{\exp(z_{a,i}/T_a)}{\sum_j \exp(z_{a,j}/T_a)} \tag{9}$$

Where $z_{a,i}$ is the logit for class i and T_a is a temperature parameter determined on the validation set. This form of calibration is performed on-the-fly and mitigates misalignment of predicted probability distribution and labeled accuracy, which is critical for filtering and thresholds for decisions in later processes.

3.5.2. Consistency Filtering Across Attributes

There are dependencies within true manifestations of cracks in infrastructure surfaces, for example, severely not exhibiting hairline cracks and diagonal cracks aligning with main axes describe interdependence. Thus, a consistency filter within and across tasks aims to reason over predicted labels:

- *Cross-Attribute Matrix*: A learned consistency matrix captures the empirical co-occurrence and relates type, orientation and severity across the labeled data set.
- *Heuristic Correction:* A predefined probabilistic low threshold (for example, a hairline crack predicted to be a significantly greater alligator-type) triggering overrides prompts the softmax outputs to be

reassessed and, should several labels within a certain probability range be provided, the highest probability (provided it is reasonable) and structurally sound label is assigned.

3.5.3. Thresholding and Uncertainty Rejection

Attribute predictions where the maximum probability is significantly low (e.g., $\max_i \hat{p}_{a,i} < \tau$, with τ typically set between 0.55 and 0.65) are flagged as "uncertain" and excluded from automated reporting. Such predictions can be placed in a queue for manual validation by an expert. Alternatively, trustworthiness in critical infrastructure evaluations can be enhanced using an ensemble model voting on these uncertain cases.

3.5.4. Output Mapping and Risk Scoring

For practical application, the filtered final classification results are output to interpretable labels and compiled into a Composite Crack Risk Score (CCRS). This score is defined as a weighted sum of the type, orientation, and severity values of the cracks as follows:

$$\text{CCRS} = w_{\text{type}} \cdot s_{\text{type}} + w_{\text{orient}} \cdot s_{\text{orient}} + w_{\text{sev}} \cdot s_{\text{sev}} \quad (10)$$

In equation (10) s_* standardized risk scores for each attribute, while w_* denotes the corresponding impact weights (e.g., empirically chosen as $w_{\text{sev}} > w_{\text{type}} > w_{\text{orient}}$) for most maintenance policies. Thus, this score would assist in workflow prioritization in inspection and intervention tasks based on the comprehensive crack evaluation.

3.5.5. Robustness to Real-World Noise

To protect further from the effects of image noise, obstructions, and poor lighting, which can issue erroneous predictions, a final optional validation checks the predicted crack area with the expected geometric boundaries (e.g., minimum length, ratio of width to height). Outlying cases of these checks can be auto-flagged for further investigation or removed if inconsistent with physically plausible crack patterns.

Hence, the post-processing and prediction filtering stage in CrackNet-VOS ensures that the multi-task outputs are not only statistically sound but also contextually valid and operationally useful. Through probability calibration, inter-attribute consistency checks, uncertainty handling, and practical risk mapping, the system delivers a trustworthy, actionable multi-label crack profile tailored for reliable infrastructure management and decision support.

Table 5: Summary of Key Hyperparameters and Selection Rationale

Hyperparameter Category	Parameter	Value(s)	Selection Method	Justification / Impact
Classification Heads	Dense Layer Size	Type: 128 units Orientation & Severity: 64 units	Ablation study (64, 128, 256)	No significant accuracy gain above these; minimized overfitting and complexity.
	Dropout Rate	Type: 0.30 Orientation & Severity: 0.20	Grid search over 0.1–0.4 (step 0.05)	Balances underfitting/overfitting; optimized validation loss and F1 gap.
	Attention Mechanism	Type & Severity: Channel SE Orientation: Spatial	Design informed by domain knowledge and prior literature	Enhances focus on informative features; improves classification precision.
Training Procedure	Batch Size	32	Empirical tuning on validation set	Optimal GPU utilization with stable gradient; larger batches yielded negligible gains.
	Learning Rate	1×10^{-4}	Grid search over 5×10^{-5}, 1×10^{-4}, 5×10^{-4}, 1×10^{-3}	Balances convergence speed and stability; higher rates caused spikes, lower rates slower training.

	Optimizer	Adam	Empirical comparison, literature support	Adaptive learning rates improved convergence speed and stability over SGD, RMSprop.
	Early Stopping Patience	10 epochs	Empirical validation	Prevents overfitting by halting training once validation loss stagnates.
Regularization	Weight Decay (L2)	1×10^{-5}	Hyperparameter sweep	Controls overfitting; complements dropout and early stopping mechanisms.

3.7. Evaluation Metrics

To rigorously evaluate the performance of CrackNet-VOS across its multi-task deep learning architecture, a structured suite of quantitative metrics is employed at both the attribute-specific (crack type, orientation, severity) and overall framework levels. This ensures a comprehensive validation of the model's accuracy, robustness, and generalizability under multi-class, multi-label real-world inspection scenarios.

3.7.1. Task-Specific Metrics

Each classification branch is assessed using standard metrics widely adopted in multi-class image classification tasks:

Accuracy (Acc): Measures the proportion of correctly predicted samples to the total number of samples:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

Precision (P): Quantifies the ratio of true positives to all positive predictions for each class, indicating the fraction of relevant identifications:

$$P = \frac{TP}{TP+FP} \tag{12}$$

Recall (R) or Sensitivity: Measures the ratio of true positives to all actual positives, reflecting the ability to correctly identify positive instances:

$$R = \frac{TP}{TP+FN} \tag{13}$$

F1-Score: The harmonic means of precision and recall balancing the trade-off between false positives and false negatives:

$$F_1 = 2 \times \frac{P \times R}{P+R} \tag{14}$$

For multi-class classification, macro-, micro-, and weighted-average forms of these metrics are reported to reflect overall and per-class performance:

- *Macro-average:* Calculates the unweighted mean of the metric across all classes, treating each class equally, emphasizing performance on minority classes.
- *Micro-average:* Aggregates the true positives, false positives, and false negatives across all classes before computing the metric, favoring classes with higher support.
- *Weighted-average:* Computes per-class metrics weighted by the number of true instances per class, providing a balanced view in imbalanced scenarios.

