



Fine-Grained Pavement Crack Segmentation Using Dual-Attention Residual U-Net with ASPP and Boundary Supervision

Oktavian Andi Cahya Nugraha^{1,*}, Guruh Fajar Shidik^{2,*}, Catur Supriyanto³

^{1,2,3}Department of Computer Science, Universitas Dian Nuswantoro, Indonesia

Abstract.

Purpose: Fine-grained pavement crack segmentation remains challenging because cracks generally occupy a small proportion of pavement images and exhibit thin, irregular, and low-contrast structures. These characteristics create severe pixel-level class imbalance and may reduce the stability of conventional U-Net predictions. This study aims to improve crack localization and segmentation consistency through an enhanced U-Net architecture.

Methods: This study proposes a pavement-specific Dual-Attention Residual U-Net integrating compact ASPP, strip pooling, and auxiliary boundary-aware supervision for thin and elongated crack structures. The model is evaluated using Crack500, CrackForest, and DeepCrack images through comparison with a U-Net baseline, repeated evaluation runs of the proposed model, and threshold-sensitivity analysis. A supporting LOCO ablation study examines the effects of component removal using seed 42.

Results: The proposed model achieved a Dice coefficient of 0.7967 and an IoU of 0.6621 in the seed-42 evaluation. Across seeds 42, 123, and 2024, the reported mean Dice was 0.7963 ± 0.0004 and the mean IoU was 0.6616 ± 0.0006 , indicating low observed variation across the evaluated runs. The component ablation provides supporting evidence about the design choices within the complete architecture.

Novelty: The architectural contribution is the pavement-specific integration of compact ASPP with strip pooling and auxiliary boundary-aware supervision within a residual dual-attention U-Net. This design combines multiscale context aggregation and elongated spatial context with feature refinement and supervision intended to preserve fine crack structures.

Keywords: Pavement crack segmentation, Dual-attention, U-Net, ASPP, Boundary supervision

Received August 2026 / **Revised** October 2026 / **Accepted** October 2026

This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).



INTRODUCTION

Pavement surface cracks are early indicators of road deterioration and may affect structural performance, maintenance cost, and traffic safety [1], [2]. Manual crack inspection is generally time-consuming, subjective, and inconsistent, especially when road networks are large and pavement conditions vary [1], [3], [4]. Therefore, automated crack segmentation using computer vision and deep learning is an important research direction for pavement monitoring systems [1], [5].

Deep learning-based semantic segmentation has become a common approach for crack detection because it can learn discriminative visual features directly from pavement images [4], [6]. U-Net is widely used in segmentation tasks due to its encoder-decoder structure and skip connections, which help preserve spatial information [7]. However, pavement cracks are often thin, irregular, and low contrast [8]–[10]. In addition, crack pixels are substantially outnumbered by background pixels, resulting in severe pixel-level class imbalance [6], [11], [12]. These characteristics can lead to unstable predictions, false positives, false negatives, and even prediction collapse where the network mostly predicts the background class [6], [13].

Recent pavement crack studies have improved U-Net-based models using residual learning, skip-connection refinement, and loss-function strategies [1], [3], [12]. The baseline paper by Ali et al. proposed Residual Sharp U-Net (RS-Net) to reduce feature mismatch between encoder and decoder representations and reported promising pavement crack segmentation performance. Nevertheless, there is still room to investigate alternative architectural designs that explicitly combine residual learning, attention-based feature refinement, and multi-scale context modeling for fine-grained cracks [9], [10], [14].

^{12*}Corresponding author.

Email addresses: vianandi60@gmail.com (Nugraha)

DOI: [10.15294/sji.v13i4.62242](https://doi.org/10.15294/sji.v13i4.62242)

To address this gap, this study proposes a Dual-Attention Residual U-Net with ASPP. The model is designed to preserve fine crack details through residual connections, emphasize informative crack regions through spatial and channel attention [9], [15], and capture cracks with different scales using ASPP [16]. The main contributions of this study are: (1) a pavement-specific residual dual-attention U-Net that redesigns multiscale context aggregation using compact ASPP with strip pooling for thin and elongated crack structures; (2) the integration of attention-gated skip connections, deep supervision, and boundary supervision to improve fine-grained crack localization; and (3) an empirical assessment of the proposed architecture through a controlled U-Net baseline comparison, component ablation, repeated runs, and threshold-sensitivity analysis to examine its segmentation performance and the contribution of its design choices.

Road pavement crack segmentation is often hindered by severe class imbalance, which can trigger prediction collapse, a condition in which the network tends to classify most pixels as pure background. As a state-of-the-art baseline, [1] proposed Residual Sharp U-Net (RS-Net), which employs residual blocks and Laplacian sharpening filters in the skip connections. At the global level, RS-Net achieves high accuracy (0.9892). The use of Laplacian sharpening motivates an examination of how high-frequency pavement textures influence feature refinement. However, aggregate overlap metrics alone cannot establish whether segmentation errors are caused by the sharpening operation. This study therefore investigates attention-based feature refinement as an alternative design choice. In contrast, the DRA-UNet framework proposed by [17] employs similar components, including Dual Attention, Residual Networks, and Atrous Spatial Pyramid Pooling (ASPP), to extract mining-induced cracks from high-resolution imagery. The present study investigates a compact ASPP configuration with strip pooling for pavement crack segmentation. This design is motivated by the spatial resolution of the bottleneck features and the elongated geometry of pavement cracks. These considerations provide a rationale for the proposed design, but do not establish an empirical performance disadvantage for DRA-UNet under the current evaluation protocol. Furthermore, square pooling windows inherently absorb excessive background noise around thin crack structures.

DRA-UNet is included as an architectural reference because it also combines residual learning, dual attention, and ASPP. Its published results were obtained on the GCCMA-UAV dataset for mining-induced ground cracks, which differs from the pavement crack datasets used in this study. DRA-UNet was not retrained under the present experimental protocol; therefore, its published scores are not included in the primary quantitative comparison. This limits the conclusions that can be drawn about its performance relative to the proposed model. The architectural discussion provides design motivation and does not establish superiority over DRA-UNet under matched evaluation conditions.

