

Article

Engineering Structure Crack Detection Method Combining TAPFormer Model and Morphological Mask Reasoning Rules

Hao Peng ¹, Lintao Zhang ², Gang Li ^{1,*}, Yu Du ³ and Han Wu ⁴

¹ School of Architectural Engineering, Xinyang Vocational and Technical College, No. 48, 24th Street, Xinyang 464000, China; penghao@xyvtc.edu.cn

² School of Horticulture, Xinyang Agriculture and Forestry University, No. 1 North Ring Road, Xinyang 464000, China; zlt.179@163.com

³ School of Civil Engineering and Architecture, Wuhan University of Technology, Wuhan 430070, China; whutdy2023@163.com

⁴ School of Infrastructure Engineering, Nanchang University, No. 999 Xuefu Avenue, Nanchang 330047, China; wuhan20170620@163.com

* Correspondence: ligang@xyvtc.edu.cn

Abstract

To address challenges such as complex background interference, limited long-range modeling capabilities of CNNs, and poor generalization in steel-concrete cross-material scenarios, this study proposes an enhanced detection framework. This framework integrates a TAPFormer with morphological reasoning rules. The method utilizes TAPFormer as the backbone network. It captures global topological features of cracks through a Task-Aware Query mechanism. This approach compensates for the deficiencies of traditional convolutional operators in modeling the continuity of thin and long cracks. Furthermore, a mask reasoning module based on geometric priors is developed to handle unstructured interferences, such as marker pen marks, welds, and concrete holes. This module defines logical criteria, including edge curvature consistency, axial aspect ratios, and endpoint extension directions. These criteria are used to perform topological repair and filter false positives in the initial segmentation masks. A hybrid dataset containing 4500 cross-material damage images was used for validation. The results show that the proposed method achieves a mean IoU of 86.72% and an F1-score of 90.36%. Notably, the method filters over 91.0% of false positives caused by manual marker pen marks in interference-rich scenarios. Compared to mainstream state-of-the-art models, the IoU improves by at least 5.48%. The results show that the proposed framework improves the robustness and logical self-consistency of crack identification in complex engineering environments.

Keywords: engineering structures; crack detection; TAPFormer; mask reasoning; topological consistency



Academic Editor: Salvatore Verre

Received: 13 May 2026

Revised: 8 June 2026

Accepted: 13 June 2026

Published: 17 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The long-term safety and stability of engineering structures, such as bridges, tunnels, dams, and industrial buildings, are essential for modern infrastructure operations [1]. However, environmental erosion, overloading, material aging, and natural disasters inevitably lead to various forms of structural damage [2,3]. Cracks are among the most intuitive and critical early signs of structural degradation. In steel structures, cracks often appear as fine fissures that threaten fatigue life. In concrete structures, they indicate a decrease in load-bearing capacity or a loss of durability [4]. If these cracks are not identified and

repaired promptly, they may evolve into catastrophic structural failures. In the broader context of structural health monitoring, the integration of damage detection with multi-source monitoring data (e.g., temperature, displacement, and strain) is crucial for comprehensive safety assessment. Recent studies have demonstrated the effectiveness of data fusion approaches in bridge cable performance warning [5] and tower stability monitoring under strong wind conditions [6]. Understanding damage evolution and bearing-deformation behavior in complex structural systems [7] provides valuable context for interpreting crack detection results within overall structural performance evaluation. Therefore, developing efficient, accurate, and automated crack detection methods is important for Structural Health Monitoring (SHM) and maintenance decision-making.

Traditional crack detection relies primarily on manual inspections. However, manual methods are time-consuming and labor-intensive. Furthermore, the results are subjective. It is also difficult to achieve full coverage in large or complex structures, such as the interior of long-span steel box girders or high-rise building facades [8,9]. With the rapid development of computer vision and photoelectric sensing technologies, automated detection methods based on image processing have become a research focus. Early studies mostly used digital image processing techniques such as edge detection and threshold segmentation. However, these methods show poor robustness in engineering sites with uneven lighting, cluttered backgrounds, and complex material textures [10,11]. In recent years, deep learning algorithms, particularly convolutional neural networks (CNNs), have achieved significant success in damage recognition and semantic segmentation. CNNs leverage strong feature extraction capabilities to learn deep semantic information through multi-layer nonlinear transformations [12,13]. However, cracks in images exhibit topological characteristics such as being thin, winding, and occupying a very low proportion of pixels. Quantitative studies show that CNN models have inherent deficiencies in capturing long-range spatial dependencies due to the limitations of the local receptive field of convolutional kernels. When processing low-contrast or complex-shaped cracks, traditional CNNs (such as U-Net and DeepLabV3+) often exhibit crack fracture rates exceeding 30%. This makes it difficult to maintain the topological continuity of the damage morphology [14–16].

To address these limitations, the Transformer architecture provides a new perspective for feature extraction of complex cracks [17–20]. This is due to its self-attention mechanism, which establishes long-range dependencies between global pixels. For example, Alshawabkeh et al. [21] proposed the MaskerTransformer hybrid model. This model integrates the precise pixel positioning of Mask R-CNN with the global context awareness of Vision Transformer to detect pavement cracks. Honarjoo et al. [22] proposed a Transformer model based on a dynamic attention mechanism. It detects crack images by integrating classification, object detection, and semantic segmentation tasks. The Task-Aware Part-based Transformer (TAPFormer) introduces part-based feature awareness and task-guided mechanisms. It can adaptively allocate attention based on cracks of different scales and shapes. This significantly improves the capture of thin crack features [23]. However, even with powerful feature extraction performance, the “cross-material” detection environment in actual engineering remains challenging. In steel structure detection, high specular reflections, interlaced welds, manual marker pen marks, and calibration scales are often highly confused with extremely fine fatigue cracks [24]. In concrete scenarios, water seepage, moss, construction joints, and rough material textures constitute complex unstructured interference under diffuse reflection environments [25,26]. These drastic fluctuations in the feature space caused by material properties make it difficult for purely data-driven models to distinguish real cracks from “logically” non-crack interferences at the pixel level. This leads to numerous false positive areas and a lack of logical self-consistency in detection results [8,27].

To solve this problem, reasoning models combined with domain prior knowledge are considered an effective supplement. The expansion of cracks follows specific topological laws in geometric morphology, such as linear continuity and edge curvature distribution. Based on this, this study proposes an automated crack detection method for cross-material engineering structures by combining the TAPFormer model with mask reasoning rules. The method first uses TAPFormer as the core feature extraction backbone to obtain initial probability maps for steel and concrete surfaces. Subsequently, a mask reasoning module adapted to engineering site characteristics is constructed. This module performs logical validation of the observed areas using geometric priors. An information feedback mechanism combines the reasoning results with the initial segmentation masks. This achieves a closed loop from pixel prediction to topological logic validation, effectively filtering noise such as welds and marker pen marks. The core contributions of this paper are summarized as follows:

(1) An architectural adaptation of TAPFormer is implemented to preserve the topological integrity of cracks in heterogeneous backgrounds. By integrating an overlapping patch embedding layer, the framework maintains sub-pixel continuity at patch boundaries, effectively preventing the crack fragmentation common in standard Transformer-based methods. Furthermore, the task-aware querying mechanism is refactored into a topological sensing logic that focuses on geometric key-points (e.g., crack tips and bifurcations). This allows for precise feature alignment and mitigates response diffusion, enabling the model to distinguish slender crack signals from highly confusing steel welds and concrete textures.

(2) A logical reasoning module based on geometric and topological priors is proposed. Unlike traditional post-processing methods that rely on simple area or aspect ratio filtering, this module constructs a perception-reasoning closed-loop. It defines criteria such as edge curvature variance, aspect ratio, and directional extension to achieve topological self-consistency. This mechanism transforms physical priors into a differentiable-like logical layer that is nonlinearly fused with visual probabilities, significantly improving the robustness of detection results in complex backgrounds.

The remainder of this paper is organized as follows: Section 2 reviews related work on crack detection and Transformer architectures. Section 3 details the proposed method combining TAPFormer and mask reasoning rules, covering the model architecture, reasoning rules, and the fusion feedback mechanism. Section 4 introduces the experimental setup, dataset construction, and provides an in-depth analysis of detection results and ablation studies. Section 5 concludes the research and presents future work.

2. Related Work

2.1. Deep Learning-Based Crack Detection

With the advancement of computing power, deep learning has made significant progress in engineering structural damage identification. Early research primarily leveraged the strong feature representation capabilities of CNNs for damage classification. For example, Cha et al. [28] implemented automated bridge crack identification using a sliding window strategy. This demonstrated the robustness of deep learning in complex lighting environments. To obtain pixel-level damage information, semantic segmentation models have gradually become mainstream. The Fully Convolutional Network (FCN) proposed by Long et al. [29] laid the foundation for refined segmentation. Subsequently, researchers improved U-Net [30], SegNet [31], and the DeepLab series [32] to account for the thin and multi-scale characteristics of cracks. Park et al. [33] combined the YOLO architecture with multi-scale feature fusion to enhance the sensitivity to minor damage in complex backgrounds. Xu et al. [34] improved the extraction accuracy of fine fatigue cracks on steel box girders by fusing deep convolutional features.