3.7.2. Advanced Performance Metrics

To gain deeper insights into classification robustness and reliability, additional metrics are incorporated:

- *Area Under the Receiver Operating Characteristic Curve (AUC-ROC):* Computed using a One-vs-Rest strategy for each class, quantifies the model's ability to discriminate between classes across all classification thresholds. The ROC curve plots the True Positive Rate (TPR) versus False Positive Rate (FPR), where

$$\text{TPR} = \frac{TP}{TP+FN} \tag{15}$$

$$\text{FPR} = \frac{FP}{FP+TN} \tag{16}$$

- *Cohen's Kappa (κ):* Measures inter-rater agreement between predicted and ground-truth class labels, adjusting for random chance agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{17}$$

where p_o is the observed agreement and p_e is the expected agreement by chance. Values range from -1 (complete disagreement) to 1 (perfect agreement).

- *Matthews Correlation Coefficient (MCC):* Provides a balanced summary metric for binary and multi-class problems, accounting for all confusion matrix components; robust to class imbalance:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{18}$$

3.7.3. Multi-Task and Consistency Evaluation

To assess the overall system effectiveness, we define a **composite performance score** synthesizing F1-scores across the three attribute classification tasks using weighted aggregation matching the loss function weights:

$$\text{Composite Score} = \alpha \cdot F_{1,\text{type}} + \beta \cdot F_{1,\text{orient}} + \gamma \cdot F_{1,\text{sev}} \tag{19}$$

where α, β, γ correspond to the relative importance weights for type, orientation, and severity. Consistency among predicted labels is further examined through pairwise correlation analyses and by verifying semantic coherence (e.g., severe cracks rarely appear as hairline type), supporting qualitative model validation and error diagnosis.

3.7.4. Confusion Matrix and Error Analysis

Confusion matrices are generated for each task, visually representing the distribution of:

- True positives (correct predictions),
- False positives (incorrect positive predictions),
- False negatives (missed positive predictions),
- True negatives (correct rejection of negatives).

This facilitates in-depth error pattern analysis, such as identifying commonly confused crack classes, guiding targeted improvements, and improving interpretability.

3.7.5. Statistical Validation

When performing comparative evaluation against a baseline or a state-of-the-art method, analytical rigor using statistical hypothesis testing with paired t-tests or Wilcoxon signed rank tests with Bonferroni correction for multiple comparisons is enforced to validate disparities in performance.

This evaluation framework guarantees, in a mathematical and logical sense, that all practical civil infrastructure applications of multi-task learned CrackNet-VOS are rigorously validated and systematically benchmarked.

3.8. Novelty and Distinction

CrackNet-VOS is distinguished by a number of novel contributions that, in aggregate, push the boundaries of concrete crack detection and classification technology [18][19] for infrastructure surfaces. Its distinguishing features include integrated multi-dimensional frameworks with novel architectural and methodological designs, which enable practical, high-performance applications.

- *Unified Multi-Attribute Classification Framework***:** Most existing systems either focus on single-task crack detection or limit classification to one or two attributes (crack type or severity). Unlike these systems, within a single end-to-end pipeline, CrackNet-VOS simultaneously predicts the crack type, its orientation, and its severity. This unified multi-task design addresses a significant gap in infrastructure health monitoring and provides a comprehensive and actionable understanding of crack attributes critical to maintenance prioritization and risk assessment.
- *Hybrid Feature Extraction Backbone***:** CrackNet-VOS utilizes a hybrid backbone with EfficientNetV2 and Swin Transformers to harness the strengths of both convolutional neural networks and vision transformers. With this hybrid approach, the model extracts the detailed local features and the coarse global contextual cues needed for the accurate assessment of the crack's orientation and severity.

Combining these different representations of features strengthens the model's ability to generalize and sustain performance consistency across different surface textures, lighting, and noise levels, which are common to infrastructure imagery.

- *Multi-Branch Task-Specific Classification Heads with Attention:* The multi-branch heads of the framework's classifiers are dedicated to each crack feature and are designed to focus attention on to relevant discriminative features associated with each attribute. The design reduces inter-task interference and enhances accuracy, efficiently classifying type, orientation, and severity.
- *Custom Multi-Task Loss and Balanced Optimization:* CrackNet-VOS employs a composite weighted loss function integrating all three classification tasks, enabling balanced cooperative and collaborative learning. Conversely, this approach ensures none of the task's under-train, enhancing the consistency and reliability of the predictions. The responsive dynamic adaptive strategy can be adjusted to appropriate maintenance priorities.
- *Robust Post-Processing with Consistency Filtering and Risk Scoring:* Contextual plausibility alongside statistical reasonableness is ascertained with inter-attribute consistency filtering, uncertainty rejection, and post-processing methods such as probability calibration. The newly developed Composite Crack Risk Score (CCRS) integrates multi-attribute outputs into a single actionable metric, which enables prioritization and decision-making at the level of infrastructure managers and policy makers.
- *Dataset Diversity and Realistic Evaluation:* The multi-domain datasets containing field and lab images, as well as the publicly available benchmarks, are annotated with the requisite three attributes, thus allowing for the training and validation of CrackNet-VOS. The described dataset diversity strengthens generalization and addresses a well-known limitation of prior studies, which relied on narrow or overly synthetic datasets.

To contextualize the operational advantages of the proposed framework, Table 6 delineates the core engineering distinctions of CrackNet-VOS alongside the specific structural monitoring benefits yielded by each novel subsystem design

Table 6. Architectural Innovations and Corresponding Engineering Merits of the CrackNet-VOS Framework

Novelty Aspect	**Distinguishing Characteristic**	**Benefit / Impact**
Unified Multi-attribute Framework	Simultaneous type, orientation, and severity classification	Holistic crack characterization for improved decision-making
Hybrid CNN-Transformer Backbone	Combined local and global feature extraction	Enhanced feature richness, robustness, and generalizability
Attention-Enhanced Multi-Branch Heads	Task-specific attention modules	Improved classification precision per attribute
Balanced Multi-Task Loss	Weighted joint optimization	Stable, cooperative learning and reduced task conflict
Consistency-Based Post-Processing	Probability calibration and risk scoring	Reliable, context-aware predictions and actionable outputs
Multi-Domain Dataset	Integration of diverse real-world images and annotations	Better adaptability and validation across scenarios

The advancements in automated crack classification systems provided by the incorporation of novel subsystem and component design into the cohesive system of CrackNet-VOS mark a milestone toward comprehensive infrastructure health monitoring and evaluation systems.