The proposed model combines residual learning, dual attention, compact ASPP, and strip pooling within a U-Net architecture. Compact multiscale context aggregation and elongated pooling are intended to support the representation of thin pavement cracks. Attention-gated skip connections, deep supervision, and boundary supervision provide additional feature-refinement and training mechanisms. The LOCO study is designed to examine the effect of each component within the complete architecture.

Although DRA-UNet also employs residual blocks, dual attention, and ASPP, the novelty of the proposed architecture does not lie in these components individually, but in their adaptation to fine-grained pavement crack segmentation. DRA-UNet was originally developed for macro-scale mining ground cracks from UAV imagery and employs ASPP dilation rates of 12, 24, and 36. In contrast, the proposed model uses a compact ASPP configuration with dilation rates of 1, 3, 6, and 9, integrated with strip pooling to capture elongated crack continuity while limiting excessive aggregation of surrounding pavement texture. In addition, attention-gated skip connections, deep supervision, and boundary supervision are incorporated to preserve thin crack structures and improve boundary localization. Therefore, the main architectural novelty is the pavement-specific integration of compact multiscale context modeling and boundary-aware auxiliary supervision within a residual dual-attention U-Net framework.

Tabel 1 Architectural comparison of related methods

Method	Residual Block	Dual Attention	ASPP	Strip Pooling	Attention Gate	Deep Supervision	Edge Supervision	Design Consideration
RS-Net	Yes	No	No	No	No	No	No	Sharpened connections skip
DRA-UNet	Yes	Yes	Yes	No	No	No	No	Mining-ground crack delineation from UAV imagery
Proposed	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Compact context aggregation and auxiliary supervision for pavement cracks

METHODS

This study adopts an experimental approach to evaluate the effectiveness of the proposed model for pavement crack segmentation. The methodology consists of several stages, including dataset preparation, baseline model implementation, proposed model design, and performance evaluation. A standard U-Net is used as the baseline to provide a fair comparison, while the proposed architecture integrates dual-attention, residual connections, and ASPP to improve segmentation performance. The overall framework of the study is designed to analyze both quantitative performance and model behavior.

Dataset and Data Preparation

This study employed the Crack500, CrackForest, and DeepCrack datasets. Each dataset was partitioned separately into training, validation, and test subsets using an approximately 70:15:15 ratio, after which the corresponding subsets were combined across sources. Partitioning preceded augmentation, which was applied only to training data. For DeepCrack, the original training and test partitions (300 and 237 image-mask pairs, respectively) were combined before repartitioning. The reported DeepCrack scores refer to the resulting internal test subset of this study. The study included 1,774 image-mask pairs: 1,119 from Crack500, 118 from CrackForest, and 537 from DeepCrack. The train/validation/test allocations were 783/170/166 for Crack500, 83/17/18 for CrackForest, and 376/80/81 for DeepCrack. The combined subsets contained 1,242 training, 267 validation, and 265 test pairs. No additional negative-only dataset was used in the final experiments.

All images were converted to RGB format and resized to 384×384 pixels, while the corresponding masks were converted to grayscale, resized to the same spatial resolution, and binarized using a threshold of 0.5. Image tensors were represented in the $[0, 1]$ range. During training, synchronized augmentations were applied to the images and masks, including random horizontal and vertical flipping, random rotation within -10° to 10° , and mild brightness and contrast adjustments. No augmentation was applied to the validation and test data. Ground truth annotations are converted into binary masks. The datasets are divided into training, validation, and testing sets using a consistent split ratio to ensure fair evaluation across model [18], [19].



Figure 1. Crack500 image



Figure 2. Crack500 masks



Figure 3. CrackForest image



Figure 4. CrackForest masks



Figure 5. DeepCrack image



Figure 6. DeepCrack masks

Baseline and Proposed Architecture

The baseline model is a standard U-Net, selected as a widely used encoder-decoder architecture for image segmentation [20], [21]. The proposed Dual-Attention Residual U-Net extends this architecture with residual connections, dual attention, attention-gated skip connections, compact ASPP, and strip pooling. Deep supervision and boundary supervision provide additional training objectives. Together, these components form a pavement-specific architecture designed for fine-grained crack segmentation. Residual blocks facilitate gradient propagation and are intended to support stable training [1], [20]. Residual connections also support feature propagation across network layers [22]. The dual-attention module combines spatial and channel attention [9]. Spatial attention emphasizes relevant image regions, while channel attention recalibrates feature responses according to their relevance [15], [23]. These mechanisms are intended to emphasize crack-related features and reduce responses to irrelevant pavement textures. Compact ASPP is positioned at the bottleneck to aggregate multiscale contextual information through parallel dilated convolutions [8], [10], [24]. Strip pooling complements this aggregation by modeling elongated spatial context associated with crack structures. Attention-gated skip connections refine encoder features before fusion with decoder features, supporting the integration of spatial detail and contextual information.

