Transport and Communications Science Journal

ENHANCING CRACK SEGMENTATION IN FUSED RGB-IR IMAGES WITH CSWIN TRANSFORMER AND SEMANTIC FEATURE PYRAMID NETWORK

Nguyen Ngoc Long, Vu Manh Trung, Phung Ngoc Hung, Nguyen Dan Le, Nguyen Ngoc Lan*

University of Transport and Communications, No 3 Cau Giay Street, Hanoi, Vietnam

ARTICLE INFO

TYPE: Research Article
Received: 21/01/2026
Revised: 11/03/2026
Accepted: 17/03/2026
Published online: 15/05/2026
https://doi.org/10.47869/tcsj.77.4.13

* *Corresponding author:*
Email: nngoclan@utc.edu.vn; Tel: 0915432768

Abstract. Surface crack segmentation is a critical task in structural health monitoring (SHM), serving as an early indicator of structural deterioration and safety risks. Recently, deep learning-based computer vision has emerged as a dominant approach for automating defect detection, gradually replacing manual inspections. However, traditional convolutional neural networks (CNNs) often struggle to capture long-range dependencies. To address this limitation, this paper introduces the CSWin-Semantic FPN, a model integrating the transformer architecture with cross-shaped window (CSWin) attention and a semantic feature pyramid network (Semantic FPN) to optimize feature extraction. Notably, this study utilizes a multi-modal fusion dataset-combining optical and thermal infrared images collected in real-world environments. This fusion approach significantly enhances crack signals against complex backgrounds, facilitating more effective model training. Experimental results demonstrate that the CSWin-Semantic FPN achieves an impressive intersection over union (IoU) of 70.53%, significantly outperforming ResUNet (59.44%), SwinUNet (57.91%), and UNet (51.79%). These findings confirm the potential of hybrid Transformer architectures combined with multi-modal data in providing reliable and automated SHM solutions.

Keywords: Crack segmentation, computer vision, cross-shaped window attention, semantic feature pyramid network, asphalt pavement.

1. INTRODUCTION

In the process of modernizing transportation infrastructure, bridge and road systems not only connect physical spaces but also serve as socio-economic lifelines. As the service life of these structures increases under the pressures of dynamic loads, humidity, and environmental factors, the central challenge has shifted from new construction to maintaining the safety of existing structures. Within this context, surface cracks are the most immediate and sensitive indicators of diminished material strength and a loss of structural continuity. From the perspective of structural health monitoring (SHM), cracks are not merely qualitative signs but are critical parameters that require precise morphological quantification [1], [2], [3]. However, the current inspection protocols in Vietnam still rely heavily on manual surveys, leading to a lack of continuous, synchronized monitoring data essential for overseeing large-scale transportation networks.

In order to overcome the mentioned limitation, the global trend is shifting greatly to SHM solutions based on computer vision [4] and deep learning [5], [6], [7], using unmanned aerial vehicle and high resolution camera to digitalize construction's current condition. Within the framework, the requirement no longer stop at detecting the existence of cracks but more urgently is to perform accurate pixel-level segmentation of these defects. Image segmentation plays a crucial role in completely separating damaged regions from the material background, thereby providing precise input data for measuring crack width, length, and propagation direction - key indicators for assessing the structural safety level.

The development of convolutional neural network made a breakthrough in automatically extracting features from raw image, overcome traditional image processing methods. Among these, UNet architecture with encoder - decoder symmetrical structure and skip connections became the benchmark for medical and structural segmentation tasks thanks to its ability to preserve detailed spatial information. However, the convolution operations at the core of CNNs and UNet inherently process information within local receptive fields. This creates a major limitation when applied to the harsh real-world conditions of bridges and pavements, where cracks are often thin, elongated, and anisotropic. Relying solely on local features makes the model susceptible to discontinuities when segmenting long cracks, or to misclassification in low-contrast environments with complex surface noise. In practical, non-laboratory environments, the information needed to fully recognize a crack is typically distributed across the entire image, rather than contained in a small local region. Purely CNN-based models therefore struggle to capture such long-range dependencies effectively [8].

The structural limitations of convolution operations highlight the urgent need for new network architectures capable of explicitly capturing global contextual information. Modern crack segmentation requires a model that not only accurately identifies local boundaries but also understands the geometric propagation pattern of cracks across the entire image. This necessity drives the adoption of self-attention mechanisms and Transformer-based architectures in computer vision, aiming to overcome the blind spots of traditional CNN methods and enhance the accuracy of structural damage quantification.

To thoroughly address the aforementioned challenges, this study proposes an advanced hybrid architecture named CSWin-Semantic FPN. This model strategically integrates the global contextual representation power of the CSWin Transformer (cross-shaped window transformer) [9] as a backbone with the robust multi-scale feature aggregation capabilities of the semantic feature pyramid network (Semantic FPN) [10].

In the proposed architecture, CSWin Transformer blocks are utilized to extract hierarchical features. By computing self-attention within cross-shaped windows arranged horizontally and vertically, CSWin effectively expands the receptive field to cover the entire image. This allows the model to overcome the local blind spots of traditional CNNs and to capture the continuity of long, discontinuous, or complex branching crack patterns more effectively.

In parallel, the Semantic FPN serves as the decoding mechanism to enhance feature reconstruction. Unlike simple skip connections, the Semantic FPN systematically merges feature maps from different stages of the transformer backbone. This mechanism effectively fuses high-level semantic information (crucial for distinguishing cracks from complex backgrounds) with low-level spatial details (essential for precise boundary delineation). The synergy between the global perception of CSWin and the hierarchical feature fusion of Semantic FPN is expected to deliver a significant breakthrough in accuracy and robustness for concrete crack segmentation under real-world conditions

Beyond model innovations, this research implements a multi-modal data fusion paradigm to bolster the framework's robustness against stochastic environmental variables [11], [12]. While high-resolution RGB imagery is proficient in delineating intricate surface textures, its reliability is frequently compromised by spectral ambiguities. Visual artifacts - such as transient shadows, inhomogeneous material backgrounds, and surface contaminants like oil stains or tire marks - often manifest as pseudo-crack patterns, leading to significant false-positive detections in purely convolutional models.

To mitigate these limitations, thermal infrared (IR) imagery is integrated to provide a distinct thermodynamic signature of structural anomalies. Surface cracks inherently exhibit anomalous heat retention or localized moisture accumulation, which generates perceptible thermal radiation disparities compared to the surrounding intact pavement. Crucially, the IR modality operates independently of ambient lighting conditions and is invariant to shadow-induced occlusions, ensuring a consistent signal-to-noise ratio regardless of the time of acquisition. Ultimately, this cross-modal fusion enables the architecture to construct a synergistic feature representation that effectively decouples the intrinsic morphological signatures of cracks from pervasive background interference, thereby establishing a more resilient and precise paradigm for automated SHM.