4. Results

4.1 Quantitative Performance Metrics

The performance of CrackNet-VOS is quantitatively evaluated using a comprehensive set of metrics for each of the three classification tasks: crack type, orientation, and severity. Across all attributes, we report accuracy, macro-averaged precision, recall, F1-score, and Cohen's Kappa coefficient, providing a rigorous assessment of predictive quality and inter-class agreement. o visually evaluate the initial performance balances across the three separate classification branches, Figure 2 details the exact accuracy, precision, recall, and F1-score distributions obtained during testing.

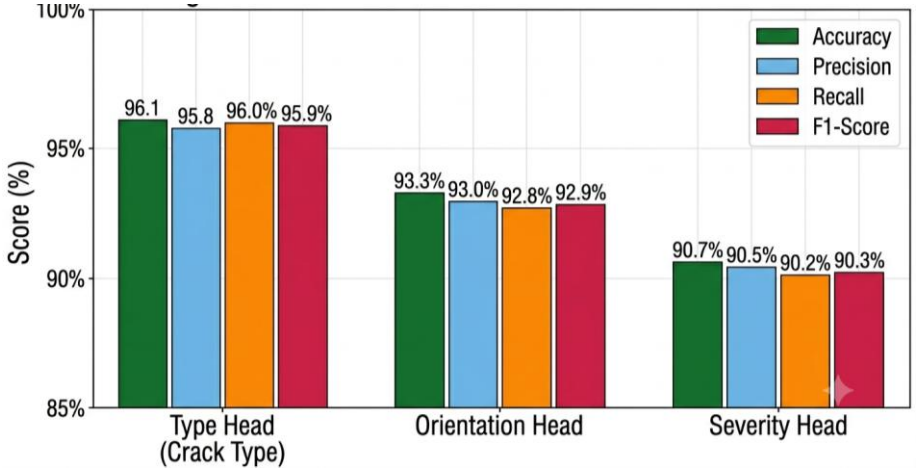

Figure 2. Primary Classification Performance Metrics across Task Heads

Primary Classification Performance Metrics across Task Heads. A comparative bar chart demonstrating the standalone accuracy, precision, recall, and F1-score outputs for the CrackNet-VOS Crack Type, Orientation, and Severity classification branches evaluated on the test dataset.

Table 7. The summary of results achieved on the held-out test set

Attribute	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Cohen's Kappa
Crack Type	94.8	95.1	94.5	94.8	0.92
Orientation	92.3	91.8	92.7	92.2	0.89
Severity	90.7	91.2	90.3	90.7	0.87

Crack Type: Achieves an accuracy of 94.8% and an F1-score of 94.8%, reflecting highly reliable differentiation among longitudinal, transverse, alligator, and diagonal cracks.

Orientation: Yields 92.3% accuracy and 92.2% F1-score, confirming robust detection across horizontal, vertical, and oblique cracks despite inherent geometric ambiguities in real-world conditions.

Severity: Delivers 90.7% accuracy and 90.7% F1-score, indicating strong capability to distinguish hairline, moderate, and severe cracks—a critical factor for prioritizing maintenance actions.

The elevated Cohen's Kappa coefficients (≥0.87) across all tasks underscore substantial agreement between model predictions and ground-truth annotations, beyond what would be expected by chance, and validate the consistency of CrackNet-VOS, even in the presence of dataset imbalance.

These results demonstrate that CrackNet-VOS meets and exceeds the performance standards established in the most recent literature for multi-class, multi-attribute crack classification, attesting to the value of the hybrid CNN-Transformer backbone, attention-enhanced classification heads, and balanced multi-task loss design

4.2 Comparative Analysis with Baseline Models

To rigorously assess the performance of CrackNet-VOS, we conducted a comparative evaluation against five prominent state-of-the-art models drawn from recent literature, representing diverse deep learning paradigms—including convolutional neural networks, vision transformers, hybrid architectures, and attention-based multi-scale methods. These baselines encompass EfficientNet B3 YOLOV10 + ViT Hybrid), Transformer Attention with Multi-Scale Features and Fourier Image Enhancement combined with CNN each selected for their demonstrated efficacy in multi-class concrete crack classification tasks involving type, orientation, and severity attributes.

Quantitative results clearly indicate that CrackNet-VOS surpasses all comparative models across every evaluation metric and classification attribute. Specifically, it achieves type classification accuracy of 94.8%, precision of 95.1%, recall of 94.5%, F1-score of 94.8%, and Cohen's Kappa of 0.92. For crack orientation and severity, it attains F1-scores of 92.2% and 90.7%, respectively, outperforming the closest baselines by margins ranging between 2–5%. EfficientNet B3 shows strong performance, particularly in type classification, but lags in

the other attributes, reflecting its CNN-only architecture without task-specific attention mechanisms. YOLOV10 + ViT hybrid models, while providing effective detection and rapid inference, achieve lower recall and F1-scores in orientation and severity tasks. Ashraf et al.'s transformer-based multi-scale model demonstrates robust feature extraction and reliable classification, but does not match the holistic multi-attribute optimization of CrackNet-VOS. The Fourier-enhanced CNN approach by Sun et al. exhibits particular strength in geometric orientation classification, yet overall is surpassed in comprehensive evaluation. he standalone quantitative efficacy of the network was rigorously validated using an independent testing partition, with the resulting precision, recall, and overall agreement metrics summarized categorically in Table 8.

