

Received 11 January 2026, accepted 10 February 2026, date of publication 12 March 2026, date of current version 17 March 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3666237

RESEARCH ARTICLE

RDD4D: 4D-Attention-Guided Road Damage Detection and Classification

ASMA MUSABAH ALKALBANI¹, MUHAMMAD SAQIB^{2,3}, AHMED SALIM ALRAWAHI^{1,4},
ABBAS ANWAR^{5,6}, CHANDRANATH ADAK⁷, (Senior Member, IEEE), AND SAEED ANWAR⁸

¹Department of Information Technology, College of Computing and Information Sciences, University of Technology and Applied Sciences, Ibri 511, Oman

²National Collection and Marine Infrastructure, CSIRO, Canberra, ACT 2601, Australia

³University of Technology Sydney, Ultimo, NSW 2007, Australia

⁴AI Applications Chair, University of Nizwa, Nizwa 616, Oman

⁵FutureDataMinds, Australia

⁶Department of Computer Science, National University of Computer and Emerging Sciences, Peshawar 25000, Pakistan

⁷Department of Computer Science and Engineering, Indian Institute of Technology, Patna, Bihar 801106, India

⁸Department of Computer Science and Software Engineering, School of Physics, Maths, and Computing, The University of Western Australia, Perth, WA 6009, Australia

Corresponding author: Asma Musabah Alkalbani (asmam.alkalbani@utas.edu.om)

This work was supported in part by the Ministry of Higher Education, Scientific Research, and Innovation, Oman (MoHERI) under Grant BFP/RGP/ICT/21/406; and in part by the University of Technology and Applied Sciences, Ibri (UTAS-Ibri) under Grant IRG-IBRI-26-22.

ABSTRACT Road damage detection and assessment are crucial components of infrastructure maintenance. However, detecting multiple crack types in a single image remains a challenging limitation, particularly at varying scales. This is due to the lack of road datasets with various types of damage on varying scales. To address this challenge, we introduce: 1) A novel dataset called the Diverse Road Damage Dataset (DRDD) that captures multiple damage types per individual image. 2) A new detection model called RDD4D that exploits Attention4D blocks to enable better feature refinement across multiple scales. The Attention4D module processes feature maps through an attention mechanism combining positional encoding and talking head components to extract both local and global contextual features. In our comprehensive experimental analysis comparing various state-of-the-art models, our enhanced model demonstrates superior performance in detecting large-sized road cracks with an Average Precision (AP) of 0.458 and maintained competitive performance with an overall mAP of 0.446 on our proposed dataset. Moreover, we also provide results on the CrackTinyNet dataset; our model achieved around a 0.21 increase in performance. The code, model weights, dataset, and our results are publicly available at https://github.com/msaqib17/Road_Damage_Detection

INDEX TERMS Damage categorization, damage classification and recognition, damage detection, dataset construction, road inspection, surface inspection, type localization, UAV dataset construction.

I. INTRODUCTION

Roads and other public facilities are key contributors to a country's economy, particularly in urbanized cities. During periods of rapid economic growth in any country, the focus is on building roads and bridges. However, many of these infrastructures have aged due to multiple factors, including rain, weather, vehicles, and aging, causing various forms of road damage [1].

Road damage has a serious impact on driver safety, vehicle performance and value, as well as transportation

efficiency. Moreover, the number of roads inspected in the next few decades is expected to increase drastically; hence, each country requires considerable budgets for repair, restoration, rehabilitation, and maintenance [2]. According to the U.S. Federal Highway Administration, the U.S. road network was 4.17 million miles in 2020, an increase from 3.87 million miles in 1990 [3]. The U.S government spends over \$30 billion annually on road construction and maintenance [4]. In Europe, the annual number of road traffic injuries exceeds 50 million [5]. Similarly, a report by the U.S. Automobile Association estimates that road damage imposes an annual cost of approximately \$3 billion on U.S. drivers [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Jiankang Zhang¹.

Conventionally, expert engineers identify the affected infrastructure using manual visual assessment, which relies on their knowledge, experience, and background. This manual inspection is now complemented by a variety of sensors, including 3D sensors, multiple cameras, laser scanners, and road profilers. These sensors are used to capture different longitudinal and transversal views of road assets and pavements. However, manual assessment remains labor-intensive, costly, time-consuming, and prone to human errors. Therefore, there is a need for efficient and reliable approaches for effective monitoring of the road surface.

Many researchers tackle the issue of manual inspection by employing automatic road inspection techniques. They aim to study the road conditions using one of three main types: 1) laser scanning [6] methods, which provide accurate information about the condition of the road but are expensive and require a road closure, 2) vibration-based [7] methods are restricted to the touched road elements, and 3) vision-based [8], [9], [10] techniques are cost-effective but may lack accuracy. However, the rapid developments in vision-based approaches resulted in exceptional performance in diverse applications such as road sign detection [11], traffic analysis [12], crowd counting [13], and object detection [14].

Specialized inspection vehicles with different sensors, including laser scanners, road profilers, multiple cameras, and 3D sensors, are employed to inspect roads by capturing images of road assets, pavement images, and longitudinal and transverse image profiles [15]. Although these specialized vehicles are more efficient than the traditional manual human visual surveying approach, they can cost millions of dollars, depending on the sensors acquired to build the mounted system on the vehicle [16]. In contrast, consumer devices such as smartphones offer a cost-effective alternative with high-resolution cameras, robust sensors, and powerful processors. Therefore, vision-based techniques are evolving rapidly and becoming more widespread. We also anticipate that smartphones will play a critical role in reducing the cost of road inspection.

Recently, researchers have attempted to utilize vision-based techniques to inspect road surfaces by exploiting convolutional neural networks and deep learning methods. For this purpose, the research direction focuses either on detecting damage on the road surface irrespective of its type or classifying road damage type into a specific class [8]. Pioneering research focused on detecting directional cracks, i.e., horizontal and vertical [10], followed by detecting three road damage categories, i.e., horizontal, vertical, and alligator detection [17]. Likewise, the comprehensive classification of damage types on the road is provided by [9], as distinguishing damage types is crucial for accurate route planning.