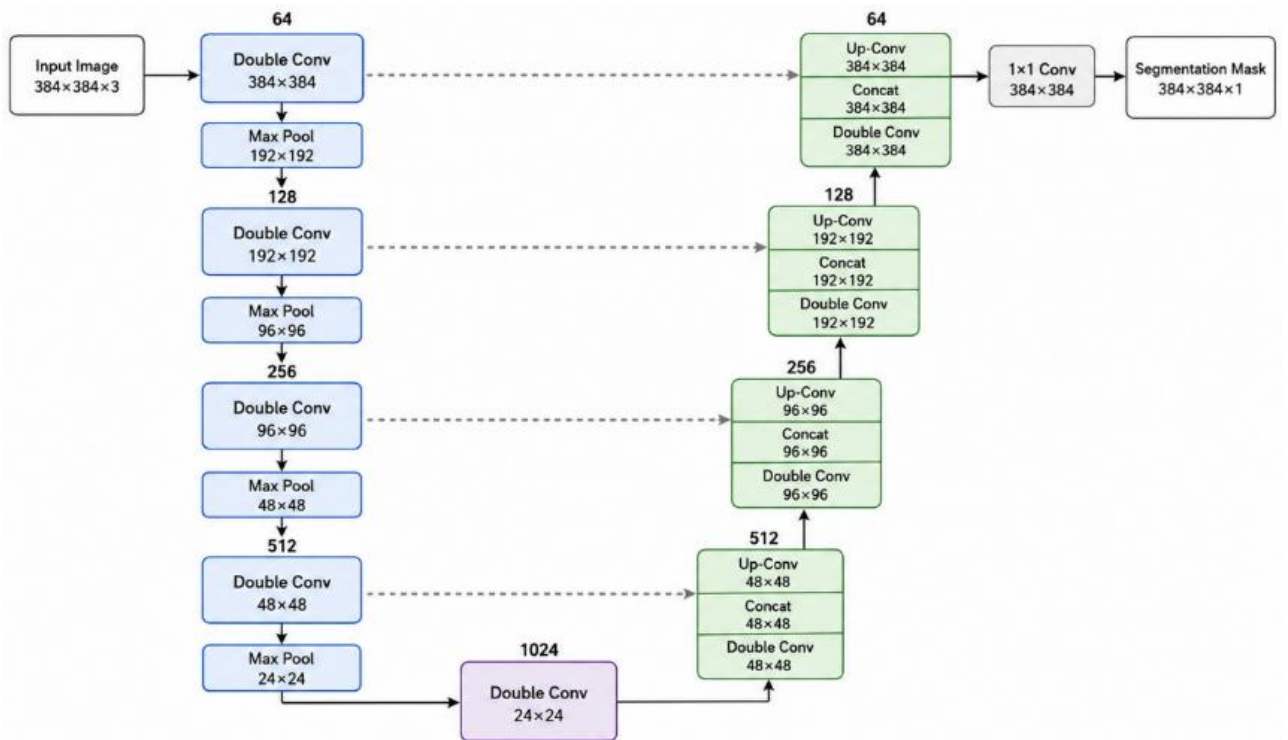


Figure 7. U-Net architecture

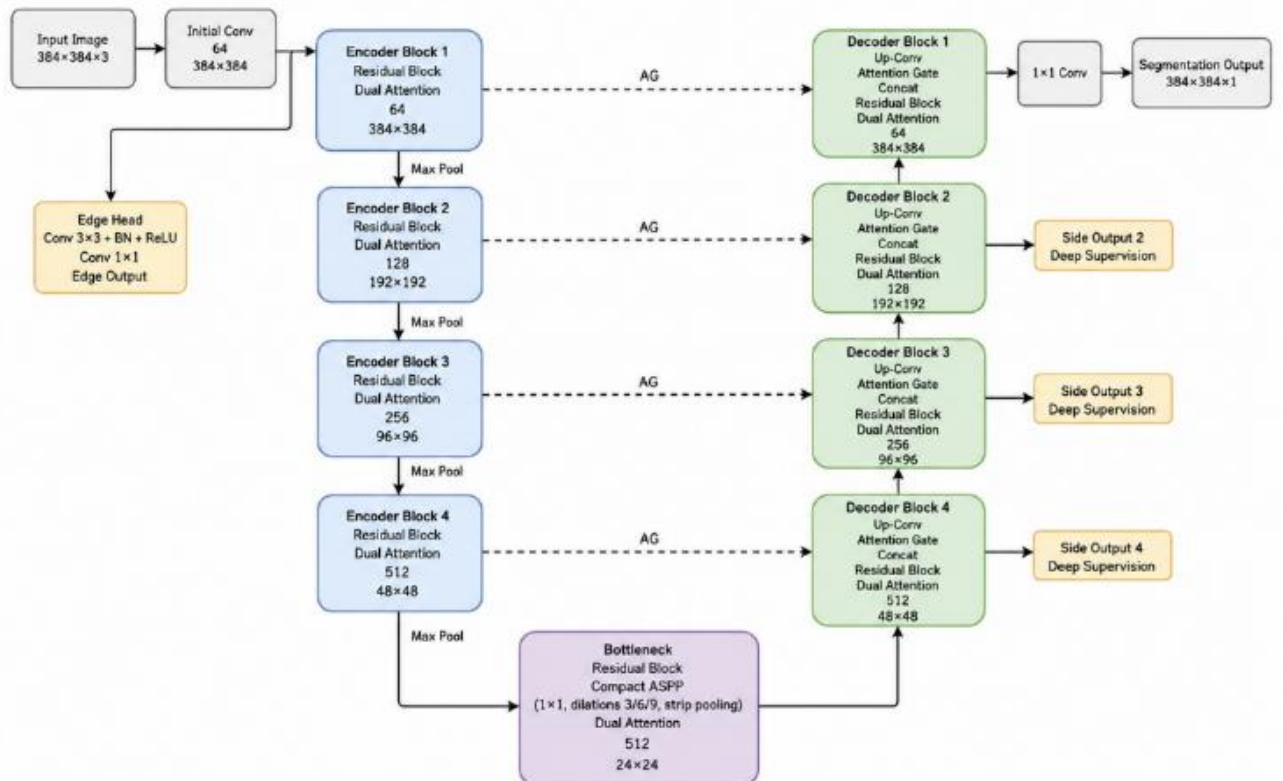


Figure 8. Proposed architecture

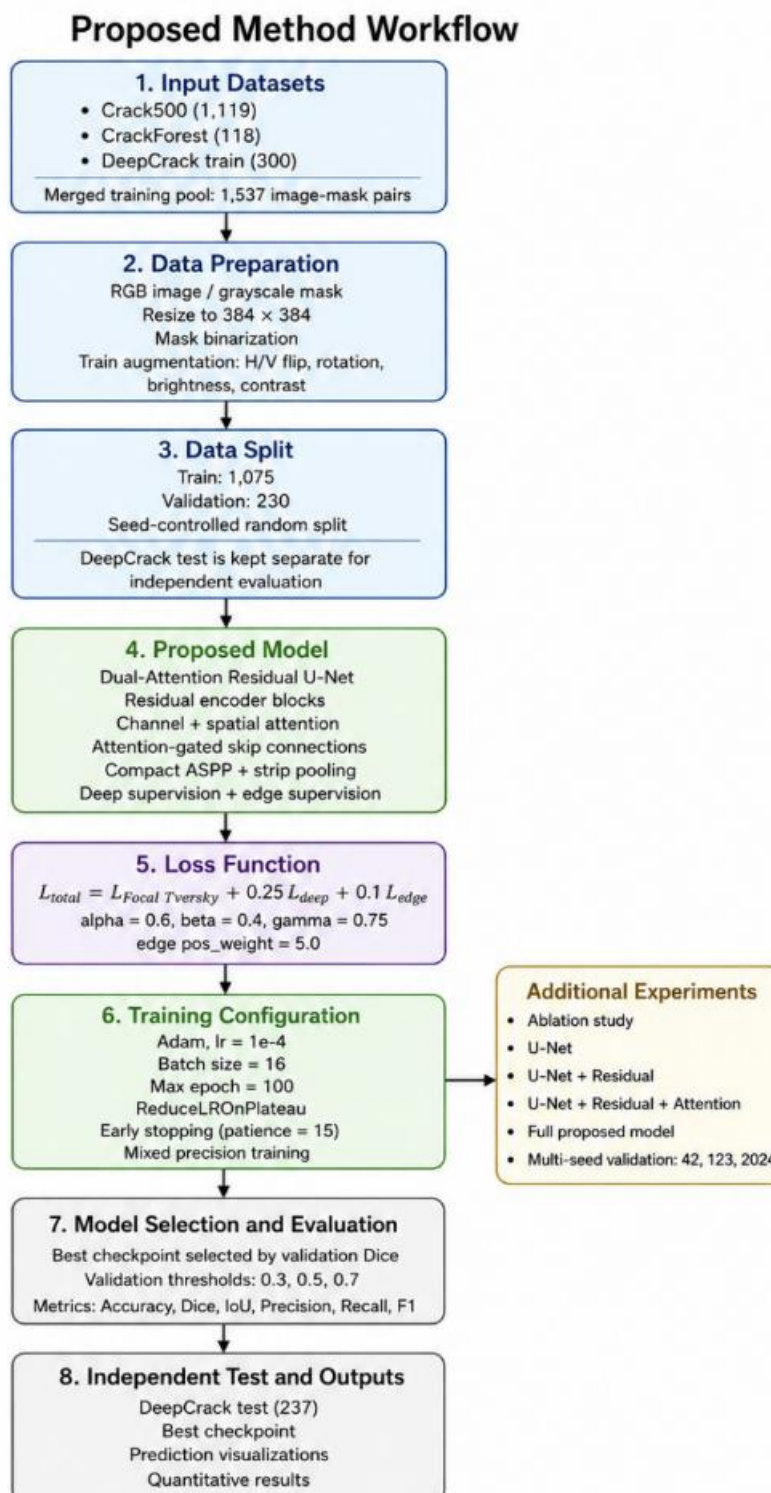


Figure 9. Overall workflow of the proposed pavement crack segmentation method

Loss Function

The proposed model was optimized using a composite loss consisting of Focal Tversky Loss, deep-supervision loss, and edge loss. The primary Focal Tversky Loss was configured with $\alpha = 0.6$, $\beta = 0.4$, and $\gamma = 0.75$. Three auxiliary segmentation outputs were weighted by 0.5, 0.3, and 0.2, and their combined contribution was scaled by 0.25. Boundary targets were generated by subtracting an eroded ground-truth mask from its dilated counterpart using a 3×3 kernel. The edge branch was optimized using

BCEWithLogitsLoss with a positive-class weight of 5.0 and an overall boundary-loss weight of 0.1. Therefore, the total objective was formulated as

$$L_{total} = L_{FT} + 0.25L_{deep} + 0.1L_{edge} \quad (1)$$

Equation (1) describes the training objective of the complete proposed model. Focal Tversky loss was used as the primary segmentation criterion for all configurations. The U-Net baseline used this criterion alone. The variant without deep supervision omitted the deep-supervision loss term, while the variant without edge supervision omitted the boundary-loss term. The remaining ablation configurations retained both auxiliary terms with the weights specified in Eq. (1). The evaluation loss reported in Table 4 included the primary and active auxiliary terms for each configuration. Consequently, these loss values reflect different objectives and are not directly comparable as measures of segmentation quality. Segmentation performance was compared using Dice, IoU, precision, recall, and F1.

Training and Configuration

All models were trained using the same experimental configuration to ensure fair comparison across all ablation settings. The experiments were conducted using the Adam optimizer with an initial learning rate of 0.0001 and a batch size of 16. Each model was trained for a maximum of 100 epochs using an early stopping strategy to prevent overfitting. Early stopping was configured with a patience value of 15 epochs and a minimum delta threshold of 0.0005.

The proposed segmentation framework used Focal Tversky loss as its primary segmentation criterion to address severe class imbalance [25], [26]. The Tversky parameters were configured with $\alpha = 0.6$, $\beta = 0.4$, and $\gamma = 0.75$. Boundary supervision was additionally incorporated with a boundary weight of 0.1 and edge positive weight of 5.0 to enhance fine-grained crack localization [27], [28].

