


Article

RoadNet: A High-Precision Transformer-CNN Framework for Road Defect Detection via UAV-Based Visual Perception

Long Gou ¹, Yadong Liang ^{1,2}, Xingyu Zhang ^{3,4} and Jianfeng Yang ^{3,4,*} ¹ Shanxi Road and Bridge Construction Group Co., Ltd., Taiyuan 710075, China² Shanxi Road and Bridge Qingyin Erguang Expressway Taiyuan Link Line Co., Ltd., Taiyuan 030000, China³ School of Electronic Information, Wuhan University, Wuhan 430072, China⁴ Pingyang Institute of Science and Technology Innovation, Wenzhou 325200, China

* Correspondence: yjf@whu.edu.cn

Abstract

Automated Road defect detection using Unmanned Aerial Vehicles (UAVs) has emerged as an efficient and safe solution for large-scale infrastructure inspection. However, object detection in aerial imagery poses unique challenges, including the prevalence of extremely small targets, complex backgrounds, and significant scale variations. Mainstream deep learning-based detection models often struggle with these issues, exhibiting limitations in detecting small cracks, high computational demands, and insufficient generalization ability for UAV perspectives. To address these challenges, this paper proposes a novel comprehensive network, RoadNet, specifically designed for high-precision road defect detection in UAV-captured imagery. RoadNet innovatively integrates Transformer modules with a convolutional neural network backbone and detection head. This design not only significantly enhances the global feature modeling capability crucial for understanding complex aerial contexts but also maintains the computational efficiency necessary for potential real-time applications. The model was trained and evaluated on a self-collected UAV road defect dataset (UAV-RDD). In comparative experiments, RoadNet achieved an outstanding mAP@0.5 score of 0.9128 while maintaining a fast-processing speed of 210.01 ms per image, outperforming other state-of-the-art models. The experimental results demonstrate that RoadNet possesses superior detection performance for road defects in complex aerial scenarios captured by drones.

Keywords: object detection; transformer; convolutional neural networks; deep learning; unmanned aerial vehicle (UAV); aerial imagery



Academic Editor: Pablo Rodríguez-González

Received: 29 August 2025

Revised: 28 September 2025

Accepted: 28 September 2025

Published: 9 October 2025

Citation: Gou, L.; Liang, Y.; Zhang, X.; Yang, J. RoadNet: A High-Precision Transformer-CNN Framework for Road Defect Detection via UAV-Based Visual Perception. *Drones* **2025**, *9*, 691. <https://doi.org/10.3390/drones9100691>

Copyright: © 2025 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) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The rapid advancement of Unmanned Aerial Vehicle (UAV) technology has revolutionized the field of infrastructure inspection, offering a safe, efficient, and cost-effective alternative to traditional manual or vehicle-based surveys [1]. Utilizing drones for road defect detection allows for the rapid collection of high-resolution imagery over large and potentially hazardous areas, minimizing traffic disruption and inspection risks. However, the vast amount of aerial image data generated by UAVs necessitates the development of robust automated analysis systems, making object detection a core technology in drone-based inspection pipelines [2].

Road defects such as potholes, cracks, and surface deterioration are inevitable due to factors like aging infrastructure, climate change, and heavy traffic loads. These defects

not only pose significant threats to transportation safety but also lead to accelerated road degradation and substantially increased long-term maintenance costs if not addressed promptly [3]. Therefore, the timely and accurate identification of these problems from UAV imagery is of paramount importance for modern urban management and smart maintenance strategies.

While deep learning models, particularly Convolutional Neural Networks (CNNs), have achieved remarkable success in various object detection tasks [4], their application to UAV-captured road imagery presents unique and formidable challenges. Firstly, the bird's-eye perspective of drones means that targets like fine cracks and small potholes occupy an extremely small proportion of pixels in the image, making them notoriously difficult to detect against complex backgrounds like asphalt texture, shadows, and occlusions [5]. Secondly, the scale of objects can vary dramatically within and across images due to changes in flight altitude and camera angle, demanding a detection model with superior multi-scale feature representation capabilities. Furthermore, deep models often require substantial computational resources, posing a challenge for real-time or on-device processing scenarios which are highly desirable for UAV applications [6]. Finally, models trained on data from one specific environment often suffer from performance degradation when deployed under different conditions (e.g., varying lighting, road types, seasons), highlighting a critical need for improved generalization ability.

In response to the aforementioned challenges specific to aerial imagery, this paper proposes a novel high-precision road defect detection framework named RoadNet. This innovative deep learning model integrates the global contextual modeling strengths of Transformer architectures with the local feature extraction efficiency of convolutional neural networks. It is specifically designed to address the shortcomings of existing technologies in handling UAV-based inspection data. After conducting experiments on a dedicated UAV road defect dataset, it has been verified that RoadNet attains a detection accuracy of over 90%. Compared with existing state-of-the-art models, it achieves significant improvements in both precision and recall. The main contributions of this paper are summarized as follows:

1. To address the critical challenge of capturing long-range dependencies and complex contextual information in expansive aerial imagery, a Transformer module is incorporated into the network. This enhancement enables the model to more accurately identify and delineate defects ranging from minute cracks to large-area potholes, which is a common scenario in UAV inspections.
2. To overcome the limitations of handling extreme scale variations inherent in UAV perspectives, a multi-level feature pyramid network is employed to effectively fuse features across different scales. Coupled with an optimized detection head structure, this ensures robust performance on targets of various sizes while maintaining computational efficiency for detecting small targets.
3. A Spatial-Channel Interaction Module (SCIM) is designed, building upon the Transformer and feature pyramid network. This module facilitates simultaneous capture of global and local features by jointly modeling spatial and channel information, significantly enhancing the feature representation power for complex aerial scenes.