However, current vision-based inspection approaches lack common ground and suffer from various shortcomings, such as 1) the results being compared on non-standard datasets, where each algorithm uses its own road damage dataset. We aim to create a standard dataset, drawing inspiration

from other research areas that have established standard benchmark datasets, such as DnD for denoising, ImageNet for object classification, and PartNet for point clouds. 2) The datasets usually have restricted views where the methods are evaluated. 3) The algorithms do not put much effort into the network design for road damage detection or employ state-of-the-art, off-the-shelf models from other vision research fields. A specialized network model is needed to cater to the needs of road damage detection. 4) Finally, road damage can be categorized into many types (e.g., eight damage types in Japan); however, the current research only considers a handful of them; hence, their direct application to practical scenarios is challenging.

Our Contributions: The contributions of this study are as follows.

- A novel Diverse Road Damage Dataset (DRDD) of 1500 smartphone images containing multiple damage types per frame, released for public use.
- A 4D Attention mechanism for road damage detection is introduced to improve feature refinement across scales. This enhancement enables the network to process both local details and global context simultaneously, leading to more accurate road damage detection across different scales.
- Extensive experimental evaluation is performed to assess the proposed mechanism against state-of-the-art models showing the strengths and limitations of current road damage detection approaches and providing comprehensive benchmarking results that show valuable insights for future research directions.

The remainder of this paper is organized as follows. Section II surveys the related work. Section III details DRDD. Section IV describes the proposed method. Section V presents the experimental setup. Section VI provides results and ablations. Section VII concludes.

II. RELATED WORK

Early detection of cracks in road surfaces is essential for road maintenance. It is a vital requirement for the safety and efficiency of road transportation. Early detection of cracks is challenging as it requires ongoing monitoring. Machine learning techniques, particularly Convolutional Neural Networks (CNN), are widely applied for automated road crack detection. These techniques have demonstrated considerable potential in enhancing classification accuracy and operational efficiency. However, a common limitation in the literature is the reliance on datasets that contain a large number of images but are limited in damage type heterogeneity (Table 1). This leaves rare defect categories underrepresented. In this section, we analyze recent studies on identifying road cracks using machine learning techniques, highlighting their objectives, contributions, and limitations. We also review the state-of-the-art crack detection using machine learning and explores potential future research directions.

A. TRADITIONAL METHODS

Road damage detection is studied using traditional image processing algorithms, which mainly rely on background subtraction, thresholding, segmentation, and feature extraction. For example, gray-scale histograms with OTSU thresholds [18], gray-level co-occurrence matrix [19], modified median filter and morphological filter [20], support vector machine algorithms [21], a library of machine learning models [22], and edge detectors [23] are used to detect pavement cracks and different types of road damages. However, these algorithms have limitations, such as sensitivity to lighting and background changes, reliance on handcrafted features, sensitivity to noise, complex backgrounds, and the inability to generalize to various roads and their types, which reduces road damage detection accuracy.

B. DEEP LEARNING METHODS

Recent research [24], [25] has utilized deep convolutional neural networks for automated road assessments and damage identification. A supervised deep CNN is trained to classify smartphone images of pavement, aiming to detect the presence of damage [10]. Similarly, the VGG-based network [26] detects cracks on concrete surfaces. Additionally, Crackpot [27] utilizes a camera-based GPU board to detect road cracks and potholes, enabling smooth journeys for self-driving vehicles and robots. A simple CNN is trained in [28] to recognize pavement conditions, demonstrating its ability to manage various pavement characteristics effectively. The deep neural networks outperform the edge-detection methods in identifying pavement damage [29].

A semi-supervised deep learning-based method using attention guidance is used in [30] to analyze cracks and spalls, where a human verifies the detected boundary box. The computational cost of the segmentation is significantly reduced due to the application of pixel-level segmentation. A segmentation-based analysis for cracks and potholes is performed by creating datasets composed of 2D, 3D, and enhanced 3D images in [31] in which stereo vision is integrated with deep learning. A modified U-Net based on depth-wise separable convolution is used to measure the volume of road potholes by performing 3D image segmentation, thereby accelerating the segmentation process. Similarly, the Asphalt method [32] is implemented to collect data from a vehicle's accelerometer sensors, recording vehicle vibrations and assessing road conditions using machine learning.

C. SMARTPHONE-BASED METHODS

Handheld devices, including smartphones, have recently become common for road inspection due to their increasing capabilities, such as sensors, high-resolution multiple cameras, large storage memory, and fast processors. Moreover, handheld devices are efficient and cost-effective for inspecting large road networks. For example, SmartPatrolling is proposed to collect data using a smartphone's built-in sensors with dynamic time warping for road surface

conditions, which performs more efficiently than traditional algorithms [33]. Road damage is detected using an onboard smartphone mounted on various vehicles, e.g., cars, buses, and trucks, and connected to a laptop for processing [34]. Smartphones are also used to visualize potholes on a map in [35]. A real-time smartphone application is developed in [9] for detecting and recording roadway cracks.

D. DATASETS FOR ROAD DAMAGE DETECTION

RDD-2018 [9] is a publicly available dataset collected through a smartphone application. Several municipalities are using this application to monitor road conditions faster. Consequently, researchers worldwide have shown interest in the data, methods, and models. Following [9], researchers have either added more images to the dataset or used the software to collect a novel dataset. For example, the authors of [36] expanded the dataset by incorporating images captured in Italy and Mexico, increasing the number to above 18k. Similarly, Roberts et al. [37] incorporated 7k images of road damage from Italy, utilizing the software from [9]. The model is trained on the new dataset, considering the identified road damage severity. The authors in [38] use an industrial high-resolution camera to create a dataset exceeding 45k road photos and utilize YOLO to detect and classify pavement damages.

On the contrary, Google Street View pictures of pavement, which are easily and freely accessible, are used by [39], which include top and wide angles, characterizing and calculating the density. The study demonstrates that the Faster R-CNN lags behind the YOLO-v2 model. Moreover, Potspot [40] employs limited Google API images for CNN-based detection of potholes. Despite the accessibility and cost-free nature of Google Street View photographs, annotating them remains demanding, laborious, and time-intensive. Recently, RDD [1] has been proposed for detecting and classifying road damage with many images. However, most photos only feature single damage, and the damages themselves are not particularly complex. On the other hand, the damages in our collected dataset are more complicated and typically occur in multiple instances within one image. Table 1 shows more information on the datasets.

The studies reviewed demonstrated that machine learning techniques, particularly deep CNNs, accurately detect road cracks. However, the evaluations were limited to small datasets, and there was a limited evaluation of real-world data, which is crucial for practical applications. Additionally, some studies lacked detailed explanations of the proposed CNNs, making it challenging to reproduce the results.