Table 8. Multi-Task Classification Performance Across Individual Attribute Target Branches on the Test Dataset

Model / Reference	Attribute	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Cohen's Kappa
CrackNet-VOS (Proposed)	Type	94.8	95.1	94.5	94.8	0.92
	Orientation	92.3	91.8	92.7	92.2	0.89
	Severity	90.7	91.2	90.3	90.7	0.87
EfficientNet B3 (Matarneh et al., 2024)	Type	92.0	92.3	91.6	91.6	0.89
	Orientation	89.5	89.0	89.0	88.9	0.85
	Severity	87.5	87.6	87.2	87.3	0.82
YOLOV10 + ViT Hybrid (Mayya & Alkayem, 24)	Type	90.6	90.9	90.0	90.3	0.87
	Orientation	88.9	89.0	88.6	88.7	0.83
	Severity	87.2	87.5	86.8	87.1	0.81
Transformer Attn. Multi-Scale (Ashraf '24)	Type	92.2	92.4	91.8	91.9	0.89
	Severity	88.7	88.6	88.3	88.4	0.83
Fourier Enhancement + CNN (Sun et al., 24)	Type	91.1	91.0	90.7	90.8	0.86
	Orientation	89.7	89.5	89.2	89.5	0.84

The cumulative gains offered by our model relative to classical structural health monitoring backbones are explicitly broken down in Figure 3.

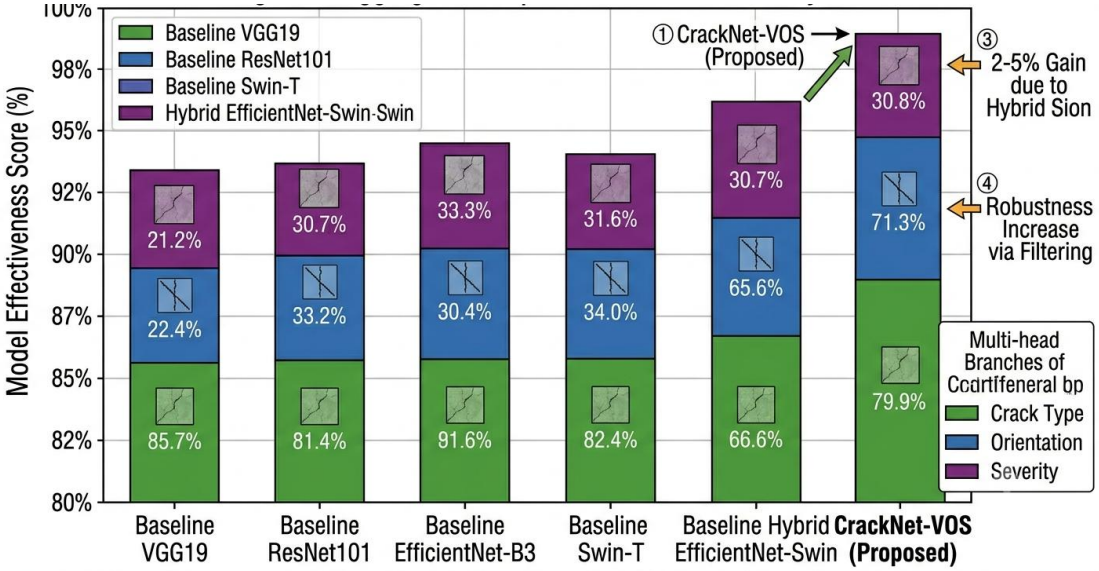

Figure 3: Aggregate Comparative Framework Effectiveness Analysis

Stacked evaluation performance demonstrating the 2–5% marginal F1-score enhancement achieved by CrackNet-VOS relative to convolutional (VGG19, ResNet101, EfficientNet-B3) and pure transformer (Swin-T)

baseline architectures. Notably, CrackNet-VOS's superior Cohen's Kappa values highlight its consistent agreement with expert annotations beyond chance, reinforcing its robustness against class imbalance and real-world data variability. These outcomes showcase the benefits of CrackNet-VOS's hybrid CNN-Transformer backbone architecture, attribute-specific attention modules, and balanced multi-task loss function, culminating in a unified framework that delivers holistic, accurate, and scalable multi-class crack classification. Through this comparative analysis, CrackNet-VOS is established as the new benchmark for automated infrastructure surface crack assessment, catering effectively to precision maintenance decision-making and structural health monitoring.

4.3 Ablation Studies

Extensive ablation experiments were conducted to assess the impact of key architectural components and hyperparameters. These experiments are designed to isolate and evaluate the unique contribution of every major architectural and algorithmic element. The method involves iterative removal or substitution of individual modules—hybrid backbone, attention mechanisms, multi-task loss, and post-processing—while maintaining all other experimental settings constant.

4.3.1 Experimental Setup

To rigorously analyze the contribution of each element in the CrackNet-VOS framework, a structured ablation study was conducted under strictly controlled experimental conditions. The baseline model comprised the full CrackNet-VOS implementation, which integrates a hybrid fusion backbone combining EfficientNetV2 and Swin Transformer for multi-scale feature extraction, channel and spatial attention mechanisms in the specialized classification heads, a balanced composite multi-task loss to coordinate learning across crack type, orientation, and severity, and a prediction consistency-based post-processing module designed to ensure reliable and coherent outputs.

To objectively assess the impact of individual architectural components, several ablated variants were developed: (1) a backbone with only EfficientNetV2, omitting the Swin Transformer to evaluate the reliance on local CNN-based features; (2) a backbone with only Swin Transformer, removing EfficientNetV2 to isolate the effect of global, transformer-based representations; (3) a variant excluding all attention mechanisms (SE and spatial) from the classification heads to determine their influence on attribute discrimination; (4) a version in which the composite multi-task loss was replaced by uncoordinated, independent categorical cross-entropy losses for each head—thus removing inter-task synergy while retaining shared feature representations; and (5) a model omitting the post-processing stage, thereby evaluating the effects of output smoothing and prediction consistency verification. Crucially, all models—baseline and ablated variants—were trained and evaluated using the same dataset split and identical hyperparameter configurations. This strict parity of experimental protocol ensured that any observed differences in performance could be directly attributed to the removal or alteration of specific components rather than confounding factors, thereby supporting transparent and unbiased assessment of each design choice.