2. Related Works

2.1. Traditional Road Inspection Methods

Traditional road defect detection primarily relies on mechanical and physical measurement approaches, including laser scanning [7], ground penetrating radar (GPR) [8], and high-resolution photogrammetry [9]. For instance, ground penetrating radar utilizes electromagnetic waves to probe the internal structure of roadways, effectively identify-

ing subsurface anomalies such as hidden voids and propagating cracks. Despite their accuracy, these methods suffer from high operational complexity and substantial costs, which severely restrict their large-scale deployment. Photogrammetry techniques, which employ high-resolution cameras to capture road surface images for subsequent analysis, offer a more cost-effective alternative. However, their performance is highly susceptible to environmental variations such as lighting changes and obstructions, leading to unstable detection outcomes. While these conventional approaches provide reliable measurements under controlled conditions, their limitations in cost, efficiency, and operational flexibility render them inadequate for modern large-scale infrastructure inspection requirements, especially from aerial platforms.

2.2. Deep Learning-Based Object Detection

The remarkable progress in artificial intelligence has established deep learning as a powerful paradigm for image processing and object detection tasks. Current deep learning-based detection methodologies can be broadly categorized into two architectural families.

The first category encompasses convolutional neural network (CNN) based approaches, which have demonstrated exceptional performance in various visual tasks including image classification, object detection, and semantic segmentation. In the domain of road defect analysis, popular architectures such as Faster R-CNN [10], YOLO [11], and SSD [12] have been extensively adopted. These models leverage convolutional operations to extract discriminative local features, enabling precise identification of road anomalies. Faster R-CNN achieves accurate defect localization through its region proposal network followed by classification and regression operations [10]. The YOLO framework, renowned for its inference efficiency, has been successfully applied in real-time road inspection systems [13]. SSD incorporates multi-scale feature fusion mechanisms to handle objects of varying dimensions, demonstrating robust performance across diverse defect sizes [12]. However, when deployed for UAV aerial imagery analysis, these CNN-based architectures face significant limitations. The characteristically small size of targets in aerial perspectives (where fine cracks may occupy merely several pixels), substantial scale variations due to altitude changes, and complex background clutter frequently cause performance deterioration, as these models lack effective global contextual modeling and efficient multi-scale representation capabilities [14,15].

The second category involves Transformer-based detection frameworks, which have recently gained prominence due to their exceptional global modeling capacities demonstrated in both natural language processing and computer vision domains. Representative models such as DETR [16] and SWIN Transformer [17] have shown remarkable capabilities in capturing long-range dependencies among image features. DETR revolutionizes object detection by implementing an end-to-end framework through self-attention mechanisms, eliminating the need for hand-designed components like region proposal networks [16]. SWIN Transformer introduces a hierarchical architecture with shifted window attention, demonstrating superior performance in processing high-resolution imagery [17]. These characteristics make Transformer-based models particularly suitable for aerial image analysis where comprehensive contextual understanding is essential for distinguishing true defects from complex background patterns. Nevertheless, their substantial computational requirements present considerable challenges for real-time deployment on resource-constrained UAV platforms or for processing large-scale aerial survey datasets [18].

2.3. UAV-Based Visual Inspection

UAV-based visual inspection has emerged as a rapidly evolving research domain, driven by the operational advantages of drone platforms for infrastructure monitoring [19].

Several investigations have explored the adaptation of existing deep learning architectures for road defect detection from aerial perspectives. For example, ref. [20] implemented an optimized YOLOv5 model on a UAV platform for automated road crack detection, achieving a balance between accuracy and computational efficiency. Similarly, ref. [21] proposed an attention-guided feature fusion network to enhance the detection of small cracks in UAV-captured images. Furthermore, ref. [22] developed a large-scale benchmark dataset for UAV-based road damage detection and provided a comprehensive evaluation of state-of-the-art models, highlighting the critical challenge of scale variation. More recently, refs. [23,24] explored the application of a lightweight Vision Transformer for real-time road inspection from drones, demonstrating its strong global feature extraction capability while addressing its computational demands. These collective efforts underscore the distinctive challenges of UAV-based road inspection—particularly in small object detection, computational efficiency, and model generalization—which demand specialized solutions beyond mere adaptation of existing ground-based models. Our proposed RoadNet framework is designed to address these specific challenges through a novel integration of convolutional and transformer architectures.

2.4. Road Defect Datasets

The development and benchmarking of deep learning models for road defect detection are heavily reliant on large-scale, high-quality annotated datasets. Several public datasets have been established, primarily focusing on ground-level vehicle perspective imagery, which presents characteristics fundamentally different from the aerial perspective of UAVs.

Among the most prominent datasets is the Road Damage Dataset (RDD2022) [22], a large-scale collection containing over 47,000 images from multiple countries, annotated with eight types of road damage. While extensive, its imagery is captured from street-level vehicles, resulting in a viewpoint and target scale that are not directly transferable to UAV-based inspection tasks. Similarly, the Crack500 dataset [25] and others like GAPS384 [26] focus specifically on pavement cracks, offering high-precision annotations but remaining constrained to the ground-level perspective and a limited range of defect types [27].

To address the growing interest in UAV applications, some datasets have been introduced that capture aerial imagery [26]. However, they often exhibit limitations in terms of image resolution, annotation granularity, or scenario diversity [28]. For instance, some datasets may utilize lower-resolution cameras or lack detailed bounding-box annotations precise enough for training models to detect extremely small defects like fine cracks from high altitudes [29]. Furthermore, many lack the variety in lighting conditions, weather, and complex urban backgrounds that are crucial for developing robust real-world applications [30].

Our self-constructed Unmanned Aerial Vehicle Road Defect (UAV-RDD) dataset is designed specifically to overcome these limitations and fill the gap in the domain of UAV-based infrastructure inspection. The major contributions of the UAV-RDD dataset are as follows:

1. All imagery is captured from a bird's-eye view using a DJI Mavic 3 Enterprise drone (DJI Innovations, Shenzhen, China), simulating real-world inspection scenarios with altitudes between 30 and 50 m. This perspective introduces the unique challenges of extreme scale variation and small target sizes, which are central to our research.
2. The dataset comprises 5842 high-resolution images (3840×2160 pixels), ensuring that minute defects remain discernible and providing rich detail for model training.
3. The dataset covers a wide spectrum of scenarios, including urban roads, highways, and rural settings under various lighting and weather conditions. This diversity is critical for enhancing the model's robustness and generalization capability.