The remainder of this paper is organized as follows. Section 2 details the proposed CSWin-Semantic FPN, focusing on the integration of the CSWin Transformer backbone for global context extraction and the Semantic FPN for multi-scale feature fusion. Section 3 describes the dataset acquisition and preprocessing steps. Section 4 presents the experimental setup and a comprehensive evaluation against benchmarks like UNet, SWinUNet and ResUNet using quantitative metrics and visual assessments. Finally, Section 5 summarizes the key contributions and outlines future research directions.

2. PROPOSED CRACK IMAGE SEGMENTATION METHOD

2.1. CSWin Transformer Encoder Block

In recent years, self-attention mechanism has becomed a dominant trend in image segmentation by overcoming the receptive field constraints of traditional CNNs. This is vital for transportation infrastructure, where cracks are often fine-scale, low-contrast, and embedded in noisy backgrounds. In these cases, capturing both pixel-level details and long-range context is imperative. CSWin Transformer addresses this via a cross-shaped window attention

mechanism, which efficiently models horizontal and vertical dependencies. This architectural strength significantly improves the reliability of extracting cracks from complex background clutter, ensuring high precision in challenging environments [13].

As illustrated in Figure 1, the multi-head self-attention (MHSA) setup projects the input feature map $X \in R^{H\times W\times C}$ into N parallel heads, where N is typically an even number. Instead of applying global attention across the entire grid, CSWin partitions the feature space into non-overlapping horizontal or vertical stripes. Each head operates solely within its assigned stripe, with all stripes processed concurrently. This organization maintains computational locality while effectively expanding the attention coverage along the primary structural axes of the image, capturing crucial long-range dependencies.

For the horizontal stripes, feature map $X \in R^{H\times W\times C}$ is divided into M non - overlap stripes $X_t^1, X_t^2, \ldots, X_t^M$, each with a height of s_w, such that $M = H/s_w$.

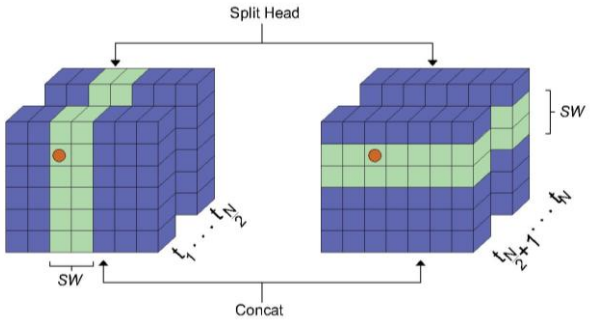


Figure 1. Illustration of CSWin self-attention mechanism along horizontal and vertical stripes.

The stripe size s_w serves as a hyperparameter that directly controls the depth of contextual information within each stripe: larger values allow the model to capture longer-range dependencies along the vertical direction, but at the cost of increased computational overhead for each attention head. For the n-th head (with head dimension $d_t = C/N$), the query-key-value projections are computed independently within each stripe as follows:

$$Q_n^i = W_n^Q X_t^i \tag{1}$$

$$K_n^i = W_n^K X_t^i \tag{2}$$

$$V_n^i = W_n^V X_t^i \tag{3}$$

And the attention within the i-th stripe is computed using the standard scaled dot-product formulation:

$$Y_n^i = Softmax(\frac{Q_n^i (K_n^i)^\top}{\sqrt{d_t}})V_n^i \tag{4}$$

Where $X_t^i \in \mathbb{R}^{s_w\times W\times C}$, and $W_n^Q, W_n^K, W_n^V \in \mathbb{R}^{C\times d_t}$denote the projection matrices of the n-th head. After the attention is computed independently within each stripe, the resulting outputs are concatenated along the spatial dimension to obtain the final representation for the n-th head:

$$X = [X_t^1, X_t^2, \dots, X_t^M] \tag{5}$$

$$H\text{-}Attention_n(X) = [Y_n^1, Y_n^2, \dots, Y_n^M] \tag{6}$$

Similarly, for the vertical stripes, the feature map is partitioned into S stripes $X_v^1, X_v^2, \dots, X_v^S$ with the same width s_w, such that $S = W/s_w$. For each head n, with $d_t = C/N$, the projections within each vertical stripe are formulated as follows:

$$Q_n^i = W_n^Q X_v^i \tag{7}$$

$$K_n^i = W_n^K X_v^i \tag{8}$$

$$V_n^i = W_n^V X_v^i \tag{9}$$

And the attention within the i-th vertical stripe is computed by:

$$Y_n^i = \text{Softmax}(\frac{Q_n^i (K_n^i)^\top}{\sqrt{d_t}}) V_n^i \tag{10}$$

Where $X_v^i \in \mathbb{R}^{H \times s_w \times C}$. When concatenating the outputs from all S vertical stripes, the self-attention along the vertical direction for the n-th head is formulated as follows:

$$X = [X_v^1, X_v^2, \dots, X_v^S] \tag{11}$$

$$V\text{-}Attention_n(X) = [Y_n^1, Y_n^2, \dots, Y_n^S] \tag{12}$$

Two attention architectures along both directions (horizontal and vertical) in this manner help CSWin exploit spatial structures along two orthogonal axes, while keeping the computational cost within a reasonable budget by computing local attention within each stripe.

To combine these two directions in a single module, the heads are divided into two types of attention. In practice, the N heads are separated into two equal groups: the first group ($N/2$ head) applies horizontal attention, while the second group ($N/2$ head) applies vertical attention. After the computation is completed, each head outputs a tensor t_n:

$$t_n = \begin{cases} H\text{-}Attention_n(X), & n = 1,2, \dots, N/2, \\ V\text{-}Attention_n(X), & n = N/2+1, \dots, N. \end{cases} \tag{13}$$

These outputs are concatenated along the channel dimension and projected through a linear matrix:

$$CSWin\text{-}Attention(X) = Concat(t_1, t_2, \dots, t_N) W^O \tag{14}$$

Where $W^O \in \mathbb{R}^{C \times C}$ serves as a channel-mixing projection that aggregates information across different heads and fuses the contextual representations obtained from both the horizontal and vertical directions. As a result, the model produces a feature representation that encodes spatial information along two orthogonal axes, while still being constructed from parallel local attention operations and therefore maintaining a manageable computational cost.

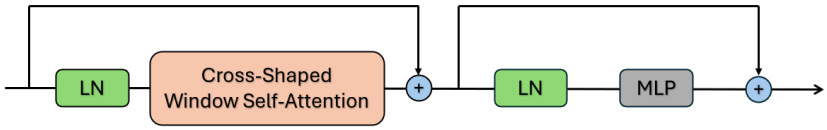


Figure 2. Block diagram of the CSWin Transformer.