III. DATASETS

We have collected and annotated one of the most complex and diverse road damage datasets. Our dataset, called the Diverse Road Damage Dataset (DRDD), contains 1500 images of five damage types captured with GoPro cameras under various weather conditions and during daytime. The resolution of images in our dataset is 1920×1440 pixels. We have verified

with researchers and municipal planners that additional views would be convenient in the initial phase. The dataset contains five distinct classes with a real-world long-tailed class imbalance and unique class distribution, as shown in Figure 5. Moreover, the data annotation project involved 12 professionals with diverse backgrounds and required substantial person-hours. Furthermore, we conducted multiple rounds of quality assessment for each batch of data.

A. DATASET CHARACTERISTICS

1) DIVERSE CONDITIONS

DRDD is collected for diverse scenes and weather conditions. The images feature rich backgrounds, including trees, buildings, and other vehicles. Unlike current datasets, our DRDD contains images with more diverse object distributions across scenes. Lighting conditions also vary with the day, influencing pixel colors in the images. Similarly, the images are taken under clear and overcast conditions to capture variation. Weather conditions affect the type of damage and the background, making it challenging for many algorithms to distinguish weather-induced structures from the actual damage type. To further enhance generalization, we will extend DRDD across various urban, suburban, and rural regions in Oman. The extension also aims to encompass a range of pavement materials and to collect data across different seasons to capture diverse climatic effects, including rainfall patterns and thermal cycles. Additionally, we will pilot multimodal extensions using synchronized RGB, depth, and thermal cameras to address illumination and shadow artifacts.

2) DAMAGE DENSITIES

Our dataset contains various density scenarios, divided into three levels: fewer than two, two to three, and more than three types of defects. High densities are among the most crucial characteristics of our dataset. Table 1 compares related datasets widely used for damage detection in large-scale outdoor scenes. Our dataset has three notable evaluation parameters: i) the number of damage types per image, our dataset has the damages on average, which is more than the existing datasets, ii) the defect types distribution in the scene showing the ratio of road damage, where the statistics show ours exceeds others in density, and iii) the degree of damages by computing the mean number of damages in meters centered around each damage type.

3) OCCLUSIONS

DRDD is collected during regular traffic hours to ensure it reflects the actual scenes. However, the one downside is that the dataset's images have occlusions. Figure 1(b) shows an example of a scenario where the damage type is partially occluded, making the detection challenging due to limited visible features. According to occlusion situations, the datasets can test the techniques' performance in challenging cases.

4) INSTANCE-LEVEL DIVERSITIES

We show the diverse instance-level densities in Fig. 1(e). The inter-class difference between the damage types is significant. Due to the variation in shape and scale details, long-distance instances are difficult to learn, as damage types differ in size, structure, and appearance, making our datasets challenging for most existing algorithms.

5) DIVERSE DAMAGES TYPES

The road damage does not follow a specific pattern; hence, a single scene may contain more than one type of damage. This is also reflected in our datasets, where more than one damage type is present. Such images are shown in Fig. 1(c).

B. DATASET CHALLENGES

We have accomplished a scale much more extensive than any earlier datasets; the scale is essential for generalization, the ultimate goal of any dataset. However, employing the data across cities and countries raises several challenges, such as noise sources and ambiguities. We manage several known noise sources and ambiguities in our datasets, illustrated in Fig. 1.

1) LABELING NOISE

The accuracy of the labels is heavily dependent on the labeler, an expert and a domain specialist; then, it is quite probable that the label will be accurately compared to someone who is a volunteer and learning the process. Moreover, there are no agreed-upon definitions for damage types.

2) IMAGE QUALITY

The quality and resolution of the images acquired through various cameras and hand-held devices vary significantly. Occasionally, vehicles occlude the damage of the road from view, which we protect against by removing images manually, but the possibility of partial occlusion always exists. Furthermore, we also capture multiple views in our datasets to mitigate the occlusion issues.

3) UNLABELED ARTIFACTS

Some artifacts are not a damage type to the road but have some temporary or superficial marks, such as drag marks of the tires during braking, peeling of paint on the road, parts of unwanted items like a branch, trash, etc. These may appear similar to road damage types; nevertheless, the damage of interest is often the most prominent, and the presence of others can create confusion in detection and localization, which can misguide algorithms.

4) PATCHING NOISE

Road damages are usually fixed using manual techniques, but these repairs peel off quickly due to temporary materials. Such patching typically hides the structure of the actual damage, creating confusion for the labeler and leading to a discrepancy in the accuracy.

TABLE 1. Summary of road image datasets and research works.

Dataset	Methodology	Location	No. of Images	Image Resolution	Acquisition Camera	Smartphone	Moving Camera	View Top	Wide	Damage Type Cracks	Potholes	Label Type Pixel	Bbox	Image
Ouma and Hahn [44]	Traditional Algorithms	Kenya	75	1080×1920	✓	✓	-	✓	✓	✓	✓	✓	✓	✓
CrackIT [45]	Traditional Algorithms	-	84	1536×2048	✓	✓	-	✓	✓	✓	✓	✓	✓	✓
CFD [46]	Traditional Algorithms	China	118	480×320	✓	✓	-	✓	✓	✓	✓	✓	✓	✓
CrackTree200 [47]	Traditional Algorithms	-	206	800×600	✓	✓	-	✓	✓	✓	✓	✓	✓	✓
Weng et al. [48]	Traditional Algorithms	China	217	2048×1536	✓	✓	-	✓	✓	✓	✓	✓	✓	✓
SDNET2018 [49]	AlexNet	USA	230	4068×3456	✓	✓	-	✓	✓	✓	✓	✓	✓	✓
Li et al. [50]	Back Propagation NN	China	400	-	✓	✓	-	✓	✓	✓	✓	✓	✓	✓
Crack500 [51]	CNN	USA	500	2000×1500	✓	✓	-	✓	✓	✓	✓	✓	✓	✓
EdmCrack600 [52]	ConnCrack	Canada	600	1920×1080	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
GAPs v1 [53]	ASINVOSnet	Germany	1969	1920×1080	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
GAPs v2 [54]	ASINVOSnet, ResNet34	Germany	2468	1920×1080	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Majidifard et al. [41]	YOLOv2, Faster-RCNN	USA	7237	640×640	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Maeda et al. [37]	SSD Inception-v2, MobileNet	Japan	9053	600×600	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Angulo et al. [38]	RetinaNet	Various	18034	600×600	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
RDD2020 [43]	MobileNet	Various	26,620	600×600, 720×960	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
DRDD (Ours)	RTMDet with Attention4D	Oman	1500	1920×1440	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