- Each image is meticulously annotated by domain experts using bounding boxes for two critical defect types: cracks and potholes. The annotations follow the standard PASCAL VOC format, ensuring compatibility with mainstream detection frameworks.

Therefore, the creation of the UAV-RDD dataset addresses a clear need for a benchmark that accurately reflects the challenges of UAV-based visual perception, providing a dedicated resource for training and evaluating models like our proposed RoadNet for high-precision aerial road defect detection.

3. Method

Object detection in UAV-captured aerial imagery presents distinct challenges that demand specialized network architectures. The bird's-eye perspective and variable flight altitudes lead to extreme scale variations, where targets like road cracks may occupy only a few pixels while large potholes span significant areas. Additionally, complex background elements such as asphalt texture, shadows, and occlusions further complicate accurate detection. Traditional convolutional neural networks (CNNs) typically achieve multi-scale representation through hierarchical feature learning, progressively capturing fine-grained to coarse-grained features across network depths. However, this approach shows inherent limitations in capturing global dependencies and detailed information simultaneously, particularly problematic for UAV-based road defect detection where both minute cracks and extensive depressions must be detected within the same framework.

To address these specific challenges of aerial imagery analysis, our RoadNet framework incorporates a hybrid architecture that combines the strengths of Transformer-based global context modeling with sophisticated multi-scale feature processing capabilities. As illustrated in Figure 1, our model integrates a multi-head self-attention mechanism (MHSA) into critical layers of the backbone network, working synergistically with traditional convolutional units to efficiently combine local feature extraction with global contextual understanding. The Transformer's proven success in natural language processing demonstrates its powerful global context modeling ability through long-range dependency capture, a characteristic that translates exceptionally well to visual tasks involving complex aerial scenes. This integration is particularly valuable for UAV imagery, where understanding global context is essential for distinguishing true defects from similar-looking background patterns.

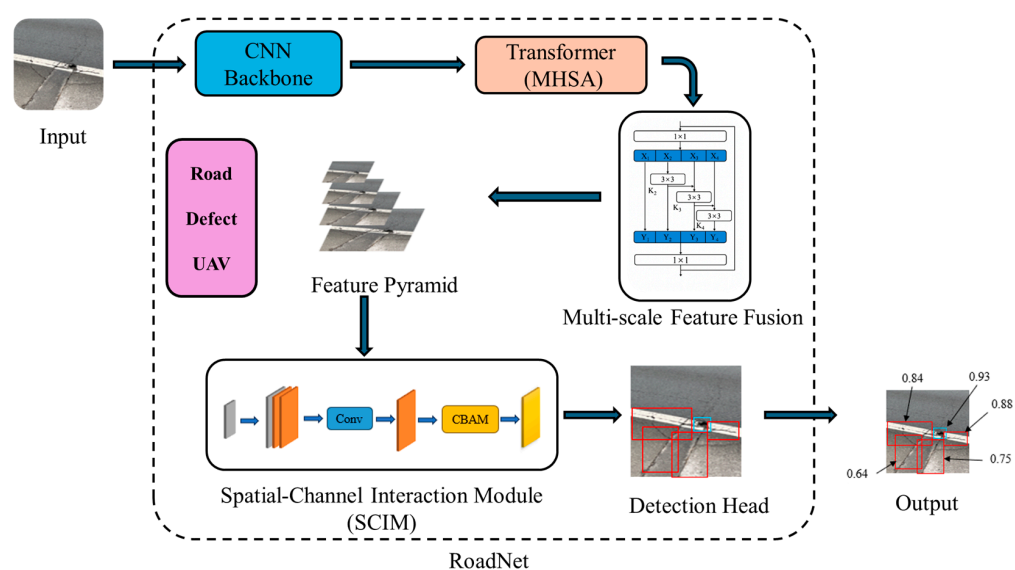


Figure 1. Overall Structure of RoadNet.

3.1. Multi-Scale Feature Representation and Optimization Strategies for Aerial Imagery

Central to RoadNet's design is the multi-scale feature fusion module that directly addresses the critical challenge of extreme scale variations in UAV-captured images. This module, now integrated within the comprehensive architecture shown in Figure 1, employs specialized multi-scale convolution operations and feature pyramid strategies to handle road defects ranging from few-pixel cracks to large-area potholes that vary with flight altitude and camera perspective. The module works in concert with the Transformer components to ensure robust feature representation across diverse scales, enabling simultaneous detection of both fine details and larger structural anomalies within the unified framework.

The Transformer module significantly enhances the model's ability to capture global dependencies through the self-attention mechanism, which is particularly crucial for analyzing expansive aerial views where defects may be distributed across wide areas. The success of this mechanism in natural language processing demonstrates its capability to effectively handle long-range dependencies, a valuable characteristic for processing high-resolution UAV imagery that requires understanding contextual relationships across the entire frame. Its core calculation method is as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

In this attention mechanism formula, Q , K , and V represent query, key, and value, respectively. The specific explanations are as follows: Q (Query): Represents an input vector, usually from the hidden state or input features of the current step. It can be understood as the "information" you want to find. K (Key): Each input vector has a corresponding key that is used to match the query vector. It is the benchmark used to calculate relevance during the query. V (Value): The value corresponding to each key. In the final calculation result, the value is the information we want to obtain. Here, $Q = XW_Q$, $K = XW_K$, $V = XW_V$, respectively, represent the linear transformation of the input feature X . W_Q , W_K , W_V are the parameter matrices learned, and d_k is the scaling factor of the feature dimension, used to prevent the values from being too large.