However, despite the excellent local feature extraction of CNNs, their performance is limited by the inherent local receptive field of convolutional kernels. This creates obvious bottlenecks when processing continuous cracks that extend over long distances in noisy backgrounds. Jing et al. [35] noted that even with optimized U-Net models, the fracture rate for cross-scale long cracks remains above 30%. This topological discontinuity severely weakens the reliability of subsequent quantitative damage assessments. Due to the lack of global context modeling, detection results often exhibit severe fracture phenomena. For instance, in scenarios with interference from marker pen marks or dark brown water seepage, CNNs often fail to establish global topological correlations between pixels. This leads to the mis-segmentation of long continuous cracks into isolated “fragmented” pixel clusters [36].

2.2. Applications of Transformers in Computer Vision and Structural Damage Identification

To overcome the inherent bottleneck of CNNs in capturing long-range dependencies, the Transformer architecture, which originated in natural language processing, has been introduced to computer vision tasks via the self-attention mechanism. The Vision Transformer (ViT) proposed by Dosovitskiy et al. [37] demonstrated the potential of pure attention mechanisms in image recognition. Subsequently, the Swin Transformer [38] significantly improved computational efficiency while maintaining global modeling capabilities through the shifted windows mechanism. In the field of structural health monitoring, Transformers show advantages in modeling the complex topological features of cracks. For example, TransUNet [39] improved the connectivity preservation of cracks by combining the local extraction capabilities of CNNs with the semantic representation of Transformers.

To meet the more challenging fine-grained detection requirements at engineering sites, this study adapts the task-aware querying mechanism originally proposed for point tracking in event cameras. While the original TAPFormer targets temporal correspondence in video sequences, we reformulate the “task-aware part” concept for spatial crack segmentation by defining “parts” as crack-specific geometric keypoints (e.g., tips, bifurcations, high-curvature points). The adapted architecture preserves the core cross-attention mechanism between learnable queries and hierarchical features but replaces the temporal tracking head with a semantic segmentation decoder. The task-aware querying mechanism enables feature alignment by learning local geometric patterns. In crack detection, this mechanism naturally captures critical morphological components such as endpoints and bifurcations, which exhibit strong topological correlations along the crack trajectory. By reconstructing the original task-aware mechanism into a spatial topological sensing logic, this model assists the system in accurately extracting the semantic representation of cracks from highly confusing backgrounds, such as welds, rust, and manual handwriting. This effectively mitigates the problem of feature response diffusion in cross-material scenarios [40].

2.3. Image Processing Methods Based on Logical Reasoning and Prior Knowledge

Despite the significant advantages of deep learning in feature representation, the lack of semantic logic caused by its “black-box” nature limits model reliability in complex engineering environments. During practical inspections, unstructured interferences—such as regular welds on steel structures, irregular moss on concrete surfaces, and manual handwriting—often cause visual ambiguity. This leads to high false positive rates in purely data-driven models. To improve the interpretability and logical self-consistency of detection results, researchers have integrated physical constraints and morphological rules into detection algorithms. Yuan et al. [41] introduced physical constraint terms into the loss function to guide models in generating more physically meaningful damage morphologies. Jing et al. [35] utilized the geometric characteristics of cracks to perform logical

validation on initial masks. This achieved the automated removal of interferences such as isolated spots.

Building on these studies, mask reasoning methods have further explored the linear continuity and spatial topological correlation of cracks [42]. However, current research still faces challenges regarding the insufficient coupling between deep semantic expression and low-level logical constraints. Specifically, establishing an efficient feedback loop between high-level semantic feature extractors (e.g., TAPFormer) and morphological reasoning rules remains a core difficulty in aligning perception with reasoning. Especially when processing heterogeneous datasets of steel-concrete structures, achieving a dynamic balance between visual perception sensitivity and morphological reasoning robustness is essential. This balance is key to improving the generalization performance of automated structural inspection systems.

3. Methodology

3.1. Overall Architecture Design

The proposed crack detection framework for cross-material engineering structures aims to establish a perception-reasoning closed-loop system that deeply integrates high-level feature extraction with high-order logical constraints. This architecture departs from the limitations of traditional end-to-end models that rely solely on pixel-driven data and treat post-processing as a disconnected filtering stage. Instead, it constructs a dual-driven path consisting of “semantic probability mapping based on TAPFormer” and “topological logic correction based on mask reasoning rules”. The overall architecture design is illustrated in Figure 1.

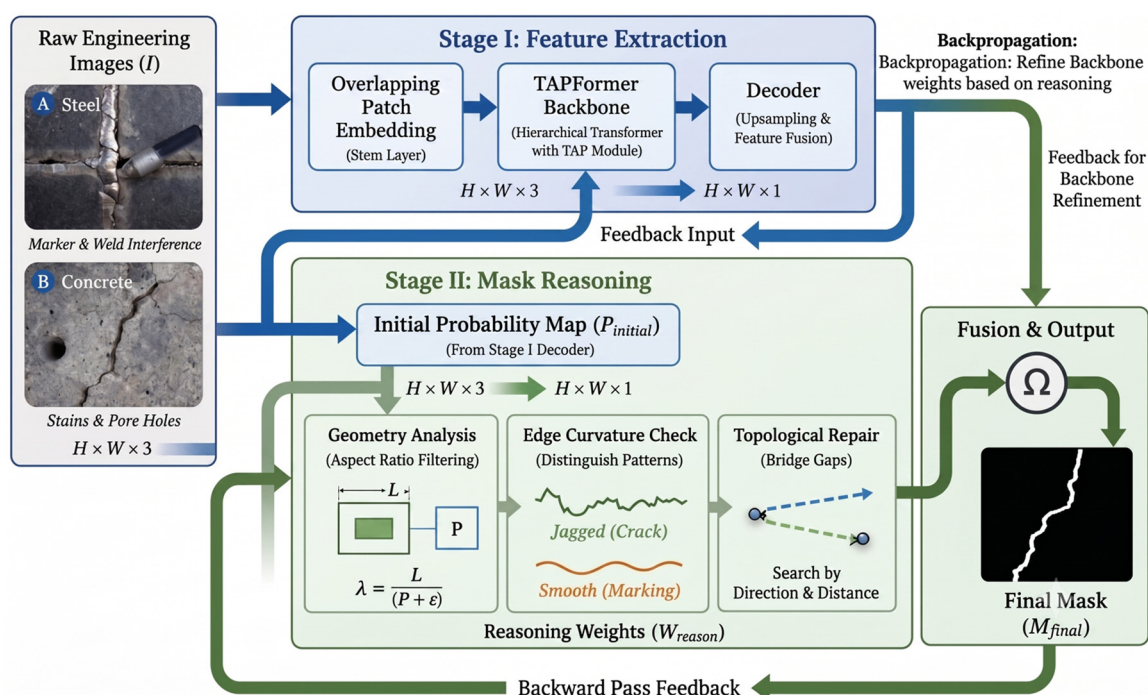


Figure 1. The flowchart of crack detection.

In the execution process, the system first performs preprocessing and standardization on the input raw engineering image $I \in \mathbb{R}^{W \times H \times 3}$. It then feeds the image into the TAPFormer feature extraction engine. Through hierarchical encoding and task-aware part mechanisms, this engine captures highly discriminative crack geometric features within complex backgrounds, such as the metallic luster of steel structures and the mottled textures

of concrete. The engine outputs an initial probability map P_{initial} with the same resolution as the original image. To further eliminate engineering-specific false positive interferences (e.g., regular weld edges, manual marker pen marks, and concrete construction joints), the system introduces a mask reasoning module. Crucially, this module is not a simple heuristic filter; rather, it acts as a high-level logical layer designed to enforce topological self-consistency. It utilizes geometric priors to perform logical validation of candidate regions. Here, we define ‘topological self-consistency’ as the adherence of detected crack morphology to fundamental physical constraints: linear continuity, directional coherence, and aspect ratio conformity. By evaluating these criteria for each observed segment, the module generates corrective weights to dynamically “enhance” or “suppress” the probability distribution. Finally, through an information feedback and fusion operator based on nonlinear weighting, the probability distribution from low-level visual features is non-linearly coupled with the correction weights generated from high-level reasoning. This results in the final refined detection mask M_{final} . This design ensures high sensitivity to minute cracks while significantly improving the robustness of detection results by ensuring they adhere to the intrinsic structural logic of cracks rather than just pixel-level patterns.

3.2. Crack Feature Extraction Model Based on TAPFormer

As the core feature extraction engine of this method, the TAPFormer is designed to solve the challenges of sparse crack features and high susceptibility to interference in complex engineering environments. In terms of implementation details, the model first employs an overlapping patch embedding layer, as shown in Figure 2. Unlike the non-overlapping slicing method used in standard Transformers, this study sets a stride s smaller than the kernel diameter k of the sliding window, creating overlapping areas between adjacent patches. Since cracks are typically thin, continuous, and randomly span across patch boundaries, traditional non-overlapping slicing can lead to topological discontinuities in feature responses at boundaries. This overlapping patch design effectively preserves sub-pixel continuity at patch boundaries, preventing the fracture of thin cracks at the bottom data level. The architectural configuration is as follows: The overlapping patch embedding layer employs a 7×7 convolutional kernel with stride 4, generating feature maps of 128×128 resolution. The hierarchical encoder consists of four stages with output resolutions of 128×128 , 64×64 , 32×32 , and 16×16 , respectively. In the TAP module, we set the number of learnable task-aware queries $N_q = 16$, query dimension $d_q = 256$, and multi-head attention with $h = 8$ heads. Each TAP block performs cross-attention between queries Q_{tap} and hierarchical features, followed by feed-forward networks with GELU activation. Subsequently, the model enters a hierarchical feature encoding stage. Through four stages of progressively reduced resolution, the model uses multi-head self-attention to model the spatial dependencies of cracks at different scales. In each encoding block, the model calculates global contextual relationships, allowing weak crack signals—which are difficult to identify from a local perspective—to be captured through their continuous spatial trajectory.