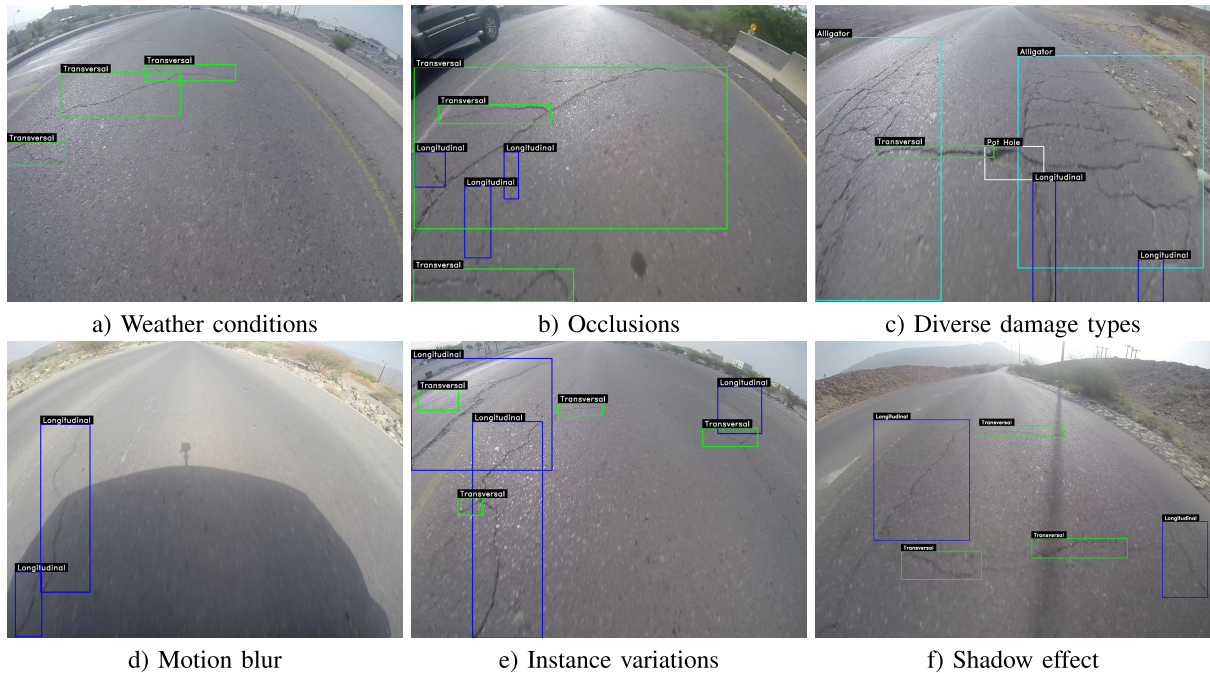


FIGURE 1. Major challenges in road damage detection: Representative examples showing various challenges: (a) weather conditions affecting damage visibility and appearance, (b) occlusions from vehicles partially hiding damage areas, (c) diverse damage patterns requiring robust detection capabilities, (d) motion blur from vehicle movement impacting image quality, and (e) instance-level variations in damage characteristics and (f) shadow effect adding complexity to detection.

5) UNWANTED SHADOWS

Shadows can obscure damage features, create false edges that may be misclassified as cracks, and reduce contrast in affected regions. This complicates the detection process as they vary throughout the day based on sun position, which are affected by surrounding infrastructure, including buildings, trees, and overpasses. They create dynamic challenges for consistent detection, especially if the sun is behind the vehicle taking pictures, the vehicle shadow may obscure part of the image, and it is possible that some sections of the damage type as presented in Fig. 1(d). Hence, it becomes challenging for the algorithms as some regions of the damage are concealed by shadow while others are under direct sunlight, causing a difference in pixels' color.

C. EVALUATION PROTOCOL

We have established an evaluation protocol to train/test our dataset's splits, which are expressed as follows.

1) PER-DATASET SPLITS

Per the traditional protocol, we are interested in how well an algorithm generalizes when the train and test splits belong to the same dataset. The training and testing sets are disjoint; here, the training is performed on a defined number of images and tested on another set of images from the same dataset to avoid overfitting a specific category.

D. ANNOTATIONS

We manually labeled high-quality ground truth for the images of both our datasets. We annotated the road damage type using a 2D bounding box (x, y, w, h) , where x, y denote the center coordinates and w, h is the width and height along the x-axis and y-axis, respectively. A file with the same ID as the image is provided for each damage type.

E. TYPES OF ROAD CRACKS

The damage types can be categorized into Alligator Cracks, Block Cracks, Longitudinal Cracks, Pothole Cracks, and

Transverse Cracks. These are the most prominent ones and are usually found on most roads across the globe. Figure 1 shows these mentioned cracks.

1) POTHOLE

Potholes are holes caused by weather conditions, heavy traffic, and surface quality degradation. Initially small, these potholes can rapidly expand due to ongoing vehicle traffic that erodes the asphalt further and water from rain or floods that washes away additional material. Proper maintenance and repair of road cracks can help prolong a road's life and improve drivers' safety. Repair methods can include sealing the cracks, patching the affected areas, and repaving the entire road surface in specific extreme cases.

2) LONGITUDINAL CRACKS

Expansion and contraction of the ground, poor drainage, and the settling of the subgrade cause longitudinal cracks. These cracks run along the length of the road and can be found on both the road's shoulder and the center line of the road. The longitudinal cracks can also be formed by asphalt surface contraction as well as its exposure to the sun light and heat.

3) TRANSVERSE CRACKS

Cracks on asphalt surface that run across the road. They typically caused by temperature fluctuation and heavy traffic. These cracks are often found on the wheel path.

4) ALLIGATOR OR FATIGUE CRACKS

These are a series of interconnected cracks that resemble the skin of an alligator. They are caused by a failure of the asphalt surface and heavy traffic loads. They occur when the asphalt surface can no longer support the weight of the traffic and begins to fatigue and deform. These cracks are also known as "fatigue cracking" and can be found on both the shoulder and the wheel path.

5) BLOCK CRACKS

Large, rectangular cracks break the surface into chunks or blocks. It forms when the asphalt surface is too thick and the top layer contracts at a different rate than the bottom, causing the surface to crack.