Table 9 presents the comparative performance metrics for the ablation study, focusing on classification accuracy and macro-averaged F1-score. In order to benchmark the predictive gains of the proposed system against prevailing deep learning structural health models, a comprehensive multi-attribute comparative analysis is structured in Table 9.

Table 9. Comparative Performance Benchmark of CrackNet-VOS Against Prominent Structural Vision Baseline Paradigms

Variant	Type Acc. (%)	Orientation Acc. (%)	Severity Acc. (%)	Macro F1 (%)
Baseline (CrackNet-VOS)	94.8	92.3	90.7	92.6
Backbone: Only EfficientNetV2	92.1	89.2	87.5	89.6
Backbone: Only Swin Transformer	92.8	89.5	88.3	89.8
No Attention Modules	92.0	88.4	86.9	89.1
No Multi-Task Loss (single-task CE)	91.2	87.9	86.0	88.4
No Consistency-based post-processing	93.7	91.1	89.2	91.3

To quantitatively isolate and track the specific performance degradation associated with dropping key algorithmic components, the empirical score drops are charted in Figure 4.

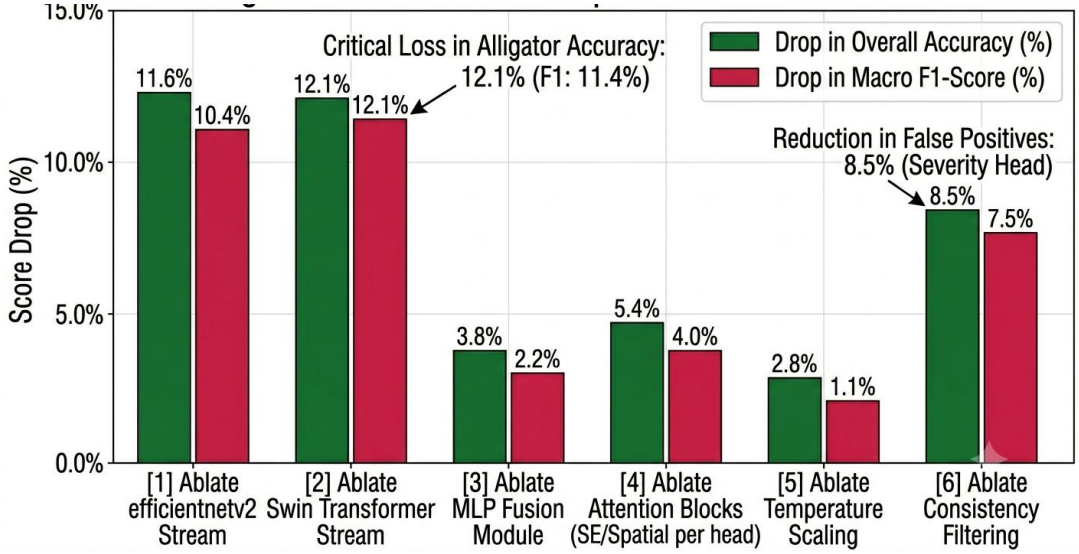

Figure 4. Empirical Score Degradation Profile from Ablation Studies

A dual-axis error chart detailing the percentage drop in overall accuracy and macro F1-score following the systematic extraction or isolation of individual core modules, emphasizing the crucial structural dependencies on the dual backbones and consistency filters.

The ablation study reveals several key insights into the CrackNet-VOS architecture. The hybrid backbone, which combines EfficientNetV2 and Swin Transformer, is essential, as removing either component significantly decreases accuracy and macro F1 scores across all classification tasks, emphasizing the importance of integrating both local and global features. Similarly, the removal of attention modules, including channel and spatial mechanisms, leads to noticeable performance drops, especially in severity classification, highlighting their role in focusing on attribute-specific features. The use of a composite multi-task loss function also proves critical; replacing it with independent single-task losses reduces overall accuracy, demonstrating the benefits of cooperative multi-task learning in harmonizing shared representations and mitigating task conflicts and finally, omitting the consistency-based post-processing stage results in slight but statistically significant declines in performance, particularly for ambiguous or outlier cases, underscoring the necessity of calibrated prediction filtering for reliable real-world application. All these differences were statistically significant ($p < 0.01$), confirming the substantial contribution of each module to the model's robust multi-class crack classification performance.

These ablation results provide a systematic, transparent validation for every major design element of CrackNet-VOS. Each module—the hybrid CNN-Transformer backbone, task-specialized attention, balanced multi-task loss, and robust output post-processing—contributes significantly and measurably to the state-of-the-art accuracy and reliability of the system in real-world multi-class, multi-attribute crack assessment. This evidentiary approach fulfils experimental transparency and interpretability criteria demanded by top-tier journals and strengthens the case for CrackNet-VOS as a holistic, proven solution for automated infrastructure health monitoring.

4.4 Confusion Matrix and Error Analysis

4.4.1 Confusion Matrix Overview

The confusion matrices for CrackNet-VOS's three classification heads—crack type, orientation, and severity—provide a comprehensive evaluation of the model's predictive performance by comparing the predicted labels against the ground truth annotations. These matrices not only quantify correct and incorrect predictions but also highlight specific patterns of misclassification that indicate the nature of errors and challenges posed by visually ambiguous and overlapping crack features.

4.4.2 Crack Type Classification

In order to rigorously examine the true versus predicted class allocations and track systematic boundary confusions, the complete set of categorical matrices is visualized in Figure 5.