In the overall architectural overview shown in Figure 2, the CSWin-Attention modules are encapsulated within a CSWin Transformer block following a residual layout in a pre-norm manner. Instead of applying normalization after the attention operation, the input of the l-th block is normalized immediately before being fed into the attention module; the resulting attention output is then added to the original input to form the first residual branch:

$$\tilde{X}^{(l)} = CSWin\text{-}Attention(LN(X^{(l-1)})) + X^{(l-1)} \quad (15)$$

Next, $\tilde{X}^{(l)}$ is normalized once again, passed through a two-layer (or deeper) MLP, and then added via a residual connection to produce the output of the block:

$$X^{(l)} = MLP(LN(\tilde{X}^{(l)})) + \tilde{X}^{(l)} \quad (16)$$

Where $X^{(l)}$ denotes the representation after the l-th Transformer block; for the first block in a stage, $X^{(l)}$ corresponds to the output of the preceding convolutional layer or down-sampling block. This pre-norm residual arrangement has been shown to improve the stability of deep Transformer models and to mitigate gradient degradation as the number of layers increases.

2.2. Semantic Feature Pyramid Network

Semantic FPN is based on changing the traditional inter-layer connections by focusing on directly consolidating multi-scale features to enrich contextual information for low-level feature maps, and improving the ability to represent semantic features at high-resolution levels, which is crucial for capturing small objects such as tumors or thin cracks [14].

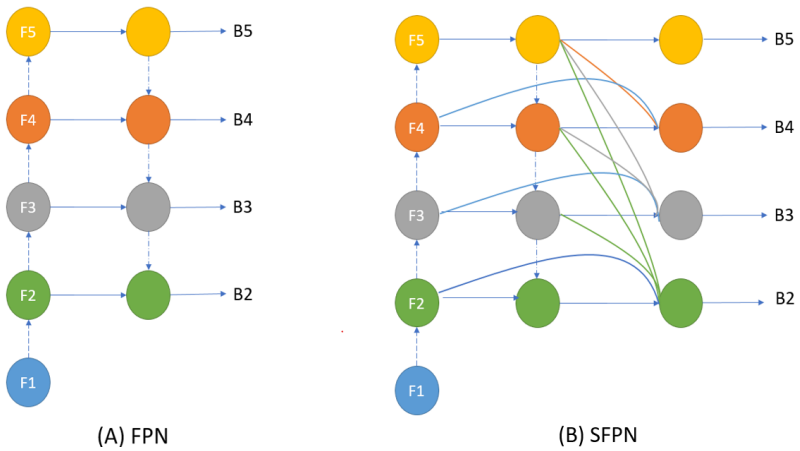


Figure 3. Comparison between the traditional FPN and the Semantic FPN.

Different from the traditional FPN (Figure 3), which passes information along top-down or horizontal directions, Semantic FPN, based on the principle of enriching contextual information, performs a direct backward propagation of semantic information from the deepest layers to shallower ones, helping the model capture small objects. The model uses a multi-scale feature aggregation function f_i to combine data from various sources through three steps, starting with

multi-dimensional lateral connections to preserve the resolution and contextual information of the current layer. In the next step, through multi-level up-sampling, the module integrates features from different scales rather than only adjacent levels. This process is formulated as follows:

$$\begin{cases} B_5 = F_5 \\ B_4 = f_4\left(F_4, S_p(F_4), up(F_5)\right) \\ B3 = f_3\left(F_3, S_p(F_3), up(F_5), up(F_4)\right) \\ B_2 = f_2\left(F_2, S_p(F_2), up(F_5), up(F_4), up(F_3)\right) \end{cases} \quad (17)$$

Where f_i is a function that performs multi-scale feature fusion at the i-th layer; B_{i+1}denotes the high-level feature propagated through the up(.) function, and S_p represents the skip connection. In the final step, Semantic FPN aggregates all feature maps and applies a single convolutional layer to fuse them, producing a unified feature map with highly accurate spatial localization and clear semantic meaning of the object. This enables the model to confidently distinguish true objects from background noise.

2.3. The overall architecture of the proposed CSWin-Semantic FPN model

The CSWin-Semantic FPN architecture (Figure 4) addresses the problem of surface crack segmentation under challenging environmental conditions such as noise, illumination variations, and inhomogeneous material backgrounds. The CSWin Transformer backbone extracts hierarchical features through four stages with gradually decreasing resolutions. The cross-shaped attention structure exploits contextual relationships along both horizontal and vertical directions, enabling the architecture to capture complex crack patterns while maintaining higher computational efficiency than global self-attention.

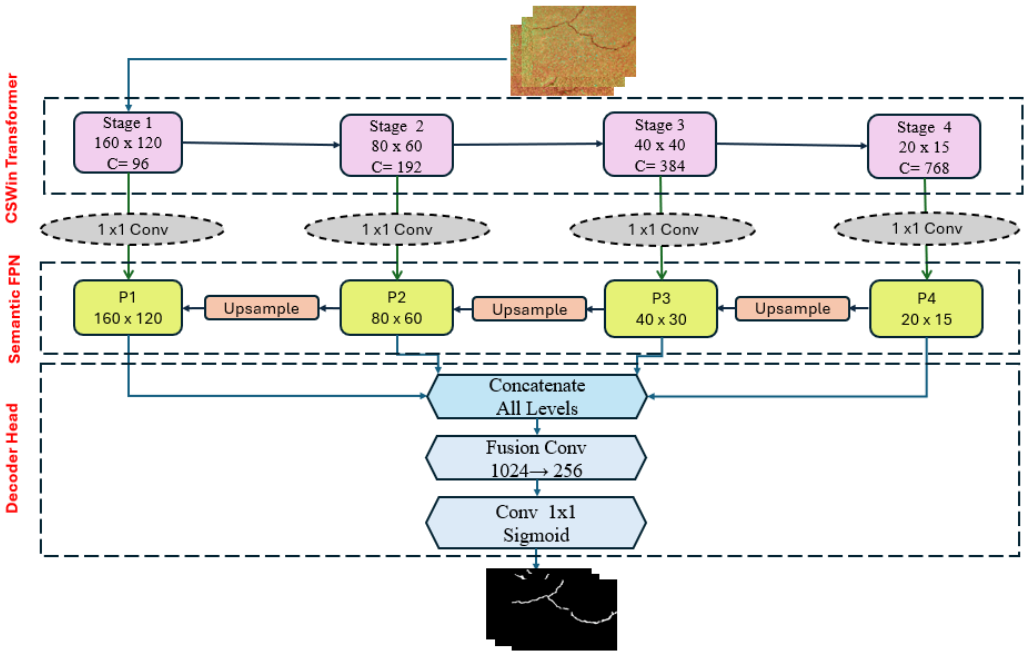


Figure 4. Overall architecture of the proposed CSWin-Semantic FPN model for crack segmentation.

The Semantic FPN block normalizes and integrates multi-scale information through a top-down pathway to produce feature maps from P1 to P4. Deeper stages provide discriminative semantic representations that separate true cracks from stains or intrinsic material textures, whereas shallower stages preserve boundary sharpness and the fine details of thin cracks. This design enables more accurate identification of crack characteristics and more reliable delineation of crack contours.