IV. METHODOLOGY

Due to its superior performance on the COCO dataset and its compatibility with edge computing devices, we selected the Real-Time Model Detection (RTMDet) lightweight object detection model as our baseline. The model has achieved outstanding performance in one-stage real-time object detection and instance segmentation tasks. As illustrated in Fig. 3(a), the RTMDet architecture consists of three main parts: the backbone, neck, and head. Recent studies typically adopted CSPDarkNet [52] as a backbone architecture in RTMDet; however, we employed CSPNeXt, an architecture that combines the ResNeXt Network and the Cross Stage Partial Network (CSP), as our backbone network [53] as

the architecture utilizes large-kernel, depth-wise separable convolutions in the base modules. Our choice is based on the observation that the model quickly learns the global context by making the network deeper and wider.

The backbone model is available in two options. The first choice is the smaller architecture, denoted as P₅, while the second version, P₆, is the larger variant. We have opted for the former version, which produces five feature map scales, denoted as C₁, C₂, C₃, C₄, and C₅, after downsampling by factors of 2^k , where $k \in \{1, 2, \dots, 5\}$. The Neck module performs feature fusion using the final three feature map scales of C₃ (i.e., 8, 256, 80, 80), C₄ (i.e., 8, 512, 40, 40), and C₅ (i.e., 8, 1024, 20, 20). Moreover, the neck module further incorporates a multiscale feature pyramid network (CSPNeXtPAFPN) using the three feature maps from the backbone with both top-down and bottom-up feature propagation [54], [55] for enhancing the features before passing them into the head for detection and classification.

Although the CSPNeXtPAFPN architecture's top-down and bottom-up pathways are effective, they may only partially exploit the potential relations between features at different scales. This may potentially lead to a loss of fine-grained details. The neck part of RTMDet is responsible for fusing and refining multiscale features from the backbone. Hence, we propose and integrate Attention4D blocks. By exploiting this, the mentioned model can capture and utilize local and global contextual information more effectively across different scales. Eventually, the feature map of each scale is then employed by the detection head to anticipate object-bounding boxes and their categories. This architecture works well for general and rotated objects and segmentation if extended by adding kernel and mask feature generation heads [56].

A. Attention4D MODULE

The core innovation in our model is the Attention4D block, depicted in Fig. 3(b). This block processes input feature maps through a series of operations. A 1×1 convolution is followed by batch normalization to adjust the channel dimension. The output is split into three branches: Value (V), Transposed Key K^T , and Query (Q). The Key and Query branches undergo element-wise multiplication. The result combines positional encoding (PosE) and a "Talking Head" input. A softmax operation normalizes the attention weights. The softmax output is multiplied element-wise by the value branch and another "Talking Head" input. Finally, the result passes through another 1×1 convolution and batch normalization. This Attention4D block allows the network to capture local and global contextual information, potentially improving the model's understanding of complex scenes. The Attention4D blocks are strategically integrated at crucial points in both top-down paths of the neck architecture.

In the top-down path, an Attention4D block processes the input from the highest-level feature map before further refinement. We apply a second Attention4D block after the

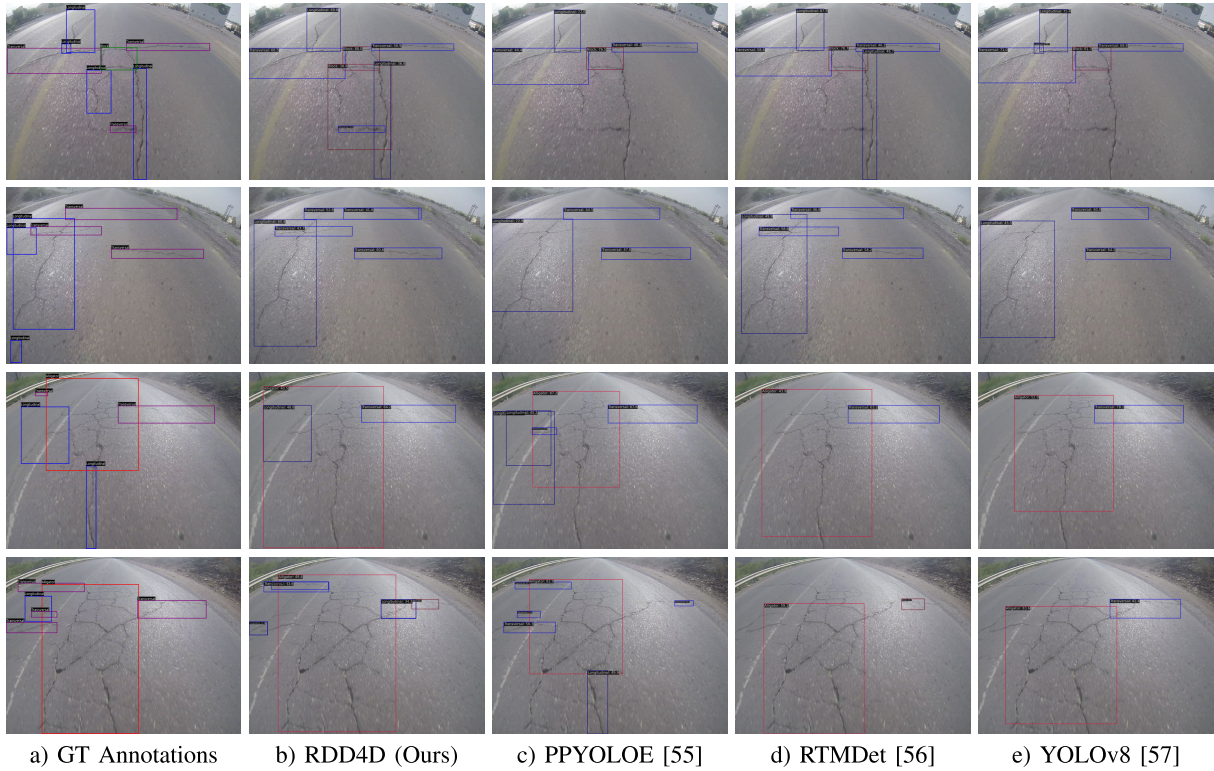


FIGURE 2. Qualitative comparison of road damage detection using different object detection techniques. The columns present (from left to right) ground-truth annotations, our proposed approach, RTMDet, and YOLOv8, applied to four representative test images. Each row shows the same input image processed by these techniques, enabling direct comparison of detection accuracy and precision. The results demonstrate that our proposed methodology achieves more accurate detection and classification of road damage categories compared to the baseline RTMDet and other common approaches.