To further address non-crack interferences in steel and concrete backgrounds, TAPFormer introduces a Task-Aware Part (TAP) module. This module adaptively allocates attention weights rather than uniformly distributing them across all pixels. By introducing a set of learnable task-aware part queries (Q_{part}), the module actively searches for and anchors “critical parts” of the image that match the topological features of cracks, such as crack tips, bifurcation points, and linear segments with specific contrast. The attention calculation formula is defined as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V \quad (1)$$

where $Q = [Q_{\text{global}}; Q_{\text{part}}]$. Through feature alignment of critical local geometric parts via Q_{part} , the model achieves more precise localization. For steel structure images, the TAP module learns micro-differences in edge roughness and linear trajectories between fatigue cracks and regular welds, guiding the attention flow to actual damage areas. For concrete structures, the module identifies the geometric aspect ratio of targets to automatically ignore spot-like water stains or holes, reinforcing only the linearly extending areas. In the decoding stage, the model adopts a symmetric skip-connection architecture to fuse high-level semantic features with low-level high-resolution details. Through up-sampling operators, the model maps the feature vectors reinforced by TAP back to the original pixel space. Finally, via a linear projection layer and a Softmax activation function, the model calculates the confidence score of each pixel belonging to a crack, outputting the initial probability map P_{initial} .

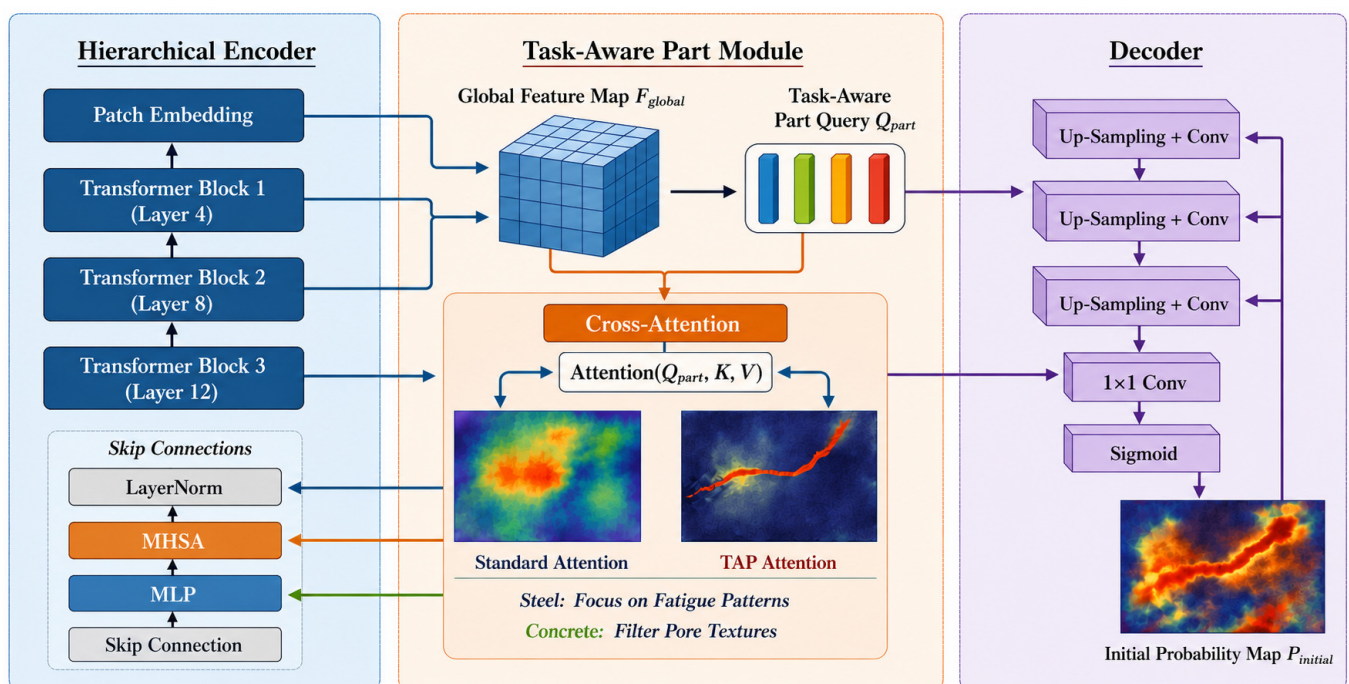


Figure 2. Crack feature extraction model structure diagram based on TAPFormer.

To ensure effective handling of the extreme pixel imbalance problem (where crack pixels are far fewer than background pixels) during training, this study designs a total loss function L_{total} combining Binary Cross Entropy (BCE) and Dice Loss. The BCE loss ensures pixel-level classification accuracy, while the Dice Loss enhances the model's ability to capture thin and sparse structures by evaluating the overlap between the predicted mask and the ground truth. The total loss function is defined as:

$$L_{\text{total}} = \lambda_{\text{bce}} L_{\text{BCE}} + \lambda_{\text{dice}} L_{\text{Dice}} \quad (2)$$

where

$$L_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (3)$$

$$L_{\text{Dice}} = 1 - \frac{2 \sum_{i=1}^N y_i p_i + \epsilon}{\sum_{i=1}^N y_i + \sum_{i=1}^N p_i + \epsilon} \quad (4)$$

In these equations, y_i is the ground truth pixel label, p_i is the predicted probability value, N is the total number of pixels, and ϵ is a smoothing factor (set to 1×10^{-5}) to prevent division by zero. The hyperparameters λ_{bce} and λ_{dice} are both set to 0.5. Compared

to Focal Loss, which is prone to gradient fluctuations and overfitting in regions with strong interference (e.g., specular reflections and handwriting), or Tversky Loss, which is highly sensitive to hyperparameter settings for fine crack widths, the combination of BCE and Dice maintains training stability while supporting high Intersection over Union (IoU) performance in complex backgrounds.

3.3. Morphological Mask Reasoning Rules

Despite the excellent performance of the TAPFormer model in global feature capture, purely pixel-probability-based deep learning models often struggle to distinguish real structural cracks from interferences with similar visual features in complex engineering environments due to a lack of semantic logic. For example, manual marker pen annotations and metal welds in steel structures, as well as construction joints, cables, or dark stains in concrete structures, frequently lead to false positive outputs. To address this, this study designs a morphological mask reasoning module. Unlike conventional connected-component filtering or morphological closing, this module operates as a feedback-guided layer that modulates the probability distribution. It can recover fractured crack segments through directional extension reasoning based on endpoint alignment and deflection angle constraints. This module performs logical validation and geometric topological repair on the initial probability map P_{initial} using domain-specific prior knowledge. The module first performs dynamic threshold segmentation on P_{initial} to extract a series of candidate connected regions $C = \{c_1, c_2, \dots, c_n\}$, and performs multi-dimensional morphological feature modeling for each region.

For interferences from manual handwriting and calibration rulers in steel structure scenarios, the module introduces edge smoothness and width constancy criteria. Real crack edges typically exhibit random micro-sawtooth shapes, and their edge curvature κ fluctuates significantly. In contrast, marker pen marks or calibration ruler edges possess high geometric smoothness. In this study, edge curvature variance Ψ is calculated as a quantitative indicator:

$$\Psi = \frac{1}{N_e} \sum_{i=1}^{N_e} (\kappa_i - \bar{\kappa})^2 \quad (5)$$

where N_e is the total number of edge pixels and κ_i is the local curvature. If Ψ is lower than a preset threshold Ψ_{th} , the region is identified as artificial interference and suppressed. In this study, threshold values were empirically determined through validation set optimization: edge curvature variance threshold $\Psi_{th} = 0.15$ for steel structures ($\Psi_{th} = 0.20$ for concrete due to higher surface roughness), aspect ratio threshold $AR_{\min} = 8.0$ for steel ($AR_{\min} = 6.0$ for concrete), endpoint distance threshold $d_{th} = 20$ pixels, deflection angle threshold $\theta_{th} = 30^\circ$, and search radius $R = 20$ pixels. The material-specific parameter sets are defined as $\varphi_{\text{steel}} = \{\Psi_{th} = 0.15, AR_{\min} = 8.0, d_{th} = 20, \theta_{th} = 30^\circ\}$ and $\varphi_{\text{concrete}} = \{\Psi_{th} = 0.20, AR_{\min} = 6.0, d_{th} = 20, \theta_{th} = 30^\circ\}$. Additionally, the width of real cracks evolves nonlinearly along the path, while the width of interferences fluctuates minimally within local scales. By calculating geometric moments and edge curvature distributions for each connected domain, the reasoning rules automatically identify and weaken these regions with high artificial regularity. For concrete structures, interferences mostly appear as isolated hole noise or linearly distributed moss. The module uses the aspect ratio $\lambda = L^2 / A$ for screening (where L is the skeleton length and A is the area). Real cracks exhibit significant thin and linear topological characteristics, whereas the λ values of holes or spot-like stains are far below the preset physical threshold. By setting global morphological filters, the system effectively removes these non-crack components from candidate masks.