In the actual implementation, the multi-head self-attention mechanism processes different feature subspaces from multiple perspectives by parallelizing multiple independent attention heads. This parallel processing capability is essential for handling the diverse visual patterns found in aerial road imagery. The final output is the concatenation of the attention results from each head.

$$MultiHead(Q, K, T) = Concat(head_1, head_2, \dots, head_h)W_0 \quad (2)$$

Here, W_0 represents the output linear transformation matrix. By introducing this mechanism at the key levels of the main network, we achieve global and local modeling of multi-scale features. This design is particularly suitable for complex road detection tasks, enabling the capture of large-scale defects while not neglecting fine cracks and edge information.

In the main network, this self-attention mechanism is integrated into specific key layers to simultaneously focus on local and global information during multi-scale feature extraction. For example, for a certain intermediate feature map X , the convolution operation extracts its local features, while the Transformer module captures the global context through multi-head self-attention. This dual approach significantly enhances the expressiveness of the features.

To fully utilize features of different resolutions, this paper adopts a bottom-up multi-scale feature fusion strategy in the main network of the network. The low-level features

generated by the convolution operation typically contain rich edge and texture information, while the high-level features capture semantic information. This paper designs a Cascaded Attention Mechanism (CAM), combining the global modeling ability of the Transformer and the local perception ability of the convolution, and achieving efficient context information integration through a multi-scale feature pyramid.

The key of the cascaded attention mechanism lies in introducing adaptive weights in each level of features to automatically adjust the importance of different feature scales. Specifically, assuming that at a certain scale s , the feature map is F_s , then the fused feature representation is:

$$F'_s = \sigma(W_1 \cdot F_s) + \phi(W_2 \cdot F_{s+1}) \quad (3)$$

Here: σ and ϕ represents a non-linear activation function; W_1 and W_2 are learnable parameters used to adjust the weights of different feature scales; F_{s+1} represents the features at a higher level.

This fusion strategy not only enhances the network's adaptability in multi-scale scenarios, but also further optimizes the network's detection performance for fine targets (such as road cracks).

3.2. Spatial and Channel Interaction Optimization

In traditional convolutional networks, the features in the spatial dimension and the channel dimension are typically processed separately. This separate processing approach limits the network's utilization of the expression ability of high-dimensional features to a certain extent. Specifically, the spatial dimension contains the geometric shape and position information of objects, while the channel dimension corresponds to the category and semantic information of the features. However, modeling these two types of information independently can easily lead to insufficient information interaction, especially in complex scenarios, which may result in improper balance between global information and detailed information.

To solve this problem, we designed a Spatial-Channel Interaction Module (SCIM). This module captures the interdependence of the spatial and channel dimensions simultaneously, achieving efficient joint feature modeling. The core idea of SCIM is to dynamically adjust the interaction strength of information from different dimensions through an adaptive weight mechanism, thereby generating more rich and more expressive features.

Spatial dimension attention is generated through the global pooling operation of the feature map and combined with convolution processing to obtain attention weights, which are used to highlight the features at specific positions. The specific formula is as follows.

$$Attention_{spatial}(F) = \sigma(Conv2D(Concat[AvgPool(F), MaxPool(F)])) \quad (4)$$

Here, AvgPool and MaxPool represent global average pooling and global maximum pooling, respectively; Concat indicates the feature concatenation operation; σ denotes the Sigmoid activation function. This dual-pooling strategy enhances the model's ability to focus on spatially significant regions in aerial imagery, where defects may be sparse and distributed unevenly.

Channel dimension attention generates channel weights through the global description of the feature map, which is used to enhance the expression ability of semantic information.

$$Attention_{channel}(F) = \sigma(W_2 \cdot ReLU(W_1 \cdot AvgPool(F))) \quad (5)$$

Here, W_1 and W_2 represent the weight matrices of the fully connected layers; ReLU denotes the ReLU activation function.

Based on the outputs of these two attention modules, the fused feature representation is calculated, and further high-dimensional features are extracted through convolution operations.

$$F_{SCIM} = \text{Conv2D}\left(\text{Attention}_{\text{spatial}}(F) + \text{Attention}_{\text{channel}}(F)\right) \quad (6)$$

This joint modeling approach effectively integrates local detail information with global contextual understanding, enabling the model to perform significantly better in multi-scale feature extraction and adaptability to complex aerial scenarios. The SCIM proves particularly valuable for UAV-based inspection tasks, where the model must simultaneously maintain sensitivity to small spatial details while understanding global scene context to reduce false positives from background clutter.

4. Experiments and Results

4.1. UAV Image Dataset and Experimental Setup

To verify the performance of RoadNet in UAV-based road defect detection, the model was trained and evaluated on a dedicated Unmanned Aerial Vehicle Road Defect (UAV-RDD) dataset, which was specifically collected for this study. All imagery was captured using a DJI Mavic 3 Enterprise drone equipped with a 4/3 CMOS 20 MP camera (DJI Innovations, Shenzhen, China) flown at altitudes between 30 and 50 m to simulate real-world inspection scenarios. This study initially focuses on the two most common types of structural road defects—cracks and potholes. These two defect categories exhibit distinct morphological characteristics and scale variations, which are sufficient to validate the model's performance in multi-scale object detection and complex backgrounds. Extending the detection scope to include more defect categories such as ruts and subsidence represents an important direction for our future work. The dataset covers these two main types of road defects critical for infrastructure assessment, featuring diverse urban, highway, and rural scenarios under various lighting and weather conditions to ensure robustness.

The UAV-RDD dataset comprises 5842 high-resolution aerial images (3840×2160 pixels), each meticulously annotated by domain experts using the Labellmg tool. The annotation process followed the standard PASCAL VOC format, with bounding boxes for both crack and pothole defects. To address the class imbalance and extreme scale variations inherent in aerial imagery, we implemented a stratified sampling strategy during dataset splitting. The dataset was divided into 70% for training (4089 images), 15% for validation (876 images), and 15% for testing (877 images), ensuring proportional representation of defect types and environmental conditions across all subsets.