Features from all levels are connected and compressed through the Fusion Convolution block to distill information, and then fed into the Sigmoid activation to generate a binary mask. The model integrates the Transformer's contextual strength with the FPN's level of detail, ensuring that the output corresponds to the correct damaged region and accurately captures the crack contour.

3. DATASET DESCRIPTION

To evaluate the effectiveness of the proposed model for crack segmentation under varying environmental and illumination conditions, this study employs a benchmark dataset of asphalt pavement cracks released by Liu et al [11]. This is one of the few publicly available datasets that provides both optical imagery and IR thermal data, enabling comprehensive exploitation of the crack's multimodal features.

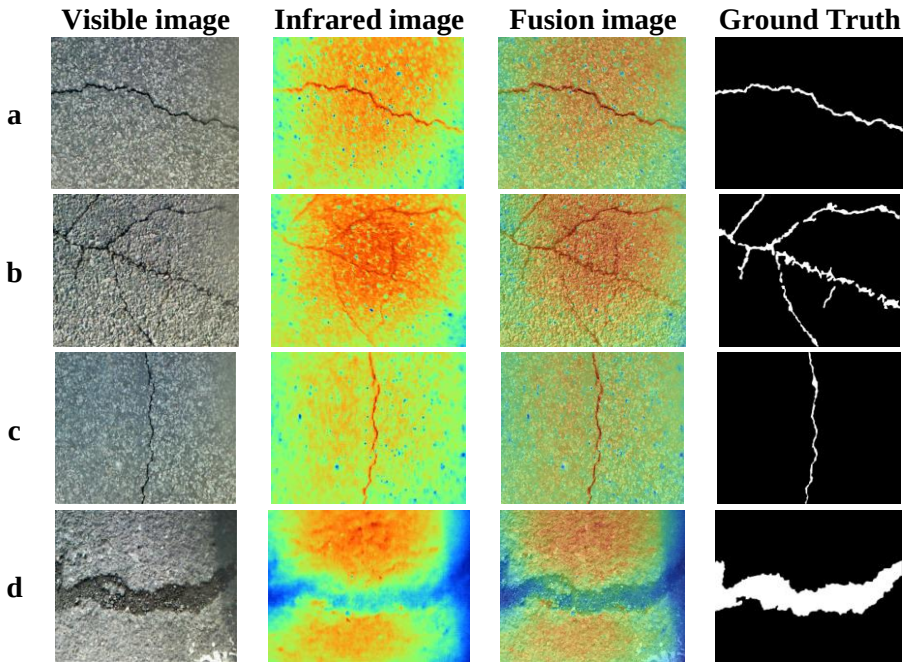


Figure 5. Crack scenarios classified by geometric characteristics: (a) single crack, (b) multi crack, (c) thin crack, (d) thick crack.

The original image data were collected in real-world conditions using the specialized infrared camera Fluke TiX580, which features an integrated dual-lens system combining an optical sensor and a thermal sensor. This configuration enables simultaneous acquisition of pavement surface conditions in two different spectra from the same viewpoint, producing multimodal input data with a standardized resolution of 640×480 pixels [11].

The dataset comprises a total of 448 samples, each consisting of four distinct image subsets designed for training and validation. Beyond the optical imagery providing chromatic details and IR imagery capturing thermal information, the dataset features fusion imagery generated

via IR-Fusion™ technology, which integrates a 50:50 ratio of RGB and IR data. The ground truth for the semantic segmentation task was expertly annotated at the pixel level using Adobe Photoshop, ensuring high-fidelity differentiation between crack features and the material substrate [11].

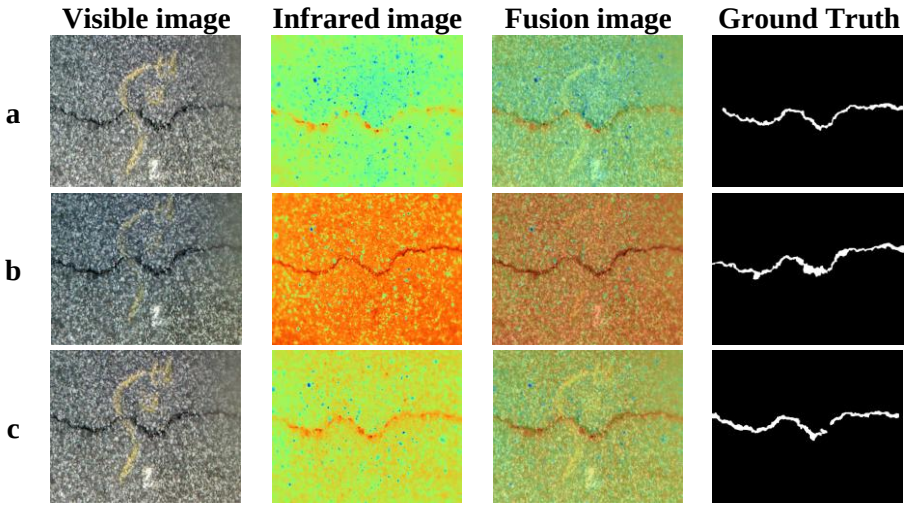
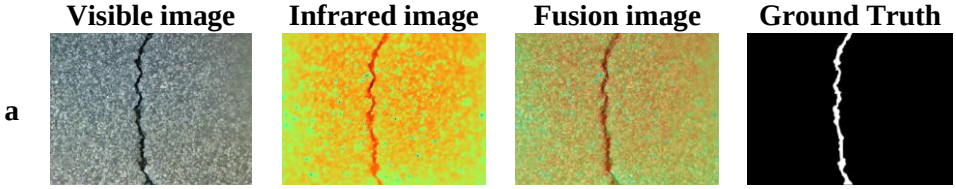


Figure 6. Crack scenarios according to the capture time: (a) morning, (b) noon, (c) afternoon.

The diversity and complexity of the data are ensured through a collection process conducted at three different times of the day: morning (08:00), noon (12:00), and afternoon (17:00). This is intended to simulate variations in pavement surface temperature, from low levels (0.5-7.5 °C in the morning) to the highest levels (17.1-31.0 °C at noon), thereby exploiting temperature differences to highlight crack features in IR images (Figure 5).

In terms of morphology, the data cover multiple damage scenarios, including single cracks, multi cracks, thin cracks, and thick cracks. The dataset also records environmental noise such as uneven pavement background, dark lighting conditions, and changes in surface texture, as shown in Figure 6 and Figure 7. Regarding the experimental setup, the dataset was randomly partitioned into two subsets: a training set comprising 80% (358 images) and a validation set accounting for the remaining 20% (90 images).