Regarding topological repair, this study constructs an automated reasoning mechanism based on directional extension laws to address detection fractures caused by lighting occlusion or material changes. For initially identified fractured crack segments, the module uses their endpoints p_i as centers to detect potential logical connection points p_j within a specific search radius R . The reasoning logic follows the principles of displacement minimization and deflection angle minimization, satisfying the following connection criteria:

$$\begin{cases} d(p_i, p_j) = \|p_i - p_j\|_2 \leq d_{th} \\ \Delta\theta = \arccos\left(\frac{v_i \cdot v_j}{\|v_i\| \|v_j\|}\right) \leq \theta_{th} \end{cases} \quad (6)$$

where v_i and v_j are the tangent direction vectors at the endpoints. If the constraints of distance threshold d_{th} and angle threshold θ_{th} are met, the two regions are determined to belong to the same crack in physical logic, and linear interpolation repair is executed in the mask layer. This enhancement mechanism compensates for the lack of observation in extremely low-contrast areas of the visual model, ensuring the overall continuity of the detection results.

3.4. Information Feedback and Integration Strategy

Upon completing visual feature extraction via TAPFormer and logical validation via mask reasoning, a dual-path information feedback and fusion mechanism is constructed. Unlike traditional crack detection methods that rely on serial post-processing (e.g., simple area or aspect ratio filtering), this mechanism treats the reasoning module as a feedback-guided logical layer. It aims to achieve deep coupling between low-level observation data and high-level prior logic, establishing a perception-reasoning closed loop based on topological self-consistency. The reasoning module operates as a non-differentiable post-processing layer applied after TAPFormer inference. Unlike end-to-end differentiable architectures, this module does not participate in gradient backpropagation during training. Instead, it acts as a logical constraint filter that refines the probability distribution $P_{initial}(x, y)$ based on geometric priors, generating corrective weights through rule-based evaluation. The fusion operator combines these two information streams in a feed-forward manner during inference. The feedback and integration strategy is shown in Figure 3.

The core objective of the fusion process is to perform pixel-level dynamic adjustment of the initial probability map $P_{initial}(x, y)$ based on the logical weights $W_{reason}(x, y)$. This weight field is constructed from high-order criteria, including edge curvature consistency, axial extension direction, and topological connection rules, rather than a mere stack of filters. The final detection probability distribution P_{final} can be calculated through the following nonlinear fusion operator:

$$P_{final}(x, y) = \text{Sigmoid}\left(\alpha \cdot P_{initial}(x, y) + \beta \cdot f_{logic}(W_{reason}, \phi)\right) \quad (7)$$

where f_{logic} denotes the mapping function that transforms geometric priors into logical responses, and ϕ is the parameter set defined for different materials. α and β are the weight coefficients for the visual feature term and the reasoning feedback term, respectively. Through cross-validation, the optimal weight allocation in this study is determined as $\alpha = 0.6$, $\beta = 0.4$.

During the feedback adjustment process, the fusion strategy follows the “logic priority” principle, enabling the system to self-correct visual ambiguities. When the mask reasoning module identifies high-smoothness edges ($\Psi < \Psi_{th}$) or regions with extremely low aspect ratios, W_{reason} exhibits significant negative suppression characteristics. This forces a reduction in the prediction confidence of the visual model by lowering the logits of the region. Crucially, for fractured areas that satisfy the directional extension and topologi-

cal connection criteria (Equation (6)), W_{reason} acts as a positive enhancement weight. This guides the model to “logically” retrieve missing crack pixels in areas where visual features are weak or occluded—a capability that simple post-processing filters do not possess. This feedback mechanism effectively transforms engineering experience into calculable mathematical constraints. Finally, the system executes global adaptive threshold segmentation on P_{final} to generate the refined segmentation mask M_{final} :

$$M_{\text{final}}(x, y) = \begin{cases} 1, & P_{\text{final}}(x, y) > \tau \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

where the discrimination threshold τ is set to 0.5 based on validation set performance. Through this fusion process, the proposed method not only preserves the powerful capability of TAPFormer to capture complex thin features but also achieves topological self-consistency, ensuring the detection results adhere to both visual evidence and physical logic.

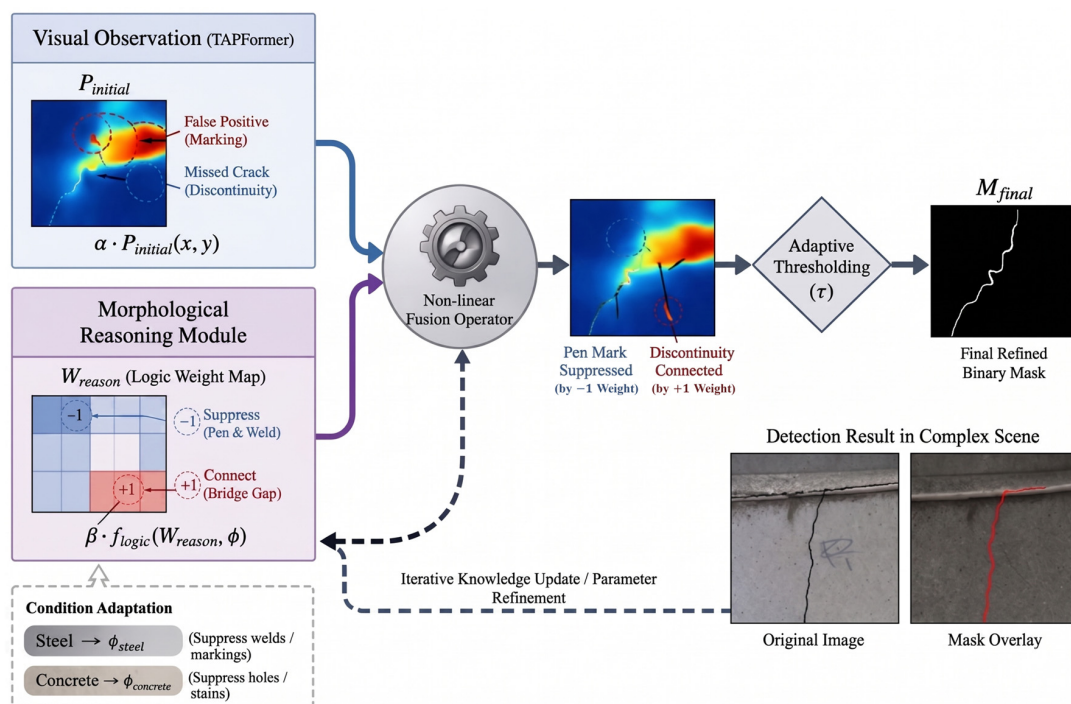


Figure 3. The information feedback and integration strategy.

4. Experiments and Analysis

4.1. Dataset Preparation and Preprocessing

To verify the robustness of the proposed method in heterogeneous engineering scenarios, this study constructed a benchmark dataset for cross-material crack detection. This dataset consists of raw images collected from real inspections and contains no synthetic samples. It covers typical interferences for both steel and concrete structures.

For the steel structure subset, samples were primarily sourced from periodic inspections of a long-span steel box girder bridge, comprising 120 high-resolution images with a resolution of 4928×3264 [34]. The detection challenges in this subset lie in the high similarity between environmental interference and structural features: (1) Manual marking interference: Marker pen marks and calibration rulers are common at inspection sites. Their high-contrast edges and linear geometric characteristics frequently trigger false positive reports. (2) Structural and environmental noise: Weld textures, pitted corrosion, and non-

uniform specular reflections on metal surfaces often blur crack edges. This increases the difficulty for the model to distinguish structural features from damage.

For the concrete structure subset, this study integrated selected samples from the SDNET2018 public dataset and field images from water conservancy dams and existing building facades [43]. The complexity of this subset arises from the significant overlap between background noise and crack features in terms of grayscale and pixel distribution. Honeycomb surfaces, water stains, moss, and pre-existing construction joints exhibit crack-like characteristics. Furthermore, crack morphology is highly uncertain due to the random influence of base material textures. This imposes higher requirements for feature extraction and logical judgment.

All images were cropped into 512×512 pixels patches before training using a sliding window approach with stride = 256 to ensure adequate overlap. For the steel structure subset, 120 original high-resolution images (4928×3264) yielded approximately 1800 patches. For the concrete subset, 280 original images from SDNET2018 and field inspections generated approximately 2700 patches, totaling 4500 image pairs. To prevent data leakage, all patches derived from the same original image were assigned to the same data split (training, validation, or test). The dataset was divided as follows: 70% for training (3150 patches), 15% for validation (675 patches), and 15% for an independent test set (675 patches). After augmentation, the effective training sample size was expanded to 15,750 samples. All images were annotated at the pixel level using LabelMe [44] by two senior structural health monitoring technicians with >5 years of inspection experience. To ensure annotation reliability, 200 randomly selected samples were independently annotated by both technicians. Inter-annotator agreement was quantified using Cohen's κ coefficient, yielding $\kappa = 0.89$ (substantial agreement) when allowing 5-pixel tolerance for crack boundary variations. The annotations were then cross-verified and consolidated through consensus discussion to ensure topological consistency in the ground truth. The dataset preparation process is shown in Figure 4.