Recognizing the challenges of small object detection in aerial imagery, we employed an extensive data augmentation pipeline tailored to UAV characteristics. This included not only standard techniques such as random flipping (horizontal and vertical), cropping ($\pm 20\%$), and color jittering (brightness, contrast, saturation adjustments of $\pm 30\%$), but also altitude simulation via random scaling ($0.5\times$ to $1.5\times$) to mimic varying flight heights, and affine transformations to simulate different drone pitch and yaw angles. Additionally, we added Gaussian noise to improve model resilience to transmission artifacts and sensor noise common in real-world drone operations.

All experiments were conducted on a uniform hardware platform equipped with an NVIDIA RTX 4060 GPU (NVIDIA Corporation, Santa Clara, CA, USA) and an Intel i9-13900K CPU (Intel Corporation, Santa Clara, CA, USA) to ensure fair comparison. The models were implemented in PyTorch 1.12 and trained with identical hyperparameters: input image size of 640×640 , batch size of 16, AdamW optimizer with an initial learning rate of 0.001, and a cosine annealing scheduler over 300 epochs. All models were trained

on the GPU platform. However, inference speed was evaluated on the CPU to simulate a more realistic deployment scenario on edge devices or ground control stations.

4.2. Ablation Study on Component Effectiveness

To rigorously evaluate the contribution of each proposed module within the context of aerial imagery analysis, we conducted a series of ablation experiments. The results, presented in Table 1, demonstrate the incremental performance gain achieved by integrating Transformer-based global context modeling and the Spatial-Channel Interaction Module (SCIM) into our baseline CNN architecture.

Table 1. Ablation study of RoadNet components on the UAV-RDD test set.

Backbone	Transformer	SCIM	mAP@0.5	Precision	Recall	Params (M)
CNN (Baseline)	✗	✗	0.7357	0.7111	0.7172	7.2
CNN	✓	✗	0.8157	0.7904	0.7831	18.5
CNN	✓	✓	0.9128	0.904	0.8328	21.3

The baseline model (a CSPDarknet backbone with PANet neck and YOLO head) achieved an mAP@0.5 of 0.7357, highlighting the inherent difficulty of detecting small, low-contrast defects in aerial imagery. The incorporation of the Transformer module led to a substantial improvement (+10.9% mAP), underscoring the critical importance of global contextual information for distinguishing defects from complex aerial backgrounds like tar streaks, shadows, and water stains. The full RoadNet model, with both Transformer and SCIM, achieved the best performance (0.9128 mAP), validating the design of SCIM for effective spatial-channel feature interaction, which is paramount for precise localization and classification of varied defect sizes in UAV images. The precision and recall rates also showed significant gains, indicating a superior balance between reducing false positives and minimizing missed detections.

Furthermore, the ablation study confirms that our architectural innovations in RoadNet directly address the core challenges of UAV-based detection. Compared to the baseline, the final model achieved a 37.55% improvement in accuracy and a 31.17% increase in recall rate, resulting in a significant leap in overall performance. Specifically, the trend of key evaluation indicators during the training process is shown in Figure 2, while the comparison of detection results on the validation set is presented in Figure 3. These results fully demonstrate that our proposed architectural innovations play a decisive role in the enhancement of model performance for aerial imagery [17].

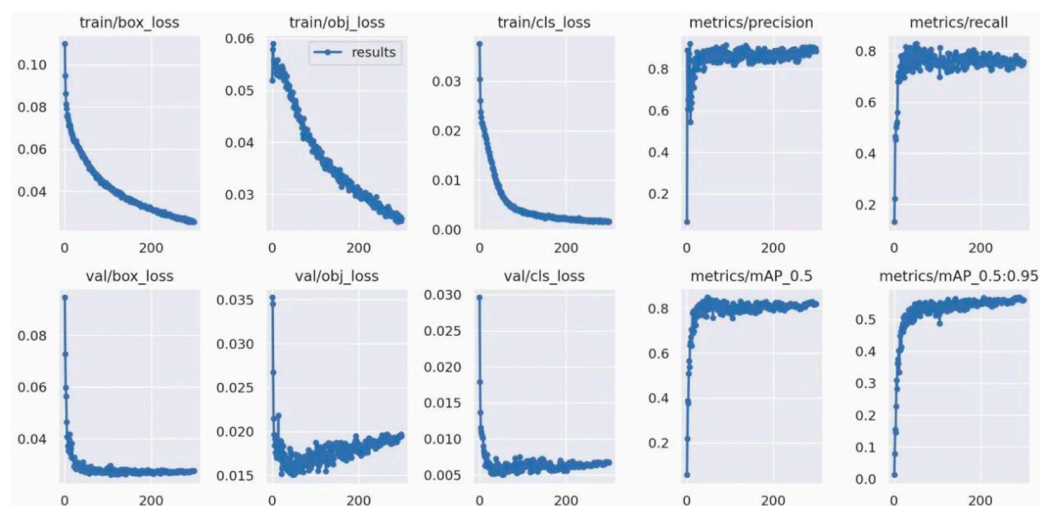


Figure 2. Trend of Key Evaluation Indicators.