To improve the model's ability to generalize and to counter the overfitting risks associated with small datasets, we utilized the AugmentationTransform class to execute a geometric data augmentation technique. The process introduces spatial diversity by applying random operations, such as vertical and horizontal flipping, discrete rotations of 90°, 180°, and 270°, and random crops with scaling factors between 0.75 and 1.0 to reflect different observational distances. Most importantly, to preserve the exact spatial relationship between the crack pixels and the ground truth during training, these geometric adjustments are performed synchronously on both the raw input images and their associated segmentation masks.



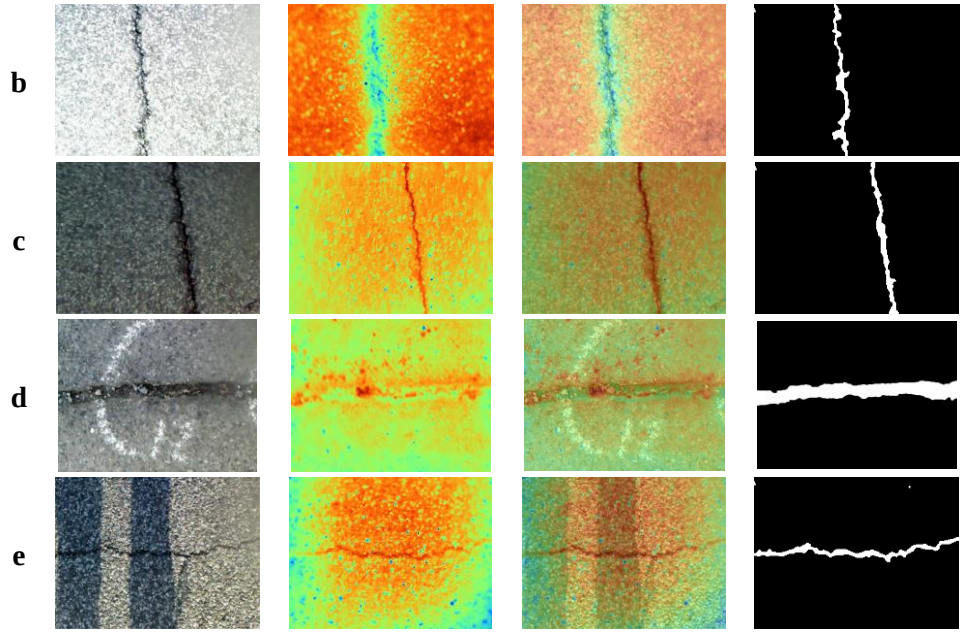


Figure 7. Representative crack scenarios categorized by background conditions: (a) normal, (b) white, (c) dark, (d) dusty, and (e) shadowed.

4. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the experimental results of the proposed model and provides an in-depth analysis of its performance and accuracy. To ensure an objective and comprehensive evaluation, the proposed architecture is compared two widely used computer vision models - UNet [15] and ResUNet [16] - both of which have demonstrated strong effectiveness in numerous image segmentation tasks. All training experiments were conducted on a workstation running Windows 11 (64-bit) equipped with an Intel Core i9-13900HX CPU, 16 GB of RAM, and an NVIDIA GeForce RTX 4060 GPU with 8 GB of VRAM. The implementation environment was based on Python 3.10.13.

Table 1. Comprehensive Implementation and Training Specifications

Category	**Technical Specification**
Backbone Architecture	CSWin Transformer (Embed dim 96, Depth [2, 4, 32, 2], Num heads [4, 8, 16, 32])
Decoder Architecture	Semantic FPN with 4 lateral stages (P1-P4) and 256 feature channels
Input Resolution	640x480 pixels
Normalization	Pixel intensity scaling to the range [0,1] (Division by 255)
Data Augmentation	Horizontal/Vertical flip (50% prob), Random Rotation (90º, 180º, 270º), Random Crop (75%-100% scale)
Optimizer	AdamW (Weight decay: 1×10^{-4})
Loss Function	Binary Cross Entropy with Logits (BCE With Logits Loss)
Learning Rate	1×10^{-3} with ReduceLROnPlateau (Factor: 0.5, Patience: 5 epochs)
Training Duration	100 Epochs
Batch Size	4 (Optimized for 640x480 resolution and memory constraints)
Binary Threshold	Fixed at 0.5 for Sigmoid output binarization

The model was trained for 100 epochs with a batch size of 4 - were selected based on a combination of established literature in crack segmentation and an empirical fine-tuning process to ensure maximum training stability. Binary Cross Entropy (BCE) was employed as the loss function. While loss functions specifically designed for imbalanced datasets, such as Dice Loss and Focal Loss, were considered and evaluated during our preliminary ablation studies, standard BCE demonstrated the most stable convergence and yielded the highest overall segmentation metrics for the proposed CSWin-Semantic FPN architecture. Optimization was performed using the Adam optimizer with an initial learning rate of 1×10^{-3} and a weight decay of 1×10^{-4} to regulate model complexity. Furthermore, a dynamic learning rate adjustment strategy, specifically the ReduceLROnPlateau scheduler, was implemented: the learning rate was decayed by a factor of 0.5 if the validation loss failed to improve for five consecutive epochs (patience = 5). Table 1 summarizes the core parameters, including the optimizer settings, data augmentation protocol, and normalization steps.

4.1. Evaluation Metrics

Based on the quantitative plots shown in Figure 8 to 9, the CSWin-Semantic FPN demonstrates superior and consistent performance compared to the ResUNet, SWinUNet and UNet benchmarks across both training and validation sets. Regarding segmentation accuracy, the CSWin-Semantic FPN consistently maintains its leading position, with the Dice Score exceeding 0.8 and the IoU Score approaching 0.7 in the final epochs. The significant margin between the orange curve and the other two models indicates that the hybrid architecture - integrating Transformer and FPN - possesses a superior ability to capture semantic features and achieve better alignment with ground truth labels compared to purely convolutional architectures.

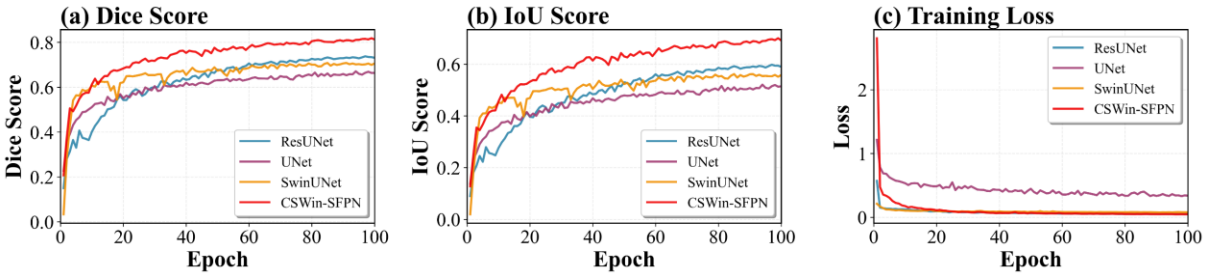


Figure 8. Convergence curves of the models on the training set.