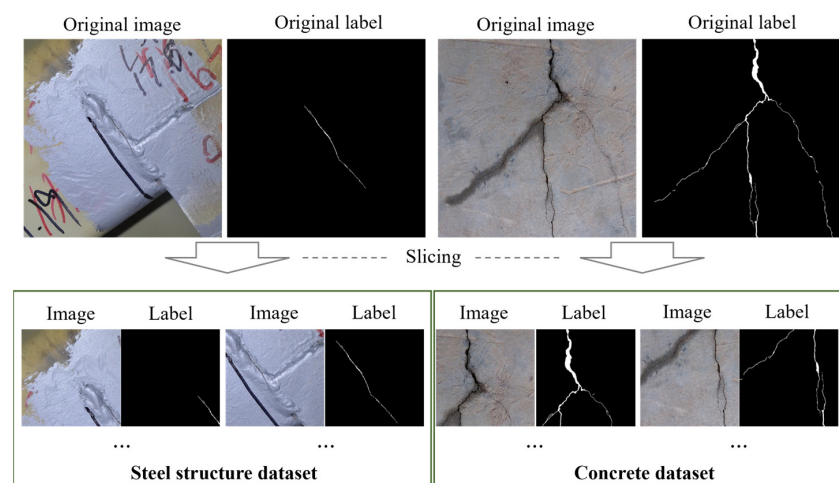


Figure 4. The process of creating dataset.

To mitigate potential overfitting and improve geometric invariance, a refined data augmentation strategy was implemented:

(1) Geometric Transformation: Including random angle rotations (0° , 90° , 180° , 270°) and horizontal/vertical flips to enhance the model's direction-independence regarding crack spatial trajectories.

(2) Photometric Distortion: Random adjustment of contrast (0.5~1.5), brightness (0.7~1.3), and saturation to simulate dynamic changes in lighting conditions at engineering sites.

(3) Noise Suppression Test: Injection of Gaussian white noise with different variances to simulate sensor noise from industrial cameras during low-light or long-distance shooting. The effective training sample size after augmentation was expanded to five times the original scale and divided into training and validation sets according to the 8:2 ratio.

4.2. Experimental Setup and Evaluation Metrics

All training and inference tasks in this study were performed on a deep learning workstation equipped with an NVIDIA GeForce RTX 3090 (24 GB VRAM), running on the Windows operating system. The core algorithm was implemented based on the PyTorch 2.0 framework with CUDA version 12.1.

To ensure stability and efficient convergence during training, the hyperparameters were strictly tuned. The AdamW optimizer was selected for its superior robustness in handling weight updates for Transformer architectures. The initial learning rate was set to 1×10^{-4} with a Cosine Annealing learning rate scheduling strategy. The complete experimental parameter configuration is shown in Table 1.

Table 1. Model training hyperparameter configurations.

Category	Parameter	Value
Software Framework	Deep Learning Library/CUDA	PyTorch 2.0/CUDA 12.1
Input Specifications	Image Resolution	512×512 pixels
Optimization Strategy	Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$)
	Base Learning Rate	1×10^{-4}
	Weight Decay	0.05
	Learning Rate Scheduler	Cosine Annealing
	Loss Function	Binary Cross Entropy + Dice Loss
Iteration Parameters	Batch Size	16
	Training Epochs	200

Regarding the evaluation system, this study focuses not only on detection accuracy but also on the algorithm's recall capability in complex backgrounds. Accordingly, IoU, Precision, Recall, and F1-score were selected as core metrics. The F1-score, as the harmonic mean of Precision and Recall, provides a comprehensive evaluation of model performance in the presence of significant negative sample interference (e.g., pen marks, weld seams). The calculation formulas for each metric are as follows:

$$\text{IoU} = \frac{TP}{TP + FP + FN} \quad (9)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (10)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (11)$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

In these formulas, TP represents the number of correctly identified crack pixels, FP represents the number of background or interference pixels incorrectly identified as cracks, and FN represents the number of missed crack pixels. Through multi-dimensional quantitative analysis, this study aims to reveal the actual contribution of TAPFormer combined with morphological reasoning rules in improving crack detection accuracy.

4.3. Comparative Analysis of Experimental Results

Model training was conducted using an end-to-end supervised learning approach, with the training process illustrated in Figure 5. In the early stages of training, benefiting from the synergistic effect of overlapping patch embedding and the TAP module, the model quickly captured low-level semantic features of cracks, resulting in a rapid downward trend in the loss function. As the number of iterations increased, the joint loss function began to guide the model to focus on the precise segmentation of fine crack edges and the logical suppression of background false positive noise. Consequently, the rate of loss reduction gradually slowed down, finally entering a stable oscillation phase around epoch 50.

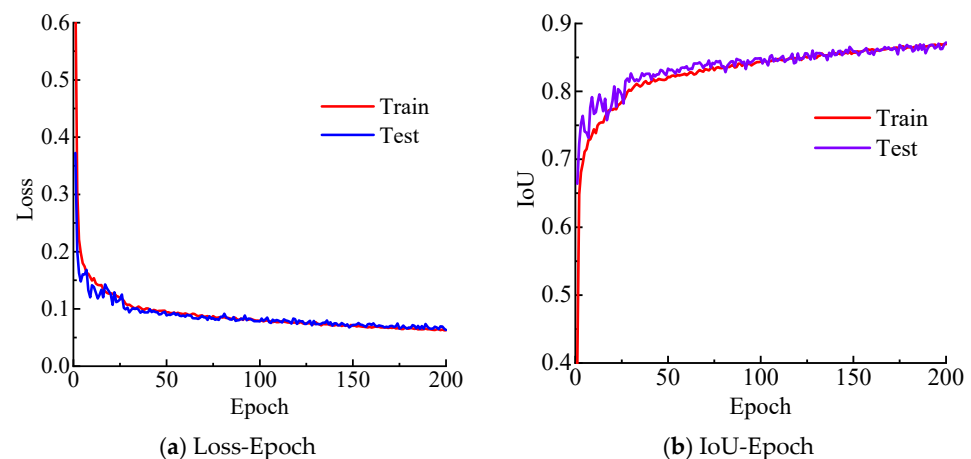


Figure 5. The monitored IoU and loss curve results.

Simultaneously, the IoU curve on the validation set exhibited excellent convergence characteristics. During the first 50 epochs, the IoU values rose rapidly, indicating that the model had initially grasped the geometric trends of cross-material cracks. Between Epoch 50 and 100, the IoU curve tended to smooth out as morphological reasoning logic was integrated and features were refined. Finally, at approximately Epoch 180, the model reached optimal performance and remained stable. The trends of the performance curves for the training and validation sets remained highly synchronized, proving that the model possesses strong generalization capability without significant overfitting.

To verify the superiority of the proposed method in cross-material scenarios, this study compared it with several representative models. These include classic convolutional neural networks U-Net [45] and DeepLabV3+ [46], the Transformer-based Swin-Unet [47], SegFormer (B5) [48] (a powerful general segmentation Transformer) and CrackFormer II [49] (a specific SOTA model for crack detection). Table 2 summarizes the performance metrics on the cross-material test set.

Table 2. Comprehensive performance metrics and computational cost on the test set (mean $\pm$ std over 3 runs).

Methods	IoU (%)	Precision (%)	Recall (%)	F1-Score (%)	Inference Time (ms)
U-Net	75.82 $\pm$ 0.63	81.56 $\pm$ 0.71	77.21 $\pm$ 0.58	79.33 $\pm$ 0.52	28 $\pm$ 2
DeepLabV3+	79.14 $\pm$ 0.55	83.22 $\pm$ 0.62	81.45 $\pm$ 0.49	82.32 $\pm$ 0.47	32 $\pm$ 2
Swin-Unet	81.05 $\pm$ 0.48	84.90 $\pm$ 0.53	83.12 $\pm$ 0.44	84.00 $\pm$ 0.41	41 $\pm$ 3
SegFormer (B5)	82.86 $\pm$ 0.44	86.15 $\pm$ 0.49	84.40 $\pm$ 0.42	85.27 $\pm$ 0.39	35 $\pm$ 2
CrackFormer II	84.12 $\pm$ 0.41	87.54 $\pm$ 0.46	86.22 $\pm$ 0.39	86.87 $\pm$ 0.37	38 $\pm$ 3
Proposed method	86.72 $\pm$ 0.38 *	91.45 $\pm$ 0.43 *	89.30 $\pm$ 0.40 *	90.36 $\pm$ 0.36 *	45 $\pm$ 3

Notes: * $p < 0.01$ (paired t -test vs. all baselines).