The input images were resized to 384×384 pixels before training. The dataset was divided into training, validation, and testing sets using a ratio of 70%, 15%, and 15%, respectively. The proposed model and the U-Net baseline were evaluated using three random seeds: 42, 123, and 2024. Performance across runs was summarized using the mean and sample standard deviation, calculated with a denominator of $n - 1$, where $n = 3$. These summaries describe the observed variation across runs and do not, by themselves, establish statistical significance. For each random seed, the checkpoint with the highest validation Dice score was retained. After completing the three independent runs, the representative final model was selected exclusively based on validation Dice. Among the evaluated seeds, seed 42 achieved the highest validation Dice and was therefore selected for the final evaluation. The test set was not used during model selection. In addition, the mean and standard deviation across the three seeds were reported to assess the robustness of the proposed architecture to random initialization. Model performance was evaluated using Evaluation Loss, Accuracy, Dice Coefficient, Intersection over Union (IoU), Precision, Recall, and F1-score. The experiments were implemented using the PyTorch deep learning framework with CUDA acceleration enabled for GPU computation.

Computational Profiling

Computational profiling used 384×384 RGB inputs, a batch size of 1, and float32 precision on a Tesla T4 GPU. The recorded profiling protocol included five warm-up iterations and twenty repetitions, with auxiliary heads disabled in the profiling configuration. FLOPs were estimated for the convolution and matrix-multiplication operations supported by the PyTorch profiler, counting one multiplication and one addition as two operations. Unsupported operations were excluded; the reported values therefore represent estimates for supported operations rather than the total operation count.

Leave-One-Component-Out Ablation Protocol

The LOCO study compares the complete proposed model with seven variants, each omitting one component: residual connections, dual attention, attention-gated skip connections, ASPP, strip pooling, deep supervision, or boundary supervision. The removals are independent, with each variant defined relative to the complete model. A U-Net baseline provides an additional reference for the overall architecture comparison.