In terms of convergence, the loss function curves reflect the robust learning capacity of the CSWin-Semantic FPN. Despite a higher initial loss value, the model undergoes a sharp decline within the first 20 epochs, subsequently maintaining stability and reaching the lowest error rate - comparable to or slightly better than ResUNet. Conversely, the UNet model exhibits lower efficiency, as its loss curve suffers from significant fluctuations and plateaus at a high level (approximately 0.5), indicating difficulties in weight optimization.

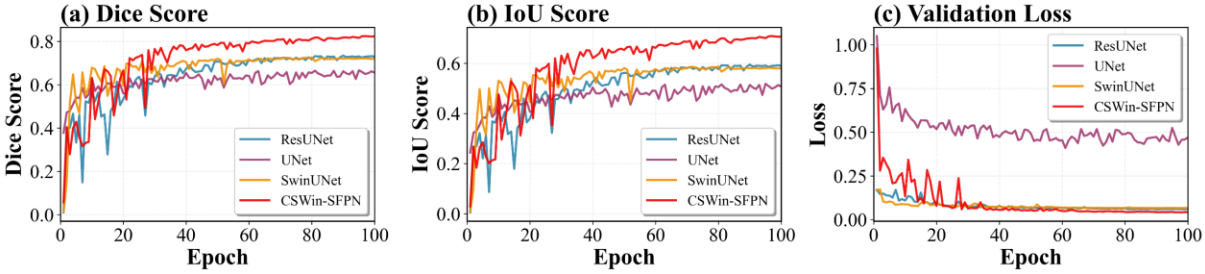


Figure 9. Convergence curves of the models on the validation set.

The high degree of consistency between the training and validation trends confirms that the CSWin-Semantic FPN possesses strong generalization capabilities, with no signs of overfitting. Achieving high accuracy alongside low and stable error rates demonstrates that the cross-window attention mechanism, coupled with the semantic feature pyramid, effectively addresses the segmentation task, yielding more reliable predictions in terms of both region identification and boundary precision compared to traditional methods.

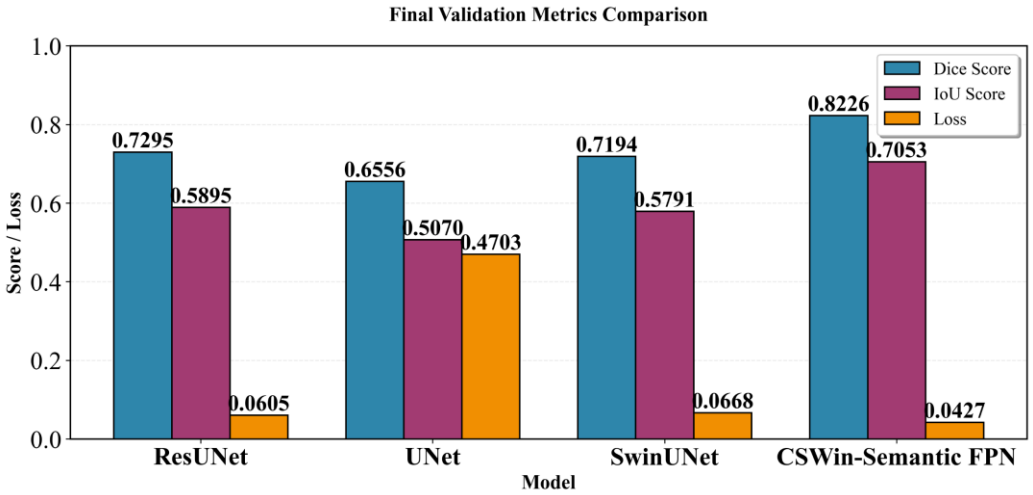


Figure 10. Final validation metric comparison across 4 models.

To provide a more comprehensive evaluation, Figure 10 presents a qualitative comparison of the segmentation results generated by the proposed CSWin-Semantic FPN against baseline models, including UNet, SWinUNet and ResUNet. The visual evidence clearly demonstrates that while traditional CNN-based architectures often suffer from fragmented predictions or fail to detect thin, low-contrast cracks, our model maintains superior structural continuity. By leveraging the cross-shaped window attention mechanism, the CSWin-Semantic FPN is able to capture the global context of crack topologies more effectively. These visual samples align with the quantitative metrics discussed previously, confirming the model's robustness in precisely delineating complex crack boundaries under diverse environmental conditions.

4.2. Quantitative Comparison

Beyond numerical evaluation, the practical effectiveness of the proposed framework is further scrutinized via visual comparisons under various crack conditions. As shown in Figures 11 to 13, the experimental outcomes reveal distinct improvements in segmentation accuracy when comparing CSWin-Semantic FPN against traditional baseline architectures.

Under simple background conditions, such as normal (Figure 13a) or dark (Figure 13c) surfaces, all three models are capable of detecting cracks. However, CSWin-Semantic FPN produces results with significantly sharper boundaries and a thickness that most closely approximates the ground truth. The performance disparity becomes particularly pronounced in the case of white backgrounds (Figure 13b), where the contrast between the crack and the surface is extremely low. While UNet and ResUNet generate broken and disconnected crack segments, CSWin-Semantic FPN maintains the continuity of the longitudinal crack. This demonstrates the superior sensitivity of the Transformer-based architecture in capturing weak semantic features.