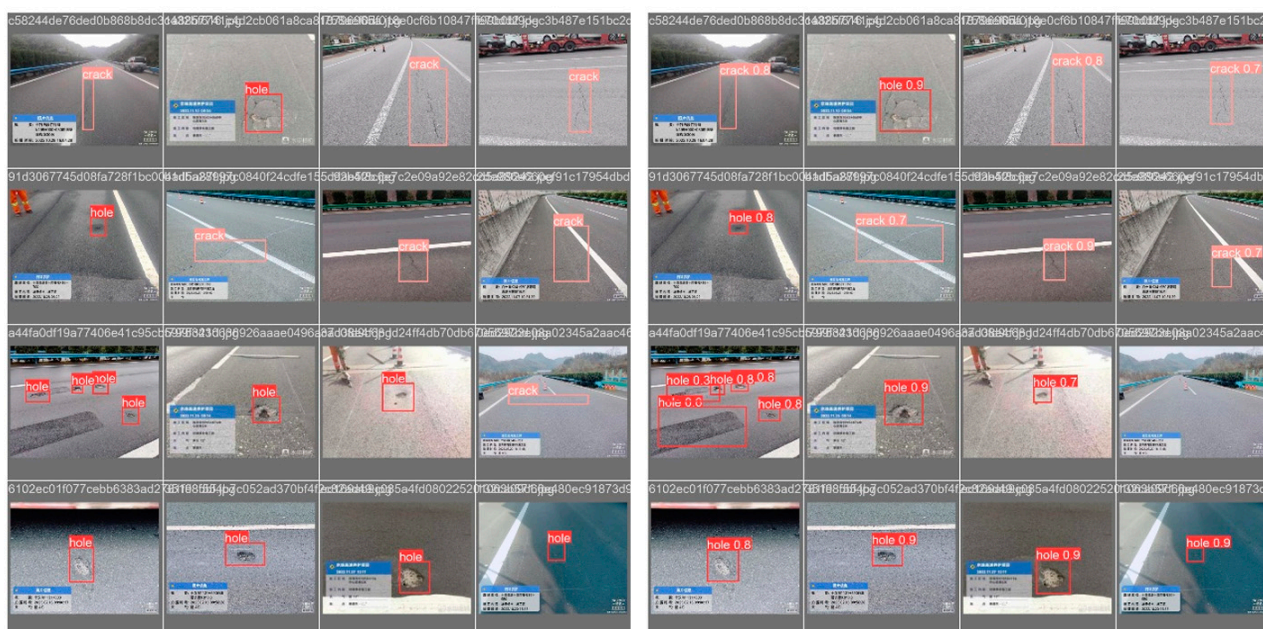


Figure 3. Comparison of Labels and Predictions on the Validation Set.

4.3. Comparative Evaluation with State-of-the-Art Models

A comprehensive comparative experiment was conducted to benchmark RoadNet against several state-of-the-art object detection models, including both established architectures and recent lightweight networks. All models were trained from scratch on our UAV-RDD dataset under the identical experimental setup described in Section 4.1 to ensure a fair and meaningful comparison. The evaluation metrics focus on both detection accuracy (mAP@0.5) and inference speed (ms per image on CPU), the latter being crucial for real-time UAV applications.

The results in Table 2 demonstrate the superior overall performance of RoadNet. Our model achieves the highest detection accuracy (0.9128 mAP@0.5), outperforming all other compared methods by a significant margin. Specifically, RoadNet surpasses the next best performer, YOLOv8s, by 25.3% in mAP, highlighting its exceptional capability in capturing both global context and fine-grained details essential for accurate defect detection in challenging aerial scenes. Notably, while lightweight architectures like YOLOv5-ShuffleNetv2 and YOLOv5-FasterNet achieve faster inference speeds (185.2 ms and 192.7 ms, respec-

tively), this comes at the cost of substantially reduced accuracy (0.6831 and 0.6945 mAP), demonstrating the typical trade-off between efficiency and precision.

Table 2. Performance comparison of different models on the UAV-RDD test set.

Model	mAP@0.5	Inference Time (ms)	Platform	Params (M)
RoadNet (Ours)	0.9128	210.1	CPU (Intel i9-13900K)	21.3
Faster R-CNN [20]	0.4283	1221.3	CPU (Intel i9-13900K)	136.7
YOLOv5s [18]	0.7225	220.4	CPU (Intel i9-13900K)	7.2
YOLOv5l [21]	0.7234	993.8	CPU (Intel i9-13900K)	46.5
YOLOv8s [19]	0.7284	330.6	CPU (Intel i9-13900K)	11.1
DETR [16]	0.7012	1850.5	CPU (Intel i9-13900K)	41.2
YOLOv5-ShuffleNetv2	0.6831	185.2	CPU (Intel i9-13900K)	5.8
YOLOv5-FasterNet	0.6945	192.7	CPU (Intel i9-13900K)	6.1

Crucially, RoadNet maintains an excellent balance between these competing objectives. With an inference speed of 210.1 ms/image on a CPU platform—the most likely deployment scenario for edge computing on UAVs or ground control stations—RoadNet is 4.7% faster than YOLOv5s while delivering dramatically superior accuracy. This represents a significant efficiency breakthrough compared to heavier architectures like Faster R-CNN (1221.3 ms) and DETR (1850.5 ms), making near-real-time analysis during flight missions practically feasible.

The combination of state-of-the-art accuracy and leading inference efficiency makes RoadNet uniquely suitable for real-world UAV-based road inspection tasks, successfully reconciling the often-contradictory goals of high precision and computational practicality.

5. Discussion

The experimental results presented in the previous section demonstrate that RoadNet achieves state-of-the-art performance on the challenging task of UAV-based road defect detection. This section provides a comprehensive discussion of these findings, highlighting the advantages of our approach, acknowledging its limitations, and outlining directions for future research.

RoadNet's exceptional performance (0.9128 mAP@0.5) can be attributed to its innovative architectural design that effectively addresses the unique challenges of aerial imagery analysis. The integration of Transformer modules enables superior global context modeling, which is crucial for distinguishing true defects from complex background patterns such as asphalt textures, shadows, and stains. Simultaneously, the Spatial-Channel Interaction Module (SCIM) facilitates effective multi-scale feature representation, allowing the model to detect both minute cracks and larger potholes within the same framework. The efficient inference speed of 210.1 ms/image on a CPU platform further demonstrates that these performance gains do not come at the cost of computational practicality, making RoadNet suitable for real-world deployment.