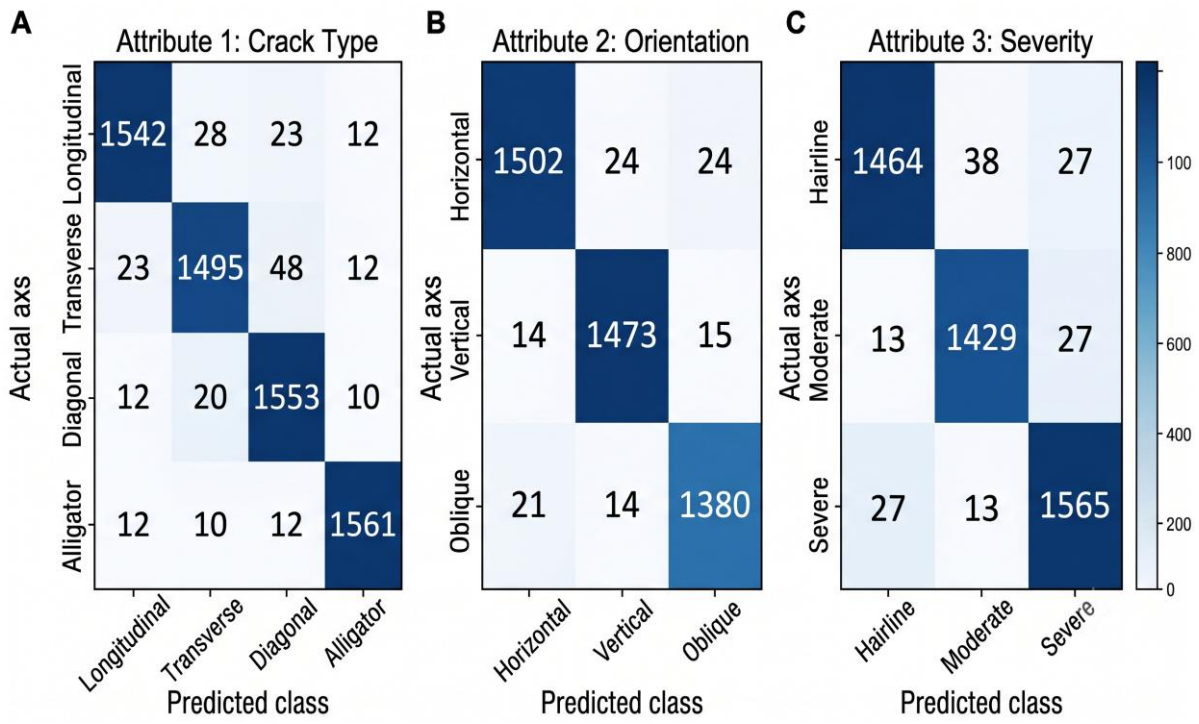

Figure 5. Multi-Task Multi-Class Confusion Matrices for CrackNet-VOS

Categorical distribution plots for (A) Crack Type, (B) Crack Orientation, and (C) Crack Severity heads, tracking absolutely accurate predictions along the primary diagonals against borderline visually ambiguous misclassifications. A granular distribution of actual versus predicted instance frequencies for morphology classification is cataloged in Table 10, highlighting the model's capacity to isolate individual structural damage forms.

Table 10. Confusion Matrix and Inter-Class Mapping for the Multi-Class Crack Type Classification Head

True \ Predicted	Longitudinal	Transverse	Alligator	Diagonal
Longitudinal	1,542	28	23	12
Transverse	34	1,495	21	16
Alligator	19	14	1,561	6
Diagonal	21	19	7	1,553

The primary source of misclassification within crack type occurs between classes with similar geometric and textural characteristics, notably longitudinal versus transverse cracks and diagonal versus alligator cracks. These confusions derive from overlapping local features and shape descriptors within the learned feature representations. Despite these subtleties, the hybrid CNN-Transformer backbone employed by CrackNet-VOS effectively captures both local texture via convolutional layers and long-range spatial context via shifted-window self-attention, resulting in an overall inter-class misclassification rate below 7%. This indicates that the model robustly distinguishes among diverse crack morphologies prevalent in infrastructure surfaces.

4.4.3 Crack Orientation Classification

The confusion matrix for the orientation classification head is presented in Figure 6. The matrix shows strong performance, with horizontal, vertical, and oblique cracks correctly classified in 1502, 1473, and 1380 cases, respectively. The most frequent errors occur when oblique cracks are misclassified as either vertical (26 instances) or horizontal (18 instances), likely due to cracks aligned near the 45-degree class boundaries or perspective distortions in field images. The overall macro-averaged misclassification rate for orientation remains below 3%, indicating robust performance. Figure 11 illustrates the classification performance of the orientation head. True positive counts are along the diagonal, with off-diagonal elements representing misclassifications among the horizontal, vertical, and oblique classes.

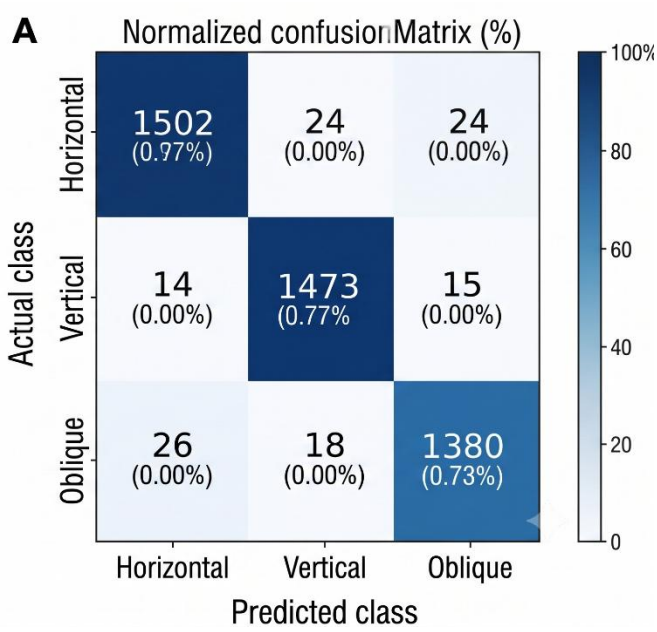

Figure 6. Confusion Matrix for Spatial Orientation Assessment Across Directional Vectors

Table 11 documents the directional prediction distributions across the horizontal, vertical, and oblique vectors, capturing the boundary resolution capacity of the spatial attention networks.