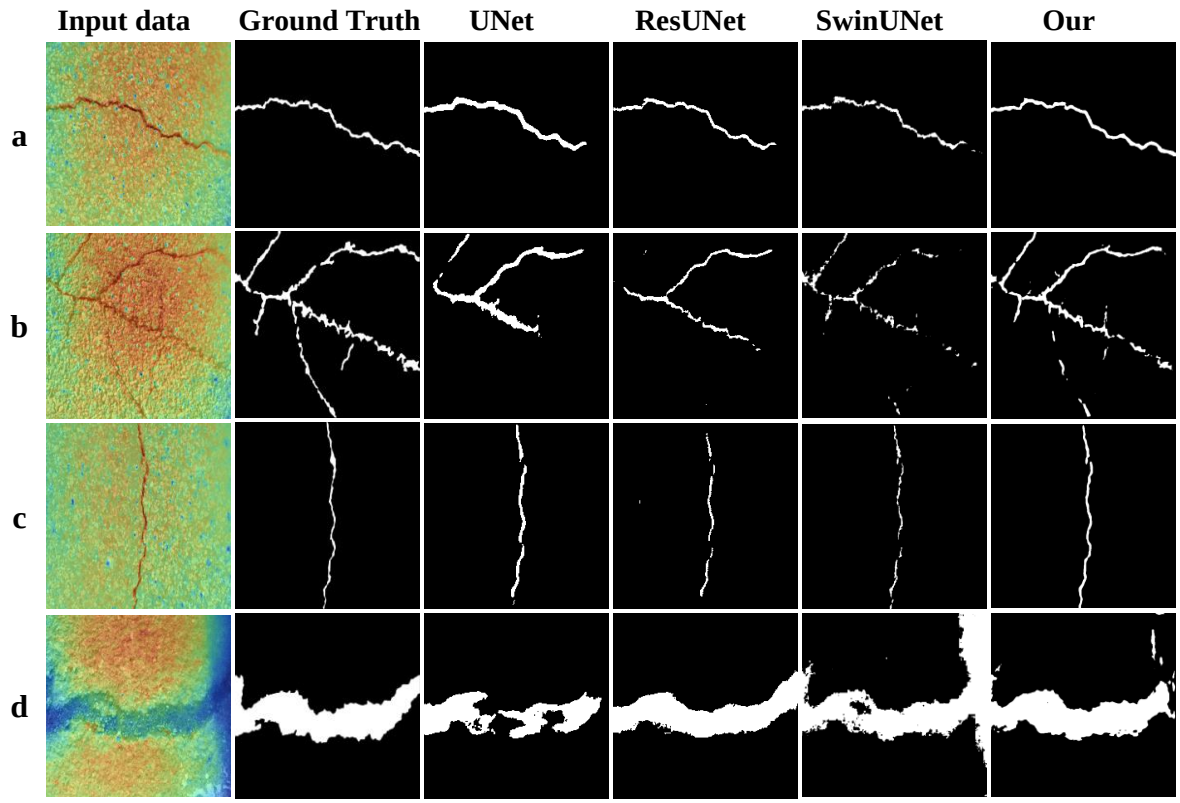


Figure 11. Prediction masks of the 4 models based on crack geometries: (a) single crack, (b) multi crack, (c) thin crack, and (d) thick crack.

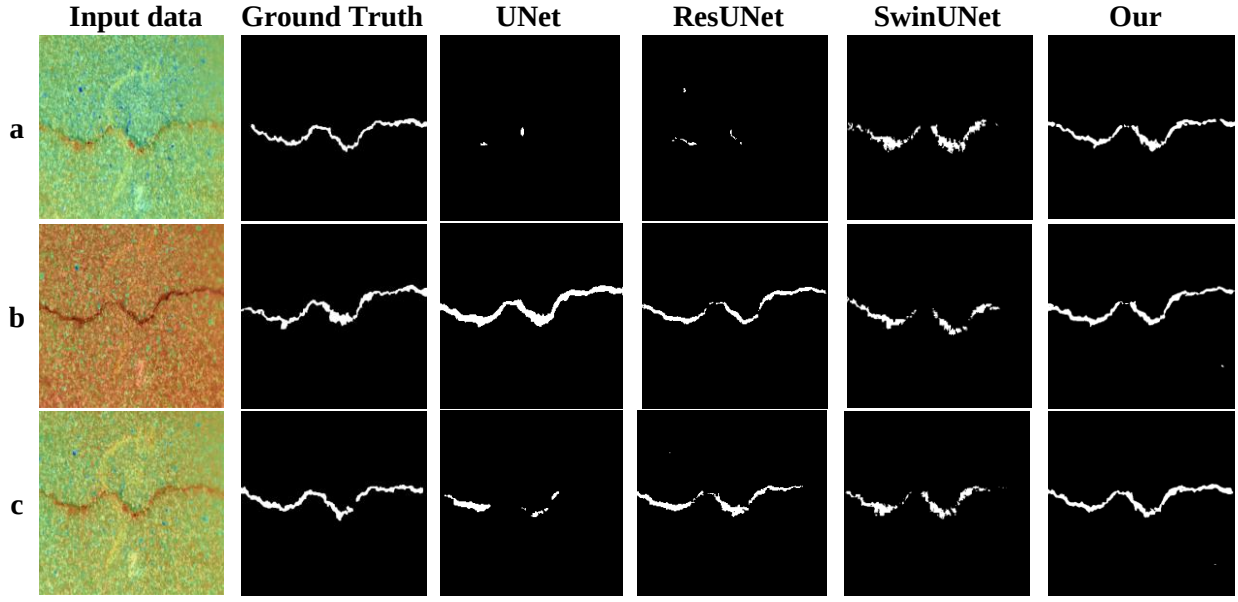


Figure 12. Prediction masks of the 4 models at different times of day: (a) morning, (b) noon, and (c) afternoon.

The advantages of the hybrid framework are further emphasized in noisy scenarios, such as the dusty surface in Figure 13d. This sample is plagued by white-painted markings and uneven textures, causing UNet to produce disorganized, ill-defined clusters, while ResUNet lacks the precision to maintain crack width consistency. Conversely, CSWin-Semantic FPN

demonstrates robust noise suppression, restoring the horizontal crack's geometry and scale in precise alignment with the ground truth.