The LOCO study comprises nine configurations evaluated using seed 42. The reported component comparisons are based on these single-seed results. The effect of a component is interpreted as the change in segmentation performance when that component is removed from the complete model. Such an effect is conditional on the remaining components and the training protocol; it does not represent an isolated or necessarily additive contribution.

Evaluation Metrics

Accuracy measures the proportion of correctly classified pixels. Dice coefficient and IoU measure spatial overlap between predicted masks and ground-truth masks, making them more informative for class-imbalanced segmentation tasks. Precision evaluates how many predicted crack pixels are correct, while Recall evaluates how many ground-truth crack pixels are detected. F1-score summarizes Precision and Recall balance.

$$Dice = \frac{2TP}{2TP + FP + FN} \quad (2)$$

Measures the degree of overlap between the segmentation results and the ground truth. This metric is widely used in image segmentation tasks because it is sensitive to the quality of the segmented object area [7].

$$IoU = \frac{TP}{TP + FP + FN} \quad (3)$$

Measures the ratio between the intersection area and union area of the prediction and ground truth. The higher the IoU value, the better the quality of the resulting segmentation [6].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Measures the proportion of all pixels that are correctly classified, both crack and background pixels.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Measures the accuracy of the model in detecting cracks, namely the percentage of pixels predicted as cracks and actually being cracks [7].

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

Measures the model's ability to find all crack areas present in the ground truth. A high recall value indicates few cracks were missed.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (7)$$

The harmonic mean between precision and recall that provides a balance between accuracy and completeness of crack detection.

RESULT AND DISCUSSION

Table 2. Proposed results

Seed	Loss	Accuracy	Dice	IoU	Precision	Recall	F1
42	0.2217	0.9880	0.7967	0.6621	0.7721	0.8228	0.7967
123	0.2241	0.9877	0.7959	0.6610	0.7705	0.8232	0.7959
2024	0.2232	0.9872	0.7964	0.6617	0.7716	0.8225	0.7963
Mean +/- std	0.2230 +/- 0.0012	0.9876 +/- 0.0004	0.7963 +/- 0.0004	0.6616 +/- 0.0006	0.7714 +/- 0.0008	0.8228 +/- 0.0004	0.7963 +/- 0.0004

Table 3. Baseline results

Seed	Loss	Accuracy	Dice	IoU	Precision	Recall	F1
42	0.6859	0.9077	0.5293	0.4733	0.6112	0.6973	0.5293

123	0.6791		0.8972		0.5145		0.4802		0.6014		0.7014		0.5264
2024	0.6845		0.9013		0.5201		0.4724		0.6106		0.6987		0.5307
Mean +/- std	0.6832	+/-	0.9021	+/-	0.5213	+/-	0.4753	+/-	0.6077	+/-	0.6991	+/-	0.5288
	0.0036		0.0053		0.0075		0.0043		0.0055		0.0021		0.0022

Tables 2 and 3 summarize the reported performance of the proposed model and the U-Net baseline across seeds 42, 123, and 2024. The baseline achieved a mean Dice score of 0.5213 ± 0.0075 and a mean IoU of 0.4753 ± 0.0043 . Seed 42 produced its highest Dice score (0.5293), whereas seed 123 produced its highest IoU (0.4802). The proposed model achieved a mean Dice score of 0.7963 ± 0.0004 and a mean IoU of 0.6616 ± 0.0006 . The reported runs therefore show higher overlap scores and smaller observed variation for the proposed model. Across the three evaluated runs, the proposed model achieved higher mean Dice and IoU scores and smaller observed variation than the U-Net baseline [13], [29].

Ablation Study

Table 4. LOCO ablation results using seed 42 at a prediction threshold of 0.5

Method	Loss	Accuracy	Dice	IoU	Precision	Recall	F1
Proposed Method	0.2217	0.9880	0.7967	0.6621	0.7721	0.8228	0.7967
Without Residual	0.2335	0.9866	0.7877	0.6486	0.7612	0.7826	0.7877
Without Attention	0.2346	0.9872	0.7865	0.6481	0.7525	0.8236	0.7865
Without Skip Attention	0.2540	0.9870	0.7737	0.6309	0.7658	0.7818	0.7737
Without ASPP	0.2313	0.9863	0.7944	0.6590	0.7955	0.7933	0.7944
Without Strip Pooling	0.2386	0.9747	0.7721	0.6366	0.7889	0.7399	0.7721
Without Deep Supervision	0.2039	0.9856	0.7627	0.6164	0.7180	0.8133	0.7627
Without Edge Supervision	0.2344	0.9875	0.7765	0.6347	0.7926	0.7610	0.7765
Baseline	0.6859	0.9077	0.5293	0.4733	0.6112	0.6973	0.5293

Table 4 presents the leave-one-component-out (LOCO) ablation results. Each variant removes one component from the complete proposed model while retaining the remaining components. This design evaluates the contribution of each component within the complete architecture, rather than the incremental effect of adding components to a baseline. The complete model achieved a Dice score of 0.7967 and an IoU of 0.6621. Removing deep supervision reduced Dice to 0.7627 and IoU to 0.6164. The variant without deep supervision reported a lower loss than the complete model (0.2039 versus 0.2217), but also a lower Dice score (0.7627 versus 0.7967). Its evaluation loss excluded the deep-supervision term, so the lower value should not be interpreted as improved segmentation performance. Removing attention-gated skip connections reduced Dice to 0.7737 and IoU to 0.6309, while removing boundary supervision yielded Dice and IoU scores of 0.7765 and 0.6347. These results suggest that auxiliary supervision and skip-feature refinement contribute to segmentation performance in the evaluated configuration. Removing dual attention yielded a Dice score of 0.7865 and an IoU of 0.6481. Removing residual connections yielded a Dice score of 0.7877 and an IoU of 0.6486. Removing strip pooling yielded scores of 0.7721 and 0.6366, respectively. These observations describe the effects of the two removals within the complete architecture under the evaluated seed-42 protocol. In comparison, removing ASPP produced a smaller change, with Dice and IoU scores of 0.7944 and 0.6590, respectively [14], [16].