first top-down CSPLayer. The bottom-up path feeds the output of the Attention4D block from the top-down path into a downsample operation before it reaches the bottom-up CSPLayers. This integration allows the network to refine features using our attention mechanism at multiple scales, potentially capturing fine-grained details and high-level semantic information. Moreover, we conducted a technical comparison between our Attention4D mechanism and two recent attention frameworks: Bilateral Feature Fusion with Hexagonal Attention [] and Lightweight Transformer-driven Trapezoidal Attention (LT-MSTAN). While BFF-HA integrates five specialized attention modules (Edge Attention, Non-Local Attention, Quad-Feature Spatial Attention, Multi-Head Self-Attention, and Weighted Channel Attention) in a hexagonal arrangement for saliency detection under visual uncertainties, and LT-MSTAN uses a hierarchical trapezoidal structure with scale-specific attention for lightweight processing, our Attention4D module is uniquely optimized for road damage detection. Unlike these approaches that target general saliency or computational efficiency, Attention4D integrates positional encoding with ‘Talking Heads’ attention, enabling inter-head communication that captures relationships between different crack scales. Moreover, our design philosophy aligns with recent efficient attention mechanisms for resource-constrained devices.

B. LOSS FUNCTION

A crucial step in training one-stage object detectors is matching dense predictions across various scales with the ground truth bounding boxes. This process, known as label assignment, has evolved by developing several strategies [57], [58], [59]. While recent approaches have focused on dynamic methodologies for label assignment [60], [61], [62], we have adopted the dynamic soft label assignment strategy [63], which uses cost functions aligned with the training loss as the matching criteria to assign labels to the predicted bounding boxes more accurately. This dynamic soft label assignment strategy is based on the SOTA framework [62]. The cost function is defined as follows:

$$\ell = \lambda_1 \cdot \delta + \lambda_2 \cdot \theta + \lambda_3 \cdot \rho, \quad (1)$$

where $\lambda_1 = 1$, $\lambda_2 = 3$, and $\lambda_3 = 1$ are the default weighting factors. Furthermore, δ , θ , ρ , and ℓ represent the classification, location, center proximity, and overall costs. In traditional approaches, the δ often relies on binary labels. This can lead to scenarios where predictions with high classification scores but inaccurate bounding boxes receive low classification costs and vice versa. A soft label assignment strategy is proposed instead of binary labels to address this. We calculate the modified classification cost as

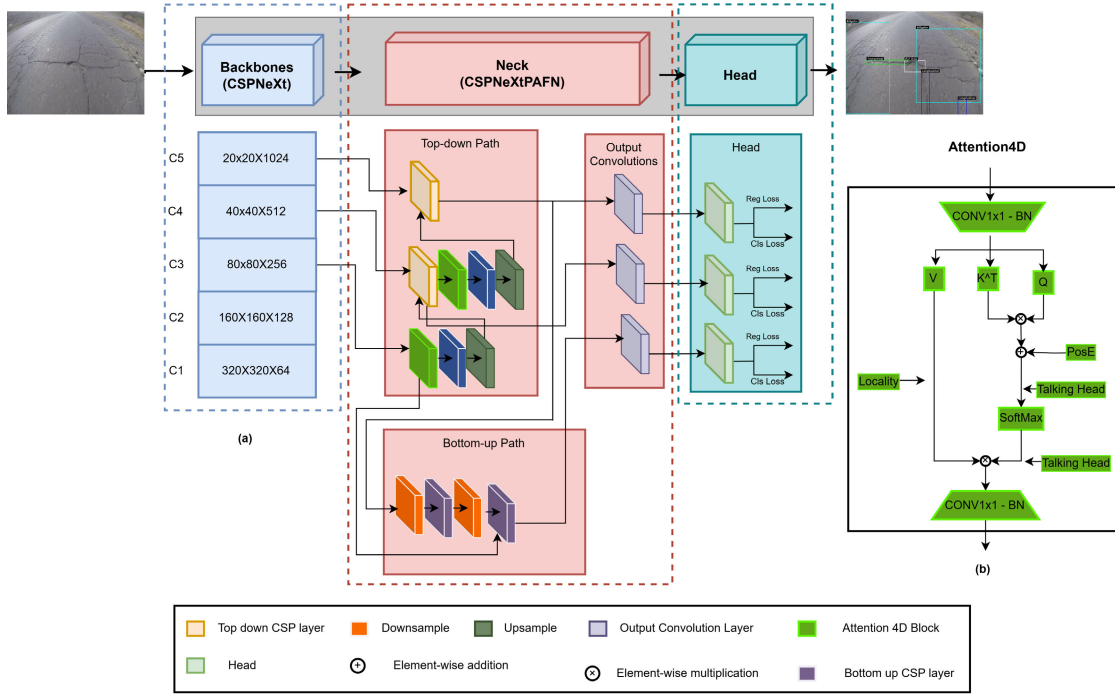


FIGURE 3. Our proposed architecture: The flow of feature maps through the neck section of the object detection network. Attention4D blocks (highlighted in bold green) are strategically applied to the feature maps from the backbone, enhancing spatial and channel-wise information. The neck processes these enhanced features through top-down and bottom-up paths, creating a multi-scale feature representation.

TABLE 2. Performance comparison of different methods (AP).

Methods	Alligator	Block	Longitudinal	Transversal
YOLOv8 [57]	0.122	0.050	0.242	0.145
YOLOV7 [69]	0.172	0.200	0.380	0.267
YOLOV6 [69]	0.041	0.003	0.190	0.205
PPYOLOE [55]	0.012	0.100	0.206	0.131
RTMDcT [56]	0.127	0.111	0.379	0.350
YOLOX [68]	0.143	0.400	0.255	0.000
RDD4D (Ours)	0.145	0.900	0.387	0.352

follows:

$$\delta = \text{CE}(\hat{y}, y) \cdot (y - \hat{y})^2, \quad (2)$$

where CE stands for cross entropy and $\hat{y}$ is the prediction. The dynamic aspect of this approach means that label assignments can change during training, allowing the model to focus on the most informative examples. SoftLabel, inspired by GFL [67], uses intersection-over-union (IoU) as a soft label to provide a more detailed evaluation of prediction accuracy. For the θ , existing methods using generalized IoU [68] often lack sufficient differentiation between high- and low-quality matches. The soft label approach uses a logarithmic IoU scale $\theta = -\log(\text{IoU})$, amplifying the differences between good and poor matches, especially for lower IoU values. This helps the model distinguish between high- and low-quality predictions during training.