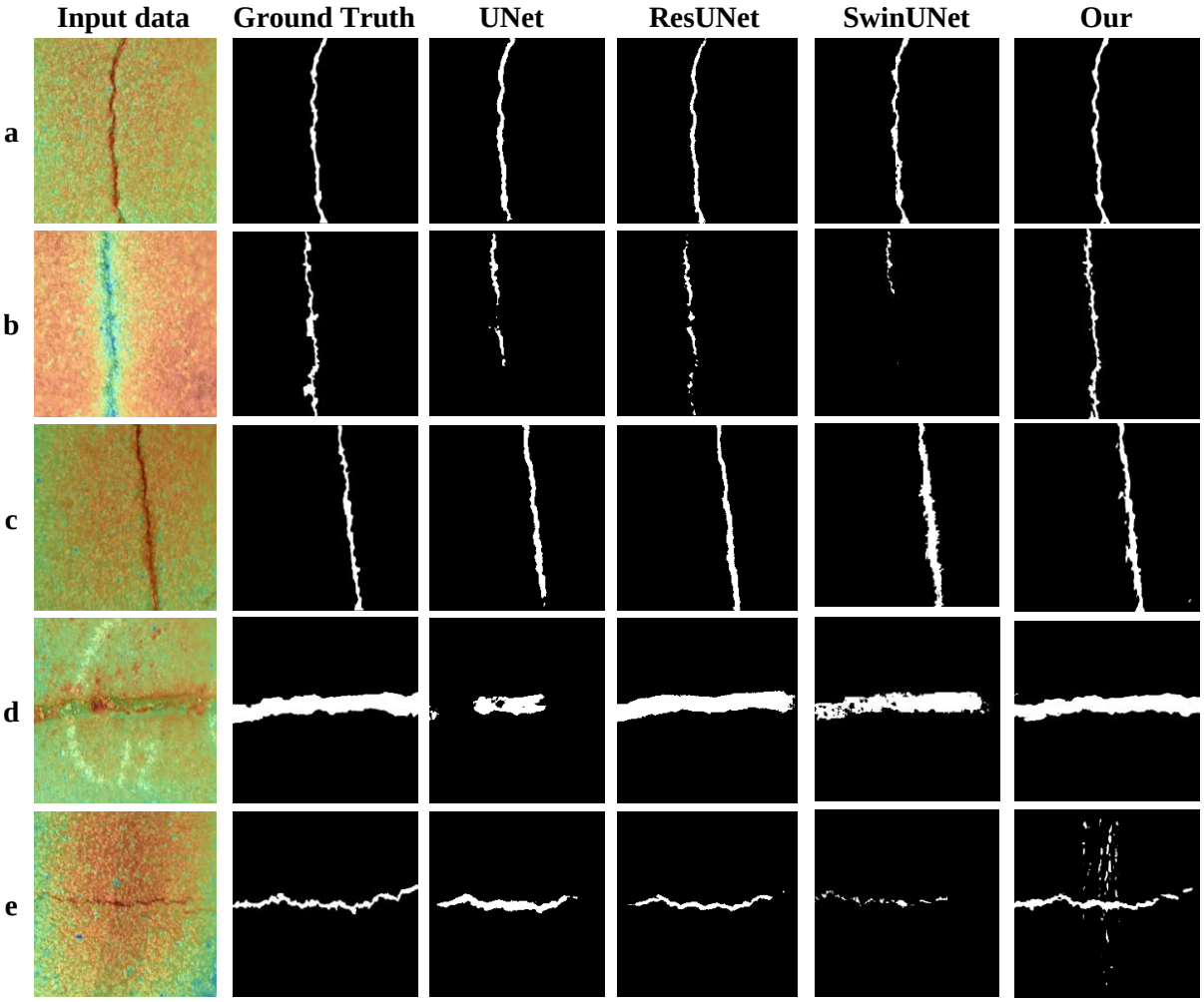


Figure 13. Prediction masks of the 4 models across different background categories: (a) normal, (b) white, (c) dark, (d) dusty, and (e) shadowed backgrounds.

5. CONCLUSION

This research proposes an advanced hybrid deep learning system designed for high-precision surface crack identification, overcoming the long-range dependency constraints of standard CNNs. By coupling the CSWin Transformer with Semantic FPN and employing a multi-modal fusion of RGB and thermal IR data, the model achieved a superior IoU of 70.53%, surpassing UNet, SwinUNet, and ResUNet benchmarks. The results validate that combining Transformer architectures with multi-modal data provides a reliable and automated paradigm for SHM under practical conditions. Future research will focus on expanding this framework to other structural surfaces, such as brick walls and cement concrete, as specialized fusion RGB-IR datasets for these materials become available. Additionally, exploring model compression techniques will be essential to facilitate deployment on edge platforms like drones and mobile devices.

ACKNOWLEDGMENT

This research is funded by University of Transport and Communications (UTC) under grant number T2026-CT-009TD

REFERENCES

[1]. W. Choi, Y.-J. Cha, SDDNet: Real-Time Crack Segmentation, IEEE Trans. Ind. Electron., 67 (2020) 8016-8025. https://doi.org/10.1109/TIE.2019.2945265
[2]. Y. Liu, J. Yao, X. Lu, R. Xie, L. Li, DeepCrack: A deep hierarchical feature learning architecture for crack segmentation, Neurocomputing, 338 (2019) 139-153. https://doi.org/10.1016/j.neucom.2019.01.036
[3]. H. Liu, J. Yang, X. Miao, C. Mertz, H. Kong, CrackFormer Network for Pavement Crack Segmentation, IEE Xplore, 24 (2023) 9240-9252. https://doi.org/10.1109/TITS.2023.3266776
[4]. Y. Yao, S.-T. E. Tung, B. Glisic, Crack detection and characterization techniques—An overview, Struct. Control Health Monit., 21 (2014) 1387-1413. https://doi.org/10.1002/stc.1655
[5]. N. C. Thi Nguyen, T. M. Vu, Damage detection in structural health monitoring using BiLSTM-1DCNN hybrid network: a case study on a large-scale steel truss bridge, Eng. Comput., 2 (2025) 2226-2242. https://doi.org/10.1108/EC-08-2024-0714
[6]. D. N. L. Minh, N. H. Xuan, T. V. Manh, B. N. K. Ngoc, Detection of damage in steel truss bridges using a hybrid 1DCNN-BIGRU model and time-series data augmentation techniques, Transp. Commun. Sci. J., 76 (2025) 1281-1295. https://doi.org/10.47869/tcsj.76.9.11
[7]. T.-V. Manh, H.-T. Ngoc, M.-T. Duc, L.-B. Phuc, L.-N. Duc, An Effective Damage Detection Approach for a Truss Bridge Using a Hybrid Deep Learning Model, Proceedings of the 5th International Conference on Sustainability in Civil Engineering, 2 (2025) 91-101. https://doi.org/10.1007/978-981-96-5206-8_10
[8]. A. Di Benedetto, M. Fiani, L. M. Gujski, U-Net-Based CNN Architecture for Road Crack Segmentation, Infrastructures, 8 (2023) 90. https://doi.org/10.3390/infrastructures8050090
[9]. X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, B. Guo, CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), (2022) 12114-12124. https://doi.org/10.1109/CVPR52688.2022.01181
[10].J. He, J. Wang, Z. Han, B. Li, M. Lv, Y. Shi, Cancer detection for small-size and ambiguous tumors based on semantic FPN and transformer, PLOS ONE, 18 (2023) e0275194. https://doi.org/10.1371/journal.pone.0275194
[11].F. Liu, J. Liu, L. Wang, Asphalt Pavement Crack Detection Based on Convolutional Neural Network and Infrared Thermography, IEEE Trans. Intell. Transp. Syst., 23 (2022) 22145-22155. https://doi.org/10.1109/IWAGPR65621.2025.11109044
[12].F. Liu, J. Liu, L. Wang, I. L. Al-Qadi, Multiple-type distress detection in asphalt concrete pavement using infrared thermography and deep learning, Autom. Constr., 161 (2024) 105355. https://doi.org/10.1016/j.autcon.2024.105355
[13].X. Liu, P. Gao, T. Yu, F. Wang, R.-Y. Yuan, CSWin-UNet: Transformer UNet with cross-shaped windows for medical image segmentation, Inf. Fusion, 113 (2025) 102634. https://doi.org/10.1016/j.inffus.2024.102634
[14].T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature Pyramid Networks for Object Detection, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2017) 936-944. https://doi.org/10.1109/CVPR.2017.106
[15].O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015, (2015) 234-241. https://doi.org/10.1007/978-3-319-24574-4_2
[16].F. I. Diakogiannis, F. Waldner, P. Caccetta, C. Wu, ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data, ISPRS J. Photogramm. Remote Sens., 162 (2020) 94-114. https://doi.org/10.1016/j.isprsjprs.2020.01.013