These differences describe the observed effects of component removal under the evaluated training and testing protocol. They are conditional on the components retained in each variant and should not be

interpreted as independent or additive contributions. As the results reported here are from seed 42, they do not establish consistency across random initializations or statistical significance.

Table 5. Result test dataset

Dataset	Accuracy	Dice	IoU	Precision	Recall	F1-Score
Crack500	0.9721	0.7997	0.6663	0.8007	0.8031	0.7997
CrackForest	0.9892	0.6586	0.4910	0.7353	0.7307	0.6586
DeepCrack	0.9910	0.8870	0.7970	0.8707	0.9039	0.8870

Table 5 presents the proposed model's test performance across the three datasets. DeepCrack achieved the highest Dice score (0.8870) and IoU (0.7970), followed by Crack500 with scores of 0.7997 and 0.6663, respectively. CrackForest yielded lower overlap scores, with Dice of 0.6586 and IoU of 0.4910. Its recall of 0.7307 was also lower than that of Crack500 (0.8031) and DeepCrack (0.9039), indicating a larger proportion of missed annotated crack pixels. Despite its high accuracy (0.9892), CrackForest showed weaker segmentation overlap, illustrating that accuracy alone does not fully describe crack segmentation quality. The evaluation does not isolate the causes of these dataset differences, which require further investigation.

Qualitative Segmentation Results

Figure 10 presents qualitative segmentation examples, including the original image, ground truth, prediction, overlay, and error map. The examples illustrate crack localization and residual errors around thin, low-contrast, and boundary regions. Selected examples do not establish the absence of all-background predictions across the evaluation set.

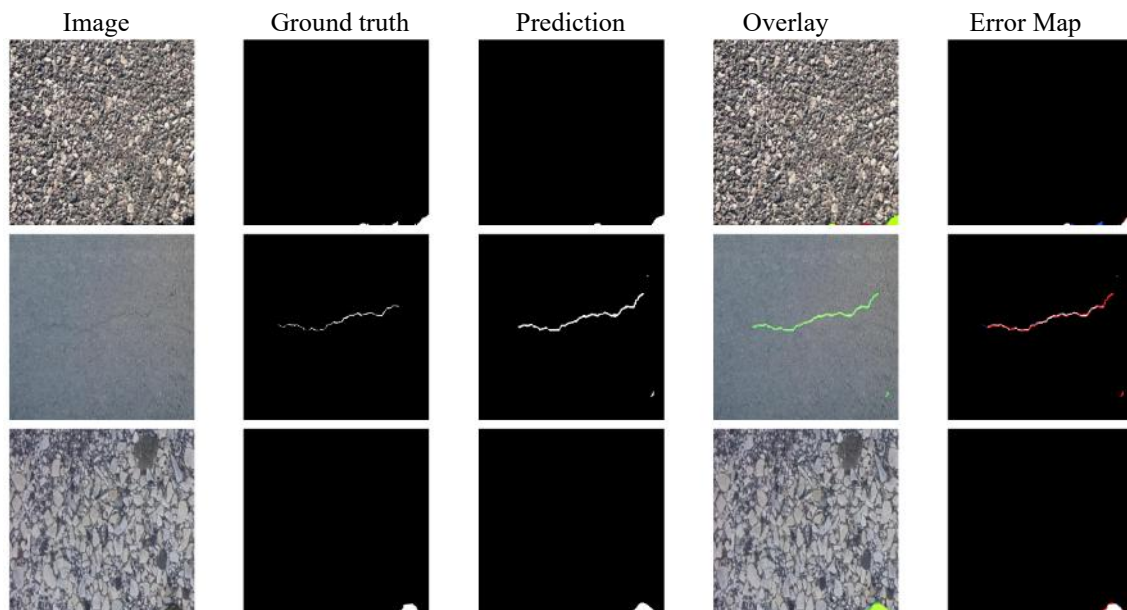


Figure 10. Segmentation results

Table 6. State of the art

Model	Dice	Loss	IoU	Accuracy	Precision	Recall	F1
U-Net	0.5293	0.6859	0.4733	0.9077	0.6112	0.6973	0.5293
RS-Net	0.6331	0.0644	0.5409	0.9892	0.7883	0.6486	NA
Proposed	0.7967	0.2217	0.6621	0.9880	0.7721	0.8228	0.7967

RS-Net values were taken from the original publication and were not obtained by retraining or evaluating RS-Net on the present study's data split. They provide cross-study context and do not establish superiority under matched evaluation conditions. NA indicates that the F1-score was not reported in the original publication. DRA-UNet is discussed as an architectural reference for mining-induced ground crack delineation using UAV imagery. Because its published evaluation concerns a different dataset domain and it was not retrained under the present protocol, its scores are not included in Table 6. No claim of superiority over DRA-UNet under matched evaluation conditions is made.

Threshold Sensitivity Analysis Proposed Method

Table 7. Threshold sensitivity analysis

Threshold	Dice	IoU	Accuracy	Precision	Recall	F1
0.3	0.7964	0.6617	0.9880	0.7618	0.8259	0.7964
0.5	0.7967	0.6621	0.9880	0.7721	0.8228	0.7967
0.7	0.7969	0.6624	0.9881	0.7752	0.8198	0.7969

Table 7 shows that the proposed model obtained similar overlap scores across thresholds of 0.3, 0.5, and 0.7. Dice increased slightly from 0.7964 to 0.7969, while IoU increased from 0.6617 to 0.6624. Precision increased while recall decreased as the threshold increased. Threshold 0.7 produced the highest Dice and IoU among the tested values, although the differences were small. These observations describe threshold sensitivity and do not establish statistical significance or an optimal threshold for other evaluation conditions. A fixed probability threshold of 0.5 was specified before test evaluation and applied consistently to the compared configurations. This common decision rule was used to evaluate the proposed method without selecting the threshold from test-set performance. Additional evaluations at thresholds of 0.3 and 0.7 were conducted to describe threshold sensitivity rather than to select the threshold for the primary comparison.