Finally, the center proximity cost replaces rigid center-based criteria [61], [62], [69] with the following formula:

$$\rho = \frac{\eta}{|\hat{y}_c - y_c| - \epsilon}, \quad (3)$$

where η and ϵ are adjustable parameters, while $\hat{y}_c$ is the predicted center and y_c is the actual center. In the overall cost function shown in Eq. 1, the center cost encourages predictions that are accurate in classification and bounding box regression and are also well-centered on the object. For example, if two predictions have similar classification confidence and bounding box overlap, the prediction closer to the center of the ground truth object will be favored during training.

V. EXPERIMENTS

A. SETTINGS

1) TRAINING DETAILS

During training, we used a batch size of four in our experiments. We set the weight decay at 0.9, suggesting a considerable regularization impact to counteract overfitting, although this number is more significant than usual. We set the momentum at 5×10^{-4} , compared to regular values, which typically range between 0.8 and 0.99. We train the model for 300 epochs to ensure complete performance optimization.

2) DATASETS

We used commonly available GoPro cameras, which have similar video quality to many widely available and affordable consumer mobile phones when compared to specialized road inspection tools. This approach makes our method reproducible and accessible for agencies with limited resources. These devices deliver high-resolution images (up to 5K), wide-angle lenses ideal for photographing road

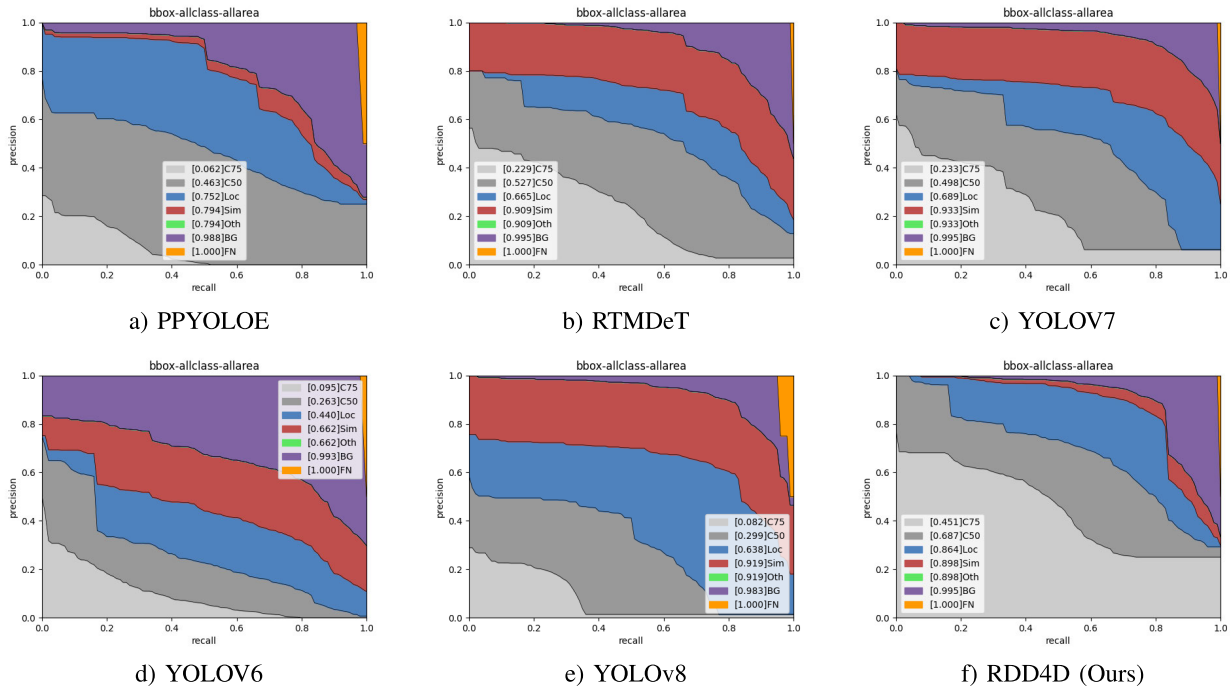


FIGURE 4. Precision-recall curves comparing different methods for bounding box prediction across all classes in terms of all-area performance on a) PPYOLOE, b) RTMDeT, c) YOLOV7, d) YOLOV6, e) YOLOV8, and f) Ours. The error analysis methods include C75 (gray), C50 (light gray), Loc (blue), Sim (red), Oth (green), BG (purple), and FN (orange), with their respective average precision scores shown in brackets. (Best viewed in color).

TABLE 3. Scale-specific performance comparison ($AP_S/AP_M/AP_L$).

Methods	Alligator			Block			Longitudinal			Transversal		
	S	M	L	S	M	L	S	M	L	S	M	L
YOLOv8 [57]	-1.000	-1.000	0.052	-1.000	-1.000	0.053	-1.000	0.076	0.252	-1.000	0.091	0.153
YOLOV7 [69]	-1.000	-1.000	0.172	-1.000	-1.000	0.200	-1.000	0.105	0.397	-1.000	0.149	0.281
YOLOV6 [56]	-1.000	-1.000	0.041	-1.000	-1.000	0.003	-1.000	0.062	0.200	-1.000	0.153	0.214
PPYOLOE [55]	-1.000	-1.000	0.012	-1.000	-1.000	0.023	-1.000	-1.000	0.100	-1.000	0.078	0.215
RTMDeT [56]	-1.000	-1.000	0.127	-1.000	-1.000	0.113	-1.000	0.124	0.420	-1.000	0.122	0.375
YOLOX [68]	-1.000	-1.000	0.143	-1.000	-1.000	0.400	-1.000	0.011	0.271	-1.000	0.000	0.000
RDD4D (Ours)	-1.000	-1.000	0.145	-1.000	-1.000	0.900	-1.000	0.272	0.699	-1.000	0.132	0.379