Quantitative analysis of the metrics in Table 2 reveals a clear trend in the technological evolution of crack detection. From the early CNN-based U-Net (75.82% IoU) to the recent Transformer-based SOTA models such as SegFormer and CrackFormer II (82.86% and 84.12% IoU, respectively), the performance gain primarily stems from the expansion of receptive fields and enhanced feature representation. However, the incremental improvement between SegFormer and CrackFormer II (only 1.26% in IoU) suggests that pure data-driven architectures are approaching a performance ceiling in complex engineering environments. The proposed method improves upon this baseline, achieving an IoU of 86.72%, representing a 2.6% jump over the strongest baseline (CrackFormer II). A critical differentiator lies in the Precision metric; while CrackFormer II achieves 87.54%, our method reaches 91.45%. This 3.91% improvement reflects better noise-handling capability. Most pure Transformer models struggle to distinguish between “visually similar” and “logically dissimilar” features, often mis-segmenting regular weld edges or sharp shadow boundaries as cracks. By introducing the mask reasoning layer, our framework filters out these high-confidence visual errors through topological self-consistency checks. Furthermore, the Recall of the proposed method (89.30%) leads the second-best model by over 3%. This indicates that the “Positive Recovery” mechanism in our logical feedback loop (Equation (7)) successfully retrieves crack segments that were visually attenuated by low contrast or occlusions—areas where even the task-aware mechanisms of CrackFormer II failed to maintain connectivity. The performance gap between the Proposed Method and mainstream models becomes more pronounced as the complexity of the background increases, particularly in cross-material scenarios where feature distributions are non-uniform. All experiments were repeated three times with different random seeds (42, 123, 456). Paired *t*-tests confirm that the proposed method’s IoU improvement over all baselines is statistically significant ($p < 0.01$). The additional 7–10 ms inference time compared to CrackFormer II results from the post-inference reasoning module, which is acceptable for non-real-time inspection scenarios.

The qualitative results in Figure 6 provide further evidence of why logical constraints are necessary for structural health monitoring. In the steel plate scenarios, the pen marks possess a linear intensity distribution nearly identical to real fatigue cracks. Consequently, even the most advanced pure vision models (SegFormer and CrackFormer II) produce “ghost” detections or merged segments. Our method, by quantifying edge curvature variance (Ψ), recognizes the artificial smoothness of the pen marks and effectively decouples them from the underlying crack morphology. In concrete scenarios characterized by shadowed “fracture zones,” the proposed method demonstrates superior topological self-consistency. While standard models produce fragmented pixel clusters that fail to represent the actual crack length, our directional extension reasoning (Equation (6)) bridges these gaps logically. This ensures that the output mask is not just a collection of classified pixels, but a structurally valid representation that supports reliable subsequent engineering assessments, such as crack width quantification and safety grading.

Finally, Table 3 further quantifies the degree of error improvement before and after the intervention of the reasoning module for the key engineering field interferences addressed in this study. Experimental results prove that for highly deceptive marker pen marks, morphological reasoning rules reduced the false positive ratio from 12.45% to 1.12%, an improvement of 91.0%. For complex steel weld seams and concrete pores, the false positive suppression rates reached 75.4% and 86.9%, respectively. This indicates that purely data-driven models face inherent bottlenecks when dealing with non-natural interferences (such as human intervention or material defects), whereas transforming engineering logic (such as linear continuity and geometric aspect ratio) into mask processing rules is a necessary path to improve the reliability of visual detection systems. Overall, the proposed method

not only reaches new heights in detection accuracy but its logical consistency in complex cross-material environments endows it with high practical engineering value.

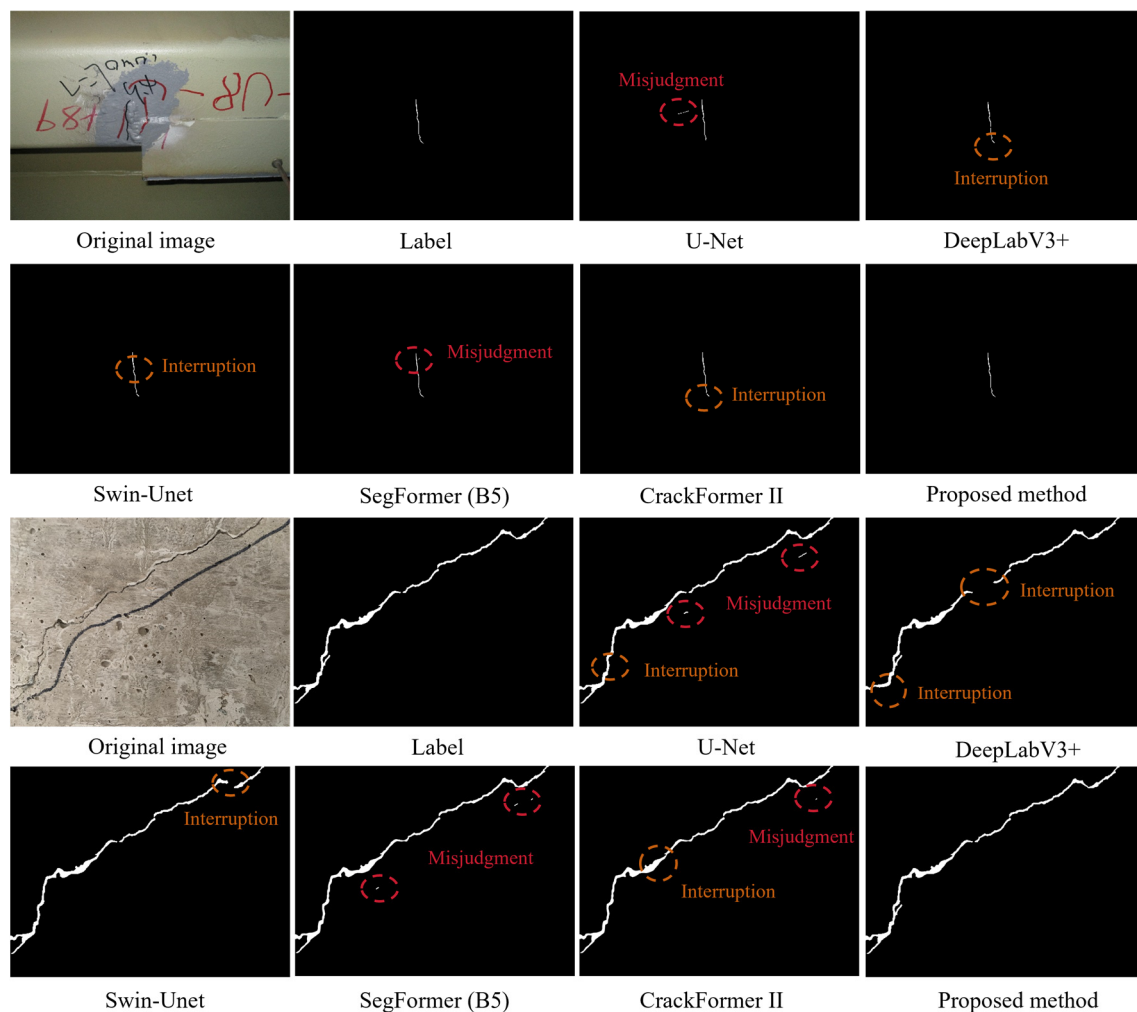


Figure 6. Some segmentation of binary image.

Table 3. Improved suppression performance against interfering substances.

Typical Interference	Initial False Positive Rate (%)	Post-Inference False Positive Rate (%)	Improvement
Marker Pen Marks	12.45	1.12	−91.0%
Steel Weld Seams	8.32	2.05	−75.4%
Concrete Pores	6.50	0.85	−86.9%
Calibration Rulers	4.12	0.50	−87.8%

Despite the overall strong performance, the proposed method exhibits limitations in specific scenarios. Visual inspection of Figure 6 reveals residual errors in heavily shadowed regions where crack contrast is extremely low, and in cases where compound interferences (e.g., weld seams overlapping with rust stains) partially evade both visual and logical filters. Additionally, highly tortuous cracks with abrupt direction changes ($>60^\circ$) occasionally trigger false connections to nearby features. To assess the risk of false topological connections, we manually reviewed all endpoint extension operations on the test set. Among 1247 evaluated endpoint pairs, the reasoning module triggered 89 connections based on the directional extension criteria (Equation (6)). Manual verification revealed 7 false connections (7.9% false connection rate), typically occurring when two spatially prox-

imate but structurally independent cracks aligned accidentally within the search radius $R = 20$ pixels. Despite these rare cases, the connectivity recovery mechanism correctly bridged 82 genuinely fractured segments, resulting in a net improvement in crack length estimation accuracy from 78.3% to 91.7%. Future work will explore adaptive search radii and multi-scale directional consensus to further reduce false connection risks.

4.4. Ablation Study and Effectiveness Analysis

To quantitatively evaluate the actual contributions of each core component in the proposed method, a series of ablation experiments was designed. The study utilized the basic hierarchical Swin-Transformer as the baseline model (M1), progressively incorporating the TAP module (M2) and the morphological reasoning module (M3) to analyze the impact of each module on detection performance. The experimental results are summarized in Table 4.

Table 4. Ablation experiment results.