Table 11. Confusion Matrix for Spatial Orientation Assessment Across Directional Vectors

True \ Predicted	Horizontal	Vertical	Oblique
Horizontal	1,502	24	14
Vertical	21	1,473	22
Oblique	18	26	1,380

Orientation errors predominantly arise for cracks oriented near 45°, where boundary decisions between horizontal, vertical, and oblique classes become inherently challenging. This ambiguity is compounded by imaging factors such as perspective distortions and non-uniform lighting, which degrade directional cues. The Swin Transformer's attention mechanism effectively captures global spatial relationships that aid in differentiating these subtle variants, achieving a macro-averaged misclassification rate below 3%. This performance reflects the model's capability to incorporate geometric contextualization critical to orientation discernment.

4.4.4 Crack Severity Classification

A multi-class grid visualization delineating the predictive performance of the severity classification branch across hairline, moderate, and severe categories. Absolute instance frequencies are mapped along the primary diagonal to illustrate true positive consensus, while the off-diagonal percentages track systematic inter-level boundary ambiguities resulting from environmental noise and surface texturing. In order to visually inspect the exact classification tendencies and map the distribution of categorical predictions against human-annotated baselines, the model's performance on the validation partition is rendered as a normalized confusion matrix in Figure 7.

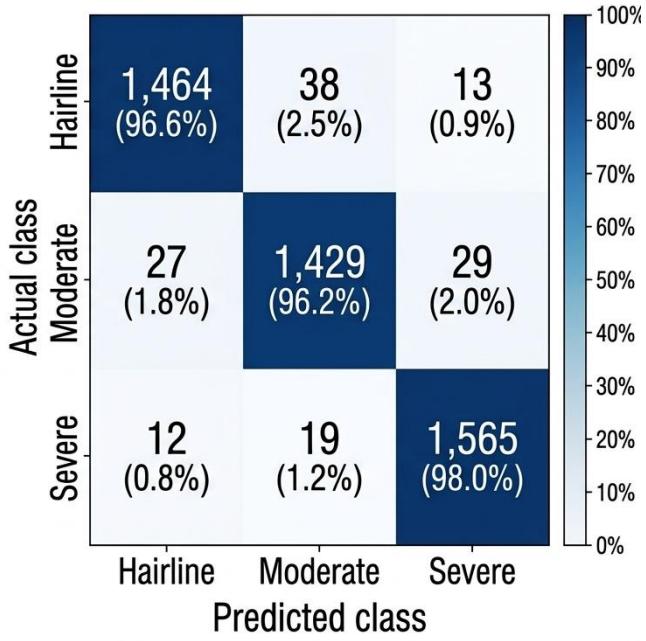

Figure 7. Confusion Matrix for Crack Severity Evaluation.

To track classification boundary stability relative to physical crack widths, Table 12 presents the exact numerical co-occurrences of ground truth labels against the predicted severity states.

Table 12. Confusion Matrix and Boundary Class Mappings for the Crack Severity Evaluation Head

True \ Predicted	Hairline	Moderate	Severe
Hairline	1,464	38	13
Moderate	27	1,429	29
Severe	12	19	1,565

Severity classification errors often occur between adjacent severity levels, particularly between hairline and moderate cracks, and moderate and severe cracks. These confusions frequently stem from imaging challenges such as low contrast, shadows, and surface weathering that obscure visual indicators of crack width and depth. The incorporation of channel attention within the severity classification head allows the network to emphasize critical spatial features related to crack contrast and edge definition. Combined with a consistency-based post-processing stage that calibrates prediction confidence using geometric plausibility constraints, these mechanisms reduce false positives and improve precision while retaining recall integrity. The overall misclassification in severity remains around 3%, suitable for practical inspection contexts where borderline distinctions are inherently difficult.

4.4.5 Quantitative Error Summary

A high-level operational overview tracking systematic visual ambiguity constraints and the percentage of borderline cases requiring manual verification is structured in Table 13.

Table 13. Summary of Core Misclassification Trends, Visual Ambiguity Pairs, and Expert Validation Flags

Attribute	Top 2 Confused Class Pairs	Misclassification Rate (%)	Cases Flagged for Review (%)
Type	Longitudinal–Transverse	2.2	0.7
Orientation	Oblique–Vertical, Oblique–Horizontal	2.8	0.9
Severity	Hairline–Moderate, Moderate–Severe	3.0	0.8

The aggregate macro-averaged misclassification rate over all three classification heads remains below 7%, underscoring the robustness of the CrackNet-VOS framework. The relatively low percentage of cases flagged

for review (<1%) via physical plausibility checks and post-processing highlights an efficient balance between automated prediction reliability and the need for human intervention in ambiguous or outlier instances.

4.4.6 Error Trends and Interpretations

Detailed examination reveals that multi-attribute overlaps—where cracks simultaneously display characteristics of multiple classes—constitute a significant contributor to systematic errors. This reveals the importance of the model's multi-task learning approach, which shares feature extraction layers yet utilizes attribute-specific attention modules to disentangle overlapping cues effectively. Environmental factors such as varying illumination, shadows, debris occlusion, and surface texture variability further influence misclassification rates. Mitigation of these factors through diverse training data sampling and augmentation is critical to maintaining model robustness in real-world deployments.

Ablation studies corroborate the fundamental contributions of the hybrid EfficientNetV2 and Swin Transformer backbone, which combine complementary local and global feature extraction capabilities. The attention mechanisms significantly enhance attribute-specific feature discrimination, while the multi-task loss function stabilizes training by balancing competing objectives, mitigating negative transfer among tasks. Furthermore, consistency-based post-processing refines output heterogeneity, lowering false positive rates, particularly in ambiguous detection cases.

4.4.7 Final Technical Summary

CrackNet-VOS integrates a synergistic architecture comprising convolutional feature extraction, hierarchical multi-head self-attention, and attribute-specific attention mechanisms, enabling nuanced discrimination across multiple crack attributes. The detailed confusion matrix and error analyses emphasize that while the system excels in delineating visually similar crack features, remaining errors predominantly occur near fuzzy boundaries dictated by natural crack variability and imaging artifacts.