From an application perspective, RoadNet offers significant practical advantages for infrastructure maintenance. Assuming a drone flight speed that captures 2 images per second with appropriate overlap, and considering RoadNet's processing speed, the system could theoretically complete imaging and analysis of a 1 km two-lane road in approximately 15–20 min. This includes both data acquisition and processing time, representing a substantial efficiency improvement over traditional manual inspection methods that might require hours for the same distance, while also eliminating safety risks associated with road-side inspections.

The selection of a UAV-based inspection approach with RoadNet offers distinct advantages over competing methodologies. Compared to traditional ground-based inspection

methods, our approach provides superior coverage, reduced traffic disruption, and enhanced safety. When compared to other deep learning models, RoadNet achieves a better balance between accuracy and computational efficiency than pure Transformer architectures (e.g., DETR), while significantly outperforming lightweight CNN models (e.g., ShuffleNet, FasterNet) in detection accuracy. This balance is particularly crucial for UAV applications where both computational resources and detection reliability are constrained.

6. Conclusions

This paper has presented RoadNet, a novel deep learning framework specifically designed to address the critical challenges of road defect detection in Unmanned Aerial Vehicle (UAV) imagery. Confronted with the inherent difficulties of aerial-based inspection—including extreme scale variations, minuscule target sizes, complex background clutter, and the demand for computational efficiency—RoadNet integrates Transformer-based global context modeling with the local feature extraction strengths of convolutional neural networks. The incorporation of a dedicated Spatial-Channel Interaction Module (SCIM) further enhances multi-scale feature representation, enabling the model to precisely localize and classify diverse road defects, from fine cracks to extensive potholes, in complex aerial scenes.

Extensive experiments conducted on a dedicated UAV road defect dataset (UAV-RDD) demonstrate the effectiveness of our approach. Ablation studies confirm the significant individual and synergistic contributions of the Transformer module and SCIM, with the full RoadNet model achieving a remarkable mAP@0.5 of 0.9128. In comparative evaluations, RoadNet not only surpassed state-of-the-art models like YOLOv8s and Faster R-CNN in detection accuracy by a considerable margin but also achieved the fastest inference speed on a CPU platform (210.1 ms/image), underscoring its suitability for real-time UAV applications.

For future work, research will proceed along several promising directions:

Enhanced Generalization: We will incorporate more diversified UAV datasets captured under a wider range of conditions (e.g., severe weather, different times of day, and various geographic locations) to further improve the model's robustness and generalization capability.

Advanced Lightweighting: While efficient, we will explore more aggressive model compression and quantization techniques, including neural architecture search (NAS) for optimal backbone design, to deploy RoadNet on the limited computational resources of UAV embedded systems without significant accuracy loss.

Edge Deployment and Real-time System Integration: The ultimate goal is to deploy an optimized version of RoadNet on edge computing devices within UAVs or ground stations, facilitating a closed-loop system for real-time detection, analysis, and reporting, thereby providing stronger technical support for the development of intelligent and automated transportation infrastructure maintenance.

This technology shows great potential for reducing infrastructure inspection costs, improving inspection efficiency, and providing data support for preventive maintenance, ultimately contributing to enhanced road safety and extended road service life.