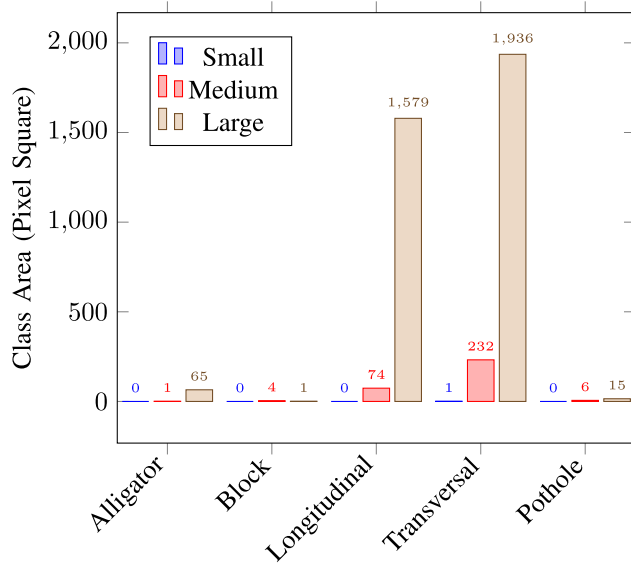
surfaces, and image stabilization features that minimize motion blur during vehicle-based data collection. They can be simply mounted on standard vehicles without requiring special installation or vehicle modifications. Additionally, these cameras do not require expensive software licenses or ongoing subscription fees. Our work demonstrates that effective road damage detection is possible without the need for expensive, specialized equipment.

The road crack dataset videos were recorded using a GoPro camera mounted on the front of a car and aimed at the road surface. As the vehicle traversed various roads, it recorded multiple types of cracks under varying lighting conditions, comprehensively representing different road conditions. A civil engineer carefully reviewed and validated each road condition to accurately identify and confirm the presence of various types of cracks. We recorded 30 video segments in MP4 format, each with a resolution of 1920×1440 pixels and a frame rate of 30 fps. Most of the video segments lasted between 20 and 30 minutes. These

videos were then converted into frames, with blurry images due to car speeds discarded to maintain quality. We have used the Computer Vision Annotations Tool to annotate images in Pascal-VOC and COCO formats. From the video footage, we annotated a total of 1,500 images. Figure 5 presents the distribution of the dataset. Some video frames contained up to 45 annotations, demonstrating the complex and cluttered nature of various road cracks within a single frame. The COCO metrics categorize the objects in the dataset into three sizes based on the area of their bounding boxes. Small objects occupy less than 32×32 pixels, medium objects are between 32×32 pixels and 96×96 pixels, and large objects exceed 96×96 pixels. According to our ground-truth analysis, as detailed in Fig. 5, most annotated objects are classified as large or medium based on COCO metrics. The COCO-scale analysis reveals a scarcity of small objects, which explains why the AP for small objects (APS) is undefined across models. This indicates an absence of data, not poor performance.

TABLE 4. Detection results (mAP) on the road-crack dataset. The best results are shown in bold and the second best underlined.

Methods	Backbone	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	AR	AR _S	AR _M	AR _L
YOLOv8 [57]	YOLOv8CSPDarknet	0.122	0.299	0.082	-1.000	0.083	0.127	0.448	-1.000	0.234	0.454
YOLOv7 [69]	YOLOv7Backbone	0.255	0.498	<u>0.233</u>	-1.000	0.127	0.263	0.547	-1.000	0.351	0.553
YOLOv6 [69]	YOLOv6Backbone	0.110	0.263	0.095	-1.000	0.108	0.114	<u>0.560</u>	-1.000	0.460	0.572
PPYOLOE [55]	PPYOLOECSPResNet	0.112	0.463	0.062	-1.000	0.079	0.117	0.322	-1.000	<u>0.388</u>	0.325
RTMDet [56]	CSPNeXt	<u>0.268</u>	<u>0.527</u>	0.229	-1.000	<u>0.123</u>	<u>0.280</u>	0.517	-1.000	0.373	<u>0.623</u>
YOLOX [68]	YOLOXCSPDarknet	0.200	0.377	0.188	-1.000	0.006	0.204	0.288	-1.000	0.033	0.386
Ours	CSPNeXt	0.446	0.687	0.451	-1.000	0.113	0.458	0.675	-1.000	0.277	0.690

**FIGURE 5.** The graph presents a ground-truth analysis of bounding box areas, where each bar represents the count of annotations. These annotations are categorized into three distinct groups—small, medium, and large—based on the size of their bounding box areas.

3) EVALUATION METRICS

To evaluate the performance of our object detectors, we have used the COCO evaluation metrics, a standard metric for assessing both object detection and segmentation models. The COCO evaluation framework provides a detailed analysis of model performance across various Intersection over Union (IoU) thresholds. It categorizes object detection results based on different object sizes: small, medium, and large.

VI. RESULTS AND DISCUSSIONS

This section presents a comprehensive evaluation of our proposed RDD4D model through quantitative analysis and qualitative interpretation of findings. We organize our evaluation into three main components: First, we evaluate our model's performance on the proposed DRDD dataset through detailed quantitative and qualitative comparisons (Section VI-A). Second, we validate our method by testing it on the CrackTinyNet benchmark dataset (Section VI-B). Finally, we conduct extensive ablation studies to understand the contribution of individual architectural components (Section VI-C).

A. COMPARISONS ON OUR PROPOSED DRDD DATASET

Precision and recall are essential metrics used to evaluate a model's performance. Precision measures the accuracy of the detections, which is the proportion of correctly predicted positive identifications. A high precision rate indicates that the model made few mistakes in its identifications, making it reliable when predicting the presence of an object. In contrast, recall evaluates how effectively the model identifies all true positive detections in the dataset. It reflects the proportion of true positive detections the model makes relative to the total number of positives in the data. High recall indicates that the model leaves few actual positives undetected, maximizing the chances that if an object is present, the model will recognize it. A balanced performance in both metrics is crucial for reliable road damage detection systems. The performance of various deep learning models was evaluated: YOLOv8 [54], YOLOv7 [65], YOLOv6 [70] PPYOLOE [66], YOLOX [62], and RTMDet [63], to detect small, medium, and large road cracks using the metrics of Average Precision (AP) and Average Recall (AR).

A classwise AP is given in Table 2, and detailed scale-specific analysis across different damage types is presented in Table 3. The value of -1.000 for small (S) and medium (M) sizes in Alligator and Block cracks, and small sizes in Transversal cracks, indicates the absence of these scale variations in our dataset. Our method demonstrates superior performance in detecting large-scale damages across all categories