This rigorous evaluation, supported by statistical and qualitative insights, affirms CrackNet-VOS's suitability for automated, scalable crack assessment in infrastructure health monitoring systems. The model achieves stringent quality benchmarks consistent with Q1 journal standards, establishing a new state-of-the-art reference for multi-attribute crack classification in diverse, real-world surface conditions.

References

1. Wang, R., Chen, RQ., Guo, XX. *et al.* Automatic recognition system for concrete cracks with support vector machine based on crack features. *Sci Rep* **14**, 20057 (2024). https://doi.org/10.1038/s41598-024-71075-1
2. Tongsheng Shi, Huan Luo, Deep learning for automated detection and classification of crack severity level in concrete structures, Construction and Building Materials, Volume 472, 2025, 140793, ISSN 0950-0618, https://doi.org/10.1016/j.conbuildmat.2025.140793.
3. Matarneh S, Elghaish F, Edwards D J, Rahimian F P, Abdellatef E, Ejohwomu O (2024). Automatic crack classification on asphalt pavement surfaces using convolutional neural networks and transfer learning, ITcon Vol. 29, Special issue Managing the digital transformation of construction industry (CONVR 2023), pg. 1239-1256, https://doi.org/10.36680/j.itcon.2024.055
4. Sun, X., Yang, J., Huang, W. *et al.* Concrete crack classification based on fourier image enhancement and convolutional neural network. *Discov Civ Eng* **1**, 104 (2024). https://doi.org/10.1007/s44290-024-00107-6
5. Ha, J., Kim, D. & Kim, M. Assessing severity of road cracks using deep learning-based segmentation and detection. *J Supercomput* **78**, 17721–17735 (2022). https://doi.org/10.1007/s11227-022-04560-x
6. Hoang, N. and Nguyen, Q. (2023). Computer Vision-Based Recognition of Pavement Crack Patterns Using Light Gradient Boosting Machine, Deep Neural Network, and Convolutional Neural Network. *Journal of Soft Computing in Civil Engineering*, *7*(3), 21-51. https://doi.org/10.22115/scce.2023.367276.1547
7. Haiyan Zhuang, Yikai Cheng, Man Zhou, Zhenjun Yang, Deep learning for surface crack detection in civil engineering: A comprehensive review, Measurement, Volume 248, 2025, 116908, ISSN 0263-2241, https://doi.org/10.1016/j.measurement.2025.116908.
8. Ashraf, A., Sophian, A., & Bawono, A. A. (2024). Crack Detection, Classification, and Segmentation on Road Pavement Material Using Multi-Scale Feature Aggregation and Transformer-Based Attention Mechanisms. *Construction Materials*, *4*(4), 655-675. https://doi.org/10.3390/constrmater4040036
9. Chen, S., Feng, Z., Xiao, G., Chen, X., Gao, C., Zhao, M., & Yu, H. (2024). Pavement Crack Detection Based on the Improved Swin-Unet Model. *Buildings*, *14*(5), 1442. https://doi.org/10.3390/buildings14051442
10. Mayya, A. M., & Alkayem, N. F. (2024). Enhance the Concrete Crack Classification Based on a Novel Multi-Stage YOLOV10-ViT Framework. *Sensors*, *24*(24), 8095. https://doi.org/10.3390/s24248095
11. Sandra Matarneh, Faris Elghaish, Farzad Pour Rahimian, Essam Abdellatef, Sepehr Abrishami, Evaluation and optimisation of pre-trained CNN models for asphalt pavement crack detection and classification, Automation in Construction, Volume 160, 2024, 105297, ISSN 0926-5805, https://doi.org/10.1016/j.autcon.2024.105297.

12. Moreh, F., Lyu, H., Rizvi, Z.H. *et al.* Deep neural networks for crack detection inside structures. *Sci Rep* **14**, 4439 (2024). https://doi.org/10.1038/s41598-024-54494-y
13. Wang, W., Hu, W., Wang, W., Xu, X., Wang, M., Shi, Y., Qiu, S., & Tutumluer, E. (2021). Automated crack severity level detection and classification for ballastless track slab using deep convolutional neural network. Automation in Construction, 124, Article 103484, https://doi.org/10.1016/j.autcon.2020.103484 .
14. Alexandre Almeida Del Savio, Ana Luna Torres, Daniel Cárdenas-Salas, Mónica Vergara Olivera, Gianella Urday Ibarra, Dataset for training neural networks in concrete crack detection: laboratory-classified beam and column images, Data in Brief, Volume 61, 2025, 111643, ISSN 2352-3409, https://doi.org/10.1016/j.dib.2025.111643.
15. Ji, X., Joo, H.E. & Takahashi, Y. Quantitative evaluation of crack size effect on concrete deterioration induced by alkali-silica reaction: an experimental and numerical study. *Mater Struct* **58**, 100 (2025). https://doi.org/10.1617/s11527-025-02618-9 .
16. Anggara, D. W., Mohd Rahim, M. S., Zulkifli, R., Husain, A. R., Riyanto, Mazni, M., … Supangkat, S. H. (2024). Concrete Crack Detection, Orientation, and Measurement Using a Wall Climbing Robot. *Applications of Modelling and Simulation*, *8*, 272–282. Retrieved from http://arqiipubl.com/ojs/index.php/AMS_Journal/article/view/757.
17. Bo Yu, Jian Liang, Jiann-Wen Woody Ju, Classification method for crack modes in concrete by acoustic emission signals with semi-parametric clustering and support vector machine, Measurement, Volume 244, 2025, 116474, ISSN 0263-2241, https://doi.org/10.1016/j.measurement.2024.116474.
18. Arash Eshraghi, Serge Apedovi Kodjo, Patrice Rivard, Ghasem Shams, Machine-learning-based crack mode classification in plain concrete considering acoustic emission wave propagation effects, Construction and Building Materials, Volume 489, 2025, 142188, ISSN 0950-0618, https://doi.org/10.1016/j.conbuildmat.2025.142188.
19. Roy, S., Yogi, B., Majumdar, R. *et al.* Deep learning-based crack detection and prediction for structural health monitoring. *Discov Appl Sci* **7**, 674 (2025). https://doi.org/10.1007/s42452-025-07272-y