Prediction Error Analysis

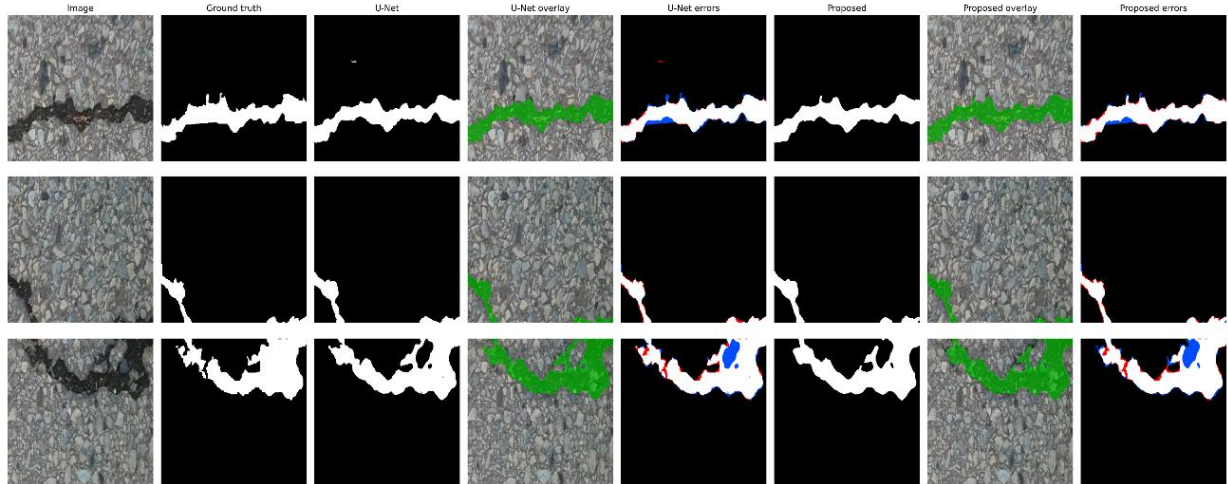


Figure 9. Qualitative comparison of U-Net and the proposed model

The seed-42 comparison in Table 4 shows a lower false discovery proportion ($FDP = 1 - \text{Precision}$) and false negative rate ($FNR = 1 - \text{Recall}$) for the proposed model than for the U-Net baseline. FDP decreased from 38.88% to 22.79%, while FNR decreased from 30.27% to 17.71%. These findings indicate smaller proportions of incorrect crack predictions and missed annotated crack pixels under the evaluated protocol. They do not establish prevention of prediction collapse.

Separately, Figure 11 illustrates prediction errors on selected test images at a probability threshold of 0.5, with false positives shown in red, false negatives in blue, and true positives in white. Both models produce non-empty crack predictions, with residual discrepancies around crack boundaries and locally misclassified regions. These examples illustrate error locations but do not establish dataset-wide superiority or exclude all-background failures on other images. The findings therefore support an analysis of segmentation errors rather than a claim that prediction collapse has been prevented.

Computational Cost Analysis

Table 8. Computational cost

Model	Parameter (M)	Estimated FLOPs for supported operations (GFLOPs)	Inference time (ms/image)
U-Net	13.428	150.013	36.91
Proposed Model	23.351	172.873	160.32

Table 8 reports the computational cost under the recorded profiling conditions. The proposed model contained 23.351 million parameters, compared with 13.428 million for U-Net. Estimated FLOPs for supported operations increased from 150.013 to 172.873 GFLOPs, while the reported inference time increased from 36.91 to 160.32 ms per image. These results indicate an increase in computational cost for the proposed architecture. Operational deployment requires further assessment under the intended hardware and processing conditions; these profiling results do not establish real-time readiness

Discussion

The proposed model integrates compact multiscale context aggregation, strip pooling, attention-based feature refinement, and auxiliary supervision to address thin and elongated pavement crack structures. Its segmentation performance is examined through comparison with U-Net, per-dataset evaluation, repeated evaluation runs, and threshold-sensitivity analysis. The results reported in Table 2 describe the observed variation across initializations, while the per-dataset results reveal remaining performance differences across data sources. The LOCO study provides supporting evidence about component contributions within the complete architecture; its single-seed findings do not establish consistency across initializations. The evaluations do not establish superiority of architectural changes over loss-function optimization or demonstrate prevention of prediction collapse [10], [14], [30].

CONCLUSION

This study proposes a pavement-specific Dual-Attention Residual U-Net integrating compact ASPP, strip pooling, and auxiliary boundary-aware supervision for fine-grained pavement crack segmentation. The complete model achieved a Dice score of 0.7967 and an IoU of 0.6621 in the seed-42 test evaluation. The supporting LOCO study examines the contribution of individual components within the complete model. The proposed framework has potential for supporting pavement inspection, while operational deployment requires further assessment of computational efficiency, inference latency, and performance under diverse field conditions. Future work will focus on broader validation of the proposed method and practical integration into pavement monitoring workflows [9], [31].

REFERENCES

[1] L. Ali, H. AlJassmi, M. Swavaf, W. Khan, and F. Alnajjar, "RS-Net: Residual Sharp U-Net Architecture for Pavement Crack Segmentation and Severity Assessment," *J. Big Data*, vol. 11, no. 1, Dec. 2024.

[2] A. Ashraf, A. Sophian, and A. A. Bawono, "Crack Detection, Classification, and Segmentation on Road Pavement Material Using Multi-Scale Feature Aggregation and Transformer-Based Attention Mechanisms," *Constr. Mater.*, vol. 4, no. 4, pp. 655–675, Dec. 2024.

[3] J. Xie, W. Li, W. Liu, and H. Chen, "A Road Crack Detection Method Based on Residual and Attention Mechanism," *Appl. Sci.*, vol. 14, no. 13, Jul. 2024.

[4] J. Wang and others, "Dual-Path Network Combining CNN and Transformer for Pavement Crack Segmentation," *Autom. Constr.*, vol. 158, Feb. 2024.

[5] M. Feng and J. Xu, "Lightweight Dual-Attention Network for Concrete Crack Segmentation," *Sensors*, vol. 25, no. 14, Jul. 2025.

[6] R. Rill-García, E. Dokladalova, and P. Dokládál, "Pixel-Accurate Road Crack Detection in Presence of Inaccurate Annotations," *Neurocomputing*, vol. 480, pp. 1–13, Apr. 2022.

[7] Z. Shi, N. Jin, D. Chen, and D. Ai, "A Comparison Study of Semantic Segmentation Networks for Crack Detection in Construction Materials," *Constr. Build. Mater.*, vol. 414, Feb. 2024.