Configuration	Baseline	TAP	Edge Smooth	Aspect Ratio	Direction Ext.	IoU (%)	Precision (%)	Recall (%)	F1 (%)
M1 (Swin-T Baseline)	✓					78.42	82.15	80.08	81.30
M2 (+TAP)	✓	✓				82.10	84.60	85.97	85.22
M3 (+Edge)	✓	✓	✓			83.85	86.42	86.15	86.28
M4 (+AR)	✓	✓	✓	✓		85.21	88.53	87.40	87.96
M5 (Full)	✓	✓	✓	✓	✓	86.72	91.45	89.30	90.36

Based on the experimental results, the introduction of the TAP module (comparing M1 and M2) improved the model's IoU by 3.68% and the F1-score by 3.92%. This validates the effectiveness of the task-aware query mechanism in capturing slender crack features. In the traditional self-attention mechanism, the model is easily influenced by large background textures (such as pores in concrete); however, the TAP module significantly enhances the semantic extraction capability under complex material backgrounds by learning local features of critical crack parts, such as bifurcation points and tips.

After incorporating the morphological reasoning module (comparing M2 and M3), the model's Precision experienced a qualitative leap, jumping from 84.60% to 91.45%, an increase of 6.85%. Simultaneously, the IoU further improved by 4.62%. This substantial growth reflects the unique advantage of morphological reasoning in handling unstructured noise. Pure deep learning models often struggle to distinguish between marker pen marks and real cracks due to a lack of physical constraints. The edge smoothness verification and topological connection logic introduced by the reasoning module effectively corrected the observation bias of the visual layer.

To isolate the contribution of each reasoning rule, we performed incremental ablation (M3–M5). Edge smoothness filtering (M3) primarily suppresses pen mark false positives, improving Precision by +1.8%. Aspect ratio screening (M4) effectively removes concrete pore noise, contributing an additional +2.1% Precision gain. Directional extension logic (M5) enhances Recall by +1.9% through topological repair, while further boosting Precision by +1.3% via false connection validation. The synergistic effect of all three rules yields the optimal performance in M5.

To validate the robustness of the fusion weight configuration, we conducted a sensitivity analysis by varying α from 0.3 to 0.9 (with $\beta = 1 - \alpha$) on the validation set. As shown in Figure 7, the F1-score reaches its peak at $\alpha = 0.6$, achieving a balance between visual perception sensitivity and reasoning robustness. When $\alpha < 0.5$, excessive reliance on reasoning weights causes the model to over-suppress weak but genuine crack signals in low-contrast regions. Conversely, when $\alpha > 0.7$, insufficient reasoning feedback fails to adequately filter false positives from marker pen marks and welds. The configuration

$\alpha = 0.6$, $\beta = 0.4$ demonstrates consistent performance across both steel and concrete subsets, with F1-score variance < 1.2%, confirming its generalization capability.

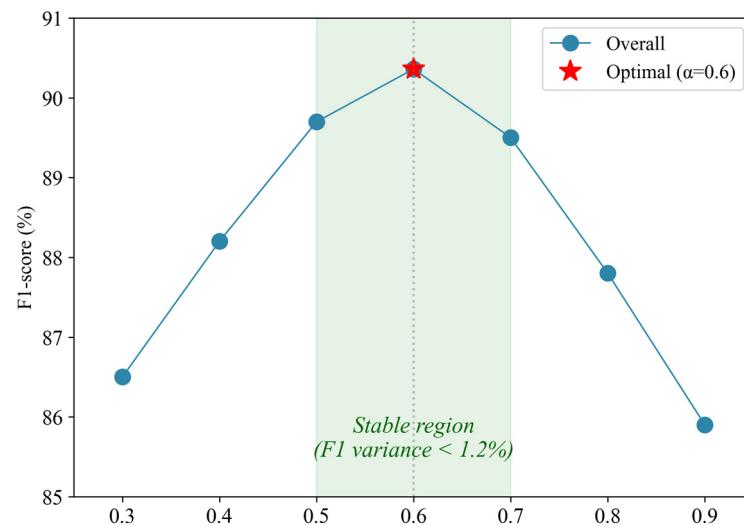


Figure 7. Sensitivity analysis of fusion weights showing F1-score vs. α .

In summary, the ablation study demonstrates that the robust visual features provided by TAPFormer serve as the foundation for detection, while the logical closed-loop provided by morphological reasoning rules is the key to achieving reliable detection under complex cross-material engineering conditions. The synergy between the two allows the proposed method to significantly suppress false positives caused by engineering interferences while maintaining a high recall rate.

5. Conclusions and Future Work

To address challenges such as severe background interference, limited long-range modeling capabilities of traditional CNNs, and poor adaptation to cross-material scenarios, this study proposes an automated crack detection method combining the TAPFormer model with mask reasoning rules. By integrating data-driven deep feature extraction with physical-prior-driven logical reasoning, robust identification of steel and concrete cracks is achieved. The primary conclusions are as follows:

(1) An architectural adaptation of TAPFormer is implemented as the feature extraction backbone. By utilizing an overlapping patch embedding layer, the model maintains sub-pixel continuity at patch boundaries, effectively mitigating the crack fragmentation common in standard Transformer architectures. Furthermore, the refactored topological sensing logic enables the precise capture of geometric key-points, such as crack tips and bifurcations, providing a robust visual foundation for feature alignment in heterogeneous (steel-concrete) environments.

(2) A mask reasoning module is developed as a logical constraint layer to enforce topological self-consistency. By quantifying geometric priors (including edge curvature variance, axial aspect ratios, and directional extension laws), this module effectively distinguishes real cracks from unstructured interferences. Unlike traditional post-processing filters, this perception-reasoning closed loop can “logically recover” fractured segments and suppress false positives. Experimental results show that this mechanism reduces handwriting-induced false positives by approximately 91.0%, significantly enhancing the logical rigor of the detection system.

(3) Extensive evaluations demonstrate that the proposed method achieves superior performance on cross-material test sets, yielding a mean IoU of 86.72% and an F1-score of 90.36%. The results consistently outperform classic CNNs (U-Net, DeepLabV3+) and

recent SOTA Transformer models (Swin-Unet, SegFormer, and CrackFormer II). Notably, the framework surpasses the strongest purely data-driven baseline (CrackFormer II) by 2.6% in IoU, showing that integrating a logical reasoning layer improves performance beyond conventional vision architectures in complex, noise-heavy engineering scenarios.

In summary, the proposed method exhibits superior detection accuracy and anti-interference capabilities in complex engineering environments. Several limitations should be acknowledged. First, the steel structure subset comprises images from a single bridge, which may limit generalizability to other steel structural types with different surface treatments or corrosion patterns. Second, while pixel-level metrics are standard in the field, crack-level quantitative metrics (length, width, connectivity) would better reflect engineering requirements; implementing such metrics requires additional skeleton extraction and sub-pixel measurement algorithms, which we identify as future work. Third, the reasoning module uses empirically determined thresholds that may require adjustment for structures with substantially different material properties or damage characteristics.

Future research will focus on the following directions: first, implementing model compression via knowledge distillation to meet the real-time monitoring requirements of embedded edge computing platforms or UAV inspection terminals; second, investigating few-shot adaptation techniques to explore the transfer learning capabilities and generalization logic of the model for damage identification in new materials, such as carbon fiber reinforced polymer.

Author Contributions: Conceptualization, H.P., H.W. and G.L.; methodology, H.P. and H.W.; software, H.P. and Y.D.; validation, H.P., G.L. and L.Z.; formal analysis, H.P., L.Z. and G.L.; investigation, H.P. and Y.D.; resources, H.P. and L.Z.; data curation, H.P.; writing—original draft preparation, H.P., L.Z. and G.L.; writing—review and editing, H.P.; visualization, H.P. and Y.D.; supervision, H.P.; project administration, H.P.; funding acquisition, G.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded by the authors.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gharehbaghi, V.R.; Noroozinejad Farsangi, E.; Noori, M.; Yang, T.Y.; Li, S.; Nguyen, A.; Málaga-Chuquitaype, C.; Gardoni, P.; Mirjalili, S. A critical review on structural health monitoring: Definitions, methods, and perspectives. *Arch. Comput. Methods Eng.* **2022**, *29*, 2209–2235.
2. Kosova, F.; Altay, Ö.; Ünver, H.Ö. Structural health monitoring in aviation: A comprehensive review and future directions for machine learning. *Nondestruct. Test. Eval.* **2025**, *40*, 1–60.
3. Mardanshahi, A.; Sreekumar, A.; Yang, X.; Barman, S.K.; Chronopoulos, D. Sensing techniques for structural health monitoring: A state-of-the-art review on performance criteria and new-generation technologies. *Sensors* **2025**, *25*, 1424. [[PubMed](#)]
4. Plevris, V.; Papazafeiropoulos, G. AI in structural health monitoring for infrastructure maintenance and safety. *Infrastructures* **2024**, *9*, 225. [[CrossRef](#)]
5. Shi, Y.; Wang, Y.; Wang, L.-N.; Wang, W.-N.; Yang, T.-Y. Bridge cable performance warning method based on temperature and displacement monitoring data. *Buildings* **2025**, *15*, 2342. [[CrossRef](#)]
6. Ma, F.; Zheng, K.; Li, Z.; Zhang, M.; Zhu, F.; Qu, X. A high-fidelity monitoring method for stress distribution in gravity-based offshore wind turbine towers using a well-conditioned orthogonal basis expansion model. *Ocean Eng.* **2025**, *325*, 120884.
7. Zhao, Y.; Jiang, S.; An, Z.; Luo, B. Experimental Investigation of the Bearing-Deformation Behavior of Broken Rocks in Goafs Under Various Influencing Factors. *Appl. Sci.* **2026**, *16*, 4276. [[CrossRef](#)]
8. Yuan, Q.; Shi, Y.; Li, M. A review of computer vision-based crack detection methods in civil infrastructure: Progress and challenges. *Remote Sens.* **2024**, *16*, 2910. [[CrossRef](#)]