Author Contributions: Software, Y.L. and L.G.; formal analysis, J.Y.; investigation, J.Y.; resources, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was fully supported by the Key Research and Development Program of Hubei Province, China (2022BCA035). The numerical calculations were performed at the Supercomputing Center of Wuhan University.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: Long Gou and Yadong Liang was employed by the company Shanxi Road and Bridge Construction Group Co., Ltd. and Shanxi Road and Bridge Qingyin Erguang Expressway Taiyuan Link Line Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Zhou, Y.; Yue, Y.; Yan, B.; Li, L.; Xiao, J.; Yao, Y. Collaborative Target Tracking Algorithm for Multi-Agent Based on MAPPO and BCTD. *Drones* **2025**, *9*, 521. [\[CrossRef\]](#)
2. Yang, B.; Tao, T.; Wu, W.; Zhang, Y.; Meng, X.; Yang, J. MultiDistiller: Efficient Multimodal 3D Detection via Knowledge Distillation for Drones and Autonomous Vehicles. *Drones* **2025**, *9*, 322. [\[CrossRef\]](#)
3. Huang, Y.; Fan, J.Y.; Hu, J.Z.Y. TBi-YOLOv5: A surface defect detection model for crane wire with Bottleneck Transformer and small target detection layer. *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.* **2024**, *238*, 2425–2438. [\[CrossRef\]](#)
4. Su, Y.; Deng, J.; Sun, R.; Lin, G.; Su, H.; Wu, Q. A Unified Transformer Framework for Group-based Segmentation: Co-Segmentation, Co-Saliency Detection and Video Salient Object Detection. *IEEE Trans. Multimed.* **2022**, *26*, 313–325. [\[CrossRef\]](#)
5. Wang, A.; Ren, C.; Zhao, S.M.S. Attention guided multi-level feature aggregation network for camouflaged object detection. *Image Vis. Comput.* **2024**, *144*, 104953. [\[CrossRef\]](#)
6. Li, H.; Zhang, R.; Pan, Y.; Ren, J.; Shen, F. LR-FPN: Enhancing Remote Sensing Object Detection with Location Refined Feature Pyramid Network. *arXiv* **2024**, arXiv:2404.01614. [\[CrossRef\]](#)
7. Ha, T.T.; Chaisomphob, T. Automated Localization and Classification of Expressway Pole-Like Road Facilities from Mobile Laser Scanning Data. *Adv. Civ. Eng.* **2020**, *2020 Pt 6*, 5016783.1–5016783.18.
8. Ma, Y.; Lei, W.; Pang, Z.; Zheng, Z.; Tan, X. Rebar Clutter Suppression and Road Defects Localization in GPR B-Scan Images Based on SuppRebar-GAN and EC-Yolov7 Networks. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [\[CrossRef\]](#)
9. Lv, Y.; Wang, G.; Hu, X. Machine learning based road detection from high resolution imagery. *ISPRS Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *41*, 891–898.
10. Zhang, H.; Shao, F.; Chu, W.; Dai, J.; Li, X.; Zhang, X.; Gong, C. Faster R-CNN based on frame difference and spatiotemporal context for vehicle detection. *Signal Image Video Process.* **2024**, *18*, 7013–7027. [\[CrossRef\]](#)
11. Mohd Yusof, N.; Sophian, A.; Mohd Zaki, H.F.; Bawono, A. Assessing the performance of YOLOv5, YOLOv6, and YOLOv7 in road defect detection and classification: A comparative study. *Bull. Electr. Eng. Inform.* **2024**, *13*, 350. [\[CrossRef\]](#)
12. Zhang, B.; Fang, S.; Li, Z. Research on Surface Defect Detection of Rare-Earth Magnetic Materials Based on Improved SSD. *Complexity* **2021**, *2021*, 1–10. [\[CrossRef\]](#)
13. Zhang, L.; Yan, S.F.; Hong, J.; Xie, Q.; Zhou, F.; Rau, S. An improved defect recognition framework for casting based on DETR algorithm. *J. Iron Steel Res. Int. Ed.* **2023**, *30*, 949–959. [\[CrossRef\]](#)
14. Zhu, W.; Zhang, H.; Zhang, C.; Zhu, X.; Guan, Z.; Jia, J. Surface defect detection and classification of steel using an efficient Swin Transformer. *Adv. Eng. Inform.* **2023**, *57*, 1572. [\[CrossRef\]](#)
15. Wu, Y.; Liao, K.; Chen, J.; Wang, J.; Chen, D.Z.; Gao, H.; Wu, J. D-former: A U-shaped Dilated Transformer for 3D medical image segmentation. *Neural Comput. Appl.* **2022**, *35*, 1931–1944. [\[CrossRef\]](#)
16. Wang, X.; Gao, H.; Jia, Z.; Zhao, J. A road defect detection algorithm incorporating partially transformer and multiple aggregate trail attention mechanisms. *Meas. Sci. Technol.* **2024**, *36*, 026003. [\[CrossRef\]](#)
17. Kim, G.I.; Yoo, H.; Cho, H.J.; Chung, K. Defect Detection Model Using Time Series Data Augmentation and Transformation. *Comput. Mater. Contin.* **2024**, *78*, 1713. [\[CrossRef\]](#)
18. Jiang, T.Y.; Liu, Z.Y.; Zhang, G.Z. YOLOv5s-road: Road surface defect detection under engineering environments based on CNN-transformer and adaptively spatial feature fusion. *Measurement* **2025**, *242*, 115990. [\[CrossRef\]](#)
19. Wang, J.; Meng, R.; Huang, Y.; Zhou, L.; Huo, L.; Qiao, Z.; Niu, C. Road defect detection based on improved YOLOv8s model. *Sci. Rep.* **2024**, *14*, 16758. [\[CrossRef\]](#)
20. Fang, Z.; Shi, Z.; Wang, X.; Chen, W. Roadbed Defect Detection from Ground Penetrating Radar B-scan Data Using Faster RCNN. In *IOP Conference Series: Earth and Environmental Science*; IOP Publishing: Bristol, UK, 2020; pp. 131–137.
21. Sadhin, A.H.; Hashim, S.Z.M.; Samma, H.; Khamis, N. YOLO: A Competitive Analysis of Modern Object Detection Algorithms for Road Defects Detection Using Drone Images. *Baghdad Sci. J.* **2024**, *21*, 2167. [\[CrossRef\]](#)
22. Arya, D.; Maeda, H.; Ghosh, S.K.; Wang, Y.; Lee, D.; Zhang, L.; Liu, H.; Xu, W.; Chen, T.; Li, X.; et al. Global Road Damage Detection: A Large-Scale Dataset and Benchmark. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 23294–23305.

23. Kim, G.I.; Yoo, H.; Cho, H.J. Lightweight vision transformer for real-time road damage detection in UAV imagery. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 6003205.
24. Xiao, J.; Guo, H.; Zhou, J.; Zhao, T.; Yu, Q.; Chen, Y.; Wang, Z. Tiny object detection with context enhancement and feature purification. *Expert Syst. Appl.* **2023**, *211*, 118665. [\[CrossRef\]](#)
25. Zhang, D.; Yang, K.; Yang, L.; Liang, H.; Zhang, Q.; Liu, C. Crack500: A pavement crack dataset for deep learning. In Proceedings of the IEEE International Conference on Big Data, Seattle, WA, USA, 10–13 December 2018.
26. Eisenbach, M.; Stricker, R.; Seichter, D.; Amende, K.; Debes, K.; Sesselmann, M.; Ebersbach, D.; Stoeckert, U.; Gross, H.-M. How to get pavement distress detection ready for deep learning? A systematic approach. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; pp. 2039–2047.
27. Arya, D.; Maeda, H.; Ghosh, S.K.; Toshniwal, D.; Mraz, A.; Kashiya, T.; Sekimoto, Y. RDD2022: A multi-national image dataset for automatic road damage detection. *IEEE Trans. Intell. Transp. Syst.* **2024**, *11*, 846–886. [\[CrossRef\]](#)
28. Shtayat, A.; Moridpour, S.; Best, B.; Shroff, A.; Raol, D. A review of monitoring systems of pavement condition in paved and unpaved roads. *J. Traffic Transp. Eng. (Engl. Ed.)* **2020**, *7*, 629–638. [\[CrossRef\]](#)
29. Alhawsawi, A.N.; Khan, S.D.; Rehman, F.U. Enhanced YOLOv8-based model with context enrichment module for crowd counting in complex drone imagery. *Remote Sens.* **2024**, *16*, 4175. [\[CrossRef\]](#)
30. Mandal, V.; Mussah, A.R.; Adu-Gyamfi, Y. Deep learning frameworks for pavement distress classification: A comparative analysis. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020.

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.