[8] G. Li, Y. Niu, X. Guo, R. Hou, L. Zhang, and J. Cao, "Topology-Aware Road Crack Semantic Segmentation via Multiscale Stripe Pyramid and Cross-Layer Directional Attention," *Appl. Soft Comput.*, p. 114536, Dec. 2025.

- [9] J. Shan, Y. Huang, and W. Jiang, "DCUFormer: Enhancing Pavement Crack Segmentation in Complex Environments with Dual-Cross/Upsampling Attention," *Expert Syst. Appl.*, vol. 264, Mar. 2025.
- [10] J. Zhang and Z. V. Beliaeva, "CrackNet++: An Enhanced UNet Architecture Employing Tri-Axial Fusion Attention and Multi-Scale Gated Fusion for Precise Crack Segmentation and Quantification," *J. Build. Eng.*, vol. 118, Jan. 2026.
- [11] Z. Wei, A. Ji, Z. Xiao, and L. Zhang, "Segment Anything Model Enhanced Dual Encoder Segmentation for Seepage Quantification in Operational Tunnels," *Adv. Eng. Informatics*, vol. 69, Jan. 2026.
- [12] Z. Pan, S. L. H. Lau, X. Yang, N. Guo, and X. Wang, "Automatic Pavement Crack Segmentation Using a Generative Adversarial Network (GAN)-Based Convolutional Neural Network," *Results Eng.*, vol. 19, Sep. 2023.
- [13] Y. Song and others, "UCSDC: Pixel-Level Rectangular Self-Calibrating Crack Segmentation Network," *Structures*, vol. 80, Oct. 2025.
- [14] L. Wu and others, "Multiscale Semantic Segmentation and Damage Assessment of Village Houses in Post-Flood Scenarios Using an Enhanced DeepLabv3 with Dual Attention Mechanism," *Adv. Eng. Informatics*, vol. 68, Nov. 2025.
- [15] Y. Yan and others, "DCMA-Net: A Dual Channel Multi-Scale Feature Attention Network for Crack Image Segmentation," *Eng. Appl. Artif. Intell.*, vol. 148, May 2025.
- [16] B. Xiong and others, "FCT-Net: A Dual-Encoding-Path Network Fusing Atrous Spatial Pyramid Pooling and Transformer for Pavement Crack Detection," *Eng. Appl. Artif. Intell.*, vol. 137, Nov. 2024.
- [17] W. Wang and others, "DRA-UNet for Coal Mining Ground Surface Crack Delineation with {UAV} High-Resolution Images," *Sensors*, vol. 24, no. 17, Sep. 2024.
- [18] Y. Xu and others, "DRCNet: A Dual-Stream Multi-Scale Information Retention Network for Crack Segmentation," *Expert Syst. Appl.*, vol. 324, Aug. 2026.
- [19] B. Garg, D. C. Lohani, and B. Rana, "A Comparative Study of U-Net-Based Segmentation Architectures for Crack Detection," in *Procedia Computer Science*, 2025, pp. 3305–3314.
- [20] J. Zhao, T. Ma, Z. Wang, P. Cachim, and M. Qin, "Pavement Crack Detection and Segmentation Using Nested U-Net with Residual Attention Mechanism," *Comput. Civ. Infrastruct. Eng.*, vol. 40, no. 24, pp. 4076–4092, Oct. 2025.
- [21] S. Pervaiz and others, "Enhanced Concrete Crack Segmentation with MSMC-U-Net: Integrating Multiscale Features and Contextual Analysis for Infrastructure Safety," *Expert Syst. Appl.*, vol. 293, Dec. 2025.
- [22] X. Shen, Y. Zhang, X. Lin, W. Zhang, J. Lin, and Y. Fu, "Visualization of Pavement Structural Cracks Using a 3D Embedding Residual Attention U-Net for GPR Data," *Autom. Constr.*, vol. 181, Jan. 2026.
- [23] A. A. P., J. J. T. J., H. K., V. K., S. Kumar, and A. Sekar, "DAEENet: A novel Dual Attention based Edge Enhanced network for asphalt pavement deterioration detection and segmentation," *Case Stud. Constr. Mater.*, Dec. 2025.
- [24] Z. Xie and others, "Semantic Segmentation for Tooth Cracks Using Improved DeepLabv3+ Model," *Heliyon*, vol. 10, no. 4, Feb. 2024.
- [25] C. Bai, W. Shi, M. Zhang, and H. Zhao, "CrackDualMamba: A Lightweight Dual-Stream Mamba with Novel Focal Dice Balanced Loss for Vehicle-Based Road Crack Segmentation," *Adv. Eng. Informatics*, vol. 71, Apr. 2026.
- [26] Y. Ma, T. Bao, Y. Li, M. Zhao, Z. Wu, and C. Fan, "GANFormerNet: A UAV-Based Concrete Crack Segmentation Model for Water-Related Structures Using Vision Transformer and Graph Attention Network," *Adv. Eng. Informatics*, vol. 68, Nov. 2025.
- [27] O. Poudel, X. Hu, R. H. Assaad, W. Wang, T. Huang, and G. Wang, "VTECSeg: Edge-Aware Hybrid CNN-Vision Transformer Network with Zero-Shot Vision-Language-Guided Region Proposals for Crack Segmentation," *Autom. Constr.*, vol. 188, Aug. 2026.
- [28] S. Joo, S. Kim, and H. Kim, "Frequency-Aware Crack Segmentation Network (FACS-Net) and Crack Topology Loss (CT-Loss) for Thin Cracks," *Autom. Constr.*, vol. 182, p. 106719, Feb. 2026.
- [29] Z. Pan and others, "High-Precision Segmentation and Quantification of Tunnel Lining Crack Using an Improved DeepLabV3+," *Undergr. Sp.*, vol. 22, pp. 96–109, Jun. 2025.
- [30] Y. Fan and others, "DME-DeepLabv3+: A Perception-Driven Semantic Segmentation Approach for Intelligent Pavement Crack Maintenance," *Expert Syst. Appl.*, Jun. 2026.
- [31] Y. Li, C. Li, P. Zhang, and H. Wang, "A Real-Time Dual-Task Defect Segmentation Network for

Grinding Wheels with Coordinate Attentioned-ASP and Masked Autoencoder,” *Machines*, vol. 12, no. 4, Apr. 2024.