9. Zhuang, H.; Cheng, Y.; Zhou, M.; Yang, Z. Deep learning for surface crack detection in civil engineering: A comprehensive review. *Measurement* **2025**, *248*, 116908. [\[CrossRef\]](#)
10. Pandey, V.; Mishra, S.S. A review of image-based deep learning methods for crack detection. *Multimed. Tools Appl.* **2025**, *84*, 35469–35511. [\[CrossRef\]](#)
11. Wu, Y.; Li, S.; Li, J.; Yu, Y.; Li, J.; Li, Y. Deep learning in crack detection: A comprehensive scientometric review. *J. Infrastruct. Intell. Resil.* **2025**, *4*, 100144. [\[CrossRef\]](#)
12. Liu, X.; Ghazali, K.H.; Han, F.; Mohamed, I.I. Review of CNN in aerial image processing. *Imaging Sci. J.* **2023**, *71*, 1–13. [\[CrossRef\]](#)
13. Salehi, A.W.; Khan, S.; Gupta, G.; Alabdullah, B.I.; Almjally, A.; Alsolai, H.; Siddiqui, T.; Mellit, A. A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope. *Sustainability* **2023**, *15*, 5930. [\[CrossRef\]](#)
14. Yao, W.; Bai, J.; Liao, W.; Chen, Y.; Liu, M.; Xie, Y. From cnn to transformer: A review of medical image segmentation models. *J. Imaging Inform. Med.* **2024**, *37*, 1529–1547. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Seol, D.; Kim, H.; Park, S.; Lee, J. Comparative study of cnn-based deep learning models for concrete crack detection. *J. Build. Eng.* **2025**, *112*, 113651. [\[CrossRef\]](#)
16. Lertrudachakul, T.; Leni, P.-E. Comparison of DeepLab v3+ base networks for crack segmentation under limited computational resources. *Sci. Eng. Health Stud.* **2025**, *19*, 25020009. [\[CrossRef\]](#)
17. Su, L.; Zuo, X.; Li, R.; Wang, X.; Zhao, H.; Huang, B. A systematic review for transformer-based long-term series forecasting. *Artif. Intell. Rev.* **2025**, *58*, 80.
18. Tang, Y.; Wang, Y.; Guo, J.; Tu, Z.; Han, K.; Hu, H.; Tao, D. A survey on transformer compression. *arXiv* **2024**, arXiv:2402.05964.
19. Zhang, E.Y.; Cheok, A.D.; Pan, Z.; Cai, J.; Yan, Y. From turing to transformers: A comprehensive review and tutorial on the evolution and applications of generative transformer models. *Sci* **2023**, *5*, 46. [\[CrossRef\]](#)
20. Xiao, H.; Li, L.; Liu, Q.; Zhu, X.; Zhang, Q. Transformers in medical image segmentation: A review. *Biomed. Signal Process. Control* **2023**, *84*, 104791. [\[CrossRef\]](#)
21. Alshawabkeh, S.; Wu, L.; Dong, D.; Cheng, Y.; Li, L. A hybrid approach for pavement crack detection using mask R-CNN and vision transformer model. *Comput. Mater. Contin.* **2025**, *82*, 561. [\[CrossRef\]](#)
22. Honarjoo, A.; Darvishan, E.; Rezazadeh, H.; Kosarieh, A.H. Damage detection and localization of structural cracks based on dynamic attention based transformer. *Int. J. Build. Pathol. Adapt.* **2026**, *44*, 339–357.
23. Liu, J.; Tan, Z.; Zhang, J.; Zhou, Y.; Shen, H.; Chen, X.; Hu, D. TAPFormer: Robust Arbitrary Point Tracking via Transient Asynchronous Fusion of Frames and Events. *arXiv* **2026**, arXiv:2603.04989.
24. He, Z.; Chen, W.; Zhang, J.; Wang, Y.-H. Crack segmentation on steel structures using boundary guidance model. *Autom. Constr.* **2024**, *162*, 105354. [\[CrossRef\]](#)
25. Gong, H.; Liu, L.; Liang, H.; Zhou, Y.; Cong, L. A state-of-the-art survey of deep learning models for automated pavement crack segmentation. *Int. J. Transp. Sci. Technol.* **2024**, *13*, 44–57.
26. El-Din Hemdan, E.; Al-Atroush, M.E. A review study of intelligent road crack detection: Algorithms and systems. *Int. J. Pavement Res. Technol.* **2025**, *18*, 1–31. [\[CrossRef\]](#)
27. Karimi, N.; Mishra, M.; Lourenço, P.B. Automated surface crack detection in historical constructions with various materials using deep learning-based YOLO network. *Int. J. Archit. Herit.* **2025**, *19*, 581–597.
28. Cha, Y.J.; Choi, W.; Büyüköztürk, O. Deep learning-based crack damage detection using convolutional neural networks. *Comput. Aided Civ. Infrastruct. Eng.* **2017**, *32*, 361–378.
29. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE Computer Society: Washington, DC, USA, 2015; pp. 3431–3440.
30. Azad, R.; Aghdam, E.K.; Rauland, A.; Jia, Y.; Avval, A.H.; Bozorgpour, A.; Karimijafarbigloo, S.; Cohen, J.P.; Adeli, E.; Merhof, D. Medical image segmentation review: The success of u-net. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 10076–10095. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Park, S.E.; Eem, S.-H.; Jeon, H. Concrete crack detection and quantification using deep learning and structured light. *Constr. Build. Mater.* **2020**, *252*, 119096. [\[CrossRef\]](#)
34. Xu, Y.; Bao, Y.; Chen, J.; Zuo, W.; Li, H. Surface fatigue crack identification in steel box girder of bridges by a deep fusion convolutional neural network based on consumer-grade camera images. *Struct. Health Monit.* **2019**, *18*, 653–674.
35. Jing, J.; Ding, L.; Yang, X.; Feng, X.; Guan, J.; Han, H.; Wang, H. Topology-informed deep learning for pavement crack detection: Preserving consistent crack structure and connectivity. *Autom. Constr.* **2025**, *174*, 106120. [\[CrossRef\]](#)

36. Shiri, F.M.; Perumal, T.; Mustapha, N.; Mohamed, R. A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU. *arXiv* **2023**, arXiv:2305.17473.
37. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
38. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE Computer Society: Washington, DC, USA, 2021; pp. 10012–10022.
39. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
40. Apedo, Y.; Tao, H.; Gao, W.; Xie, C.; Zhao, S. Unsupervised Domain Adaptation for Crack Segmentation via Cross-Domain Stylization and Dual Adversarial Feature Learning. *J. Comput. Civ. Eng.* **2026**, *40*, 04025146. [[CrossRef](#)]
41. Yuan, L.; Ni, Y.-Q.; Rui, E.-Z.; Zhang, W. Structural damage inverse detection from noisy vibration measurement with physics-informed neural networks. In *Proceedings of the Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2024; p. 192013.
42. Wu, J.; Zhang, X. Tunnel crack detection method and crack image processing algorithm based on improved retinex and deep learning. *Sensors* **2023**, *23*, 9140. [[CrossRef](#)] [[PubMed](#)]
43. Dorafshan, S.; Thomas, R.J.; Maguire, M. SDNET2018: An annotated image dataset for non-contact concrete crack detection using deep convolutional neural networks. *Data Brief* **2018**, *21*, 1664–1668. [[CrossRef](#)] [[PubMed](#)]
44. Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A database and web-based tool for image annotation. *Int. J. Comput. Vis.* **2008**, *77*, 157–173.
45. Lau, S.L.H.; Chong, E.K.P.; Yang, X.; Wang, X. Automated pavement crack segmentation using u-net-based convolutional neural network. *IEEE Access* **2020**, *8*, 114892–114899. [[CrossRef](#)]
46. Song, F.; Wang, D.; Dai, L.; Yang, X. Concrete bridge crack semantic segmentation method based on improved DeepLabV3+. In *Proceedings of the 2024 IEEE 13th Data Driven Control and Learning Systems Conference (DDCLS)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1293–1298.
47. Chen, S.; Feng, Z.; Xiao, G.; Chen, X.; Gao, C.; Zhao, M.; Yu, H. Pavement crack detection based on the improved Swin-Unet model. *Buildings* **2024**, *14*, 1442. [[CrossRef](#)]
48. Li, H.; Zhang, H.; Zhu, H.; Gao, K.; Liang, H.; Yang, J. Automatic crack detection on concrete and asphalt surfaces using semantic segmentation network with hierarchical Transformer. *Eng. Struct.* **2024**, *307*, 117903. [[CrossRef](#)]
49. Ullah, R.; Wahab, F.; Bilal, H.; Ayouni, S.; Phanomchoeng, G. CrackFormer: A reinforcement learning-enhanced Swin Transformer V3 framework for high-fidelity detection and segmentation of pavement surface cracks. *Alex. Eng. J.* **2025**, *131*, 183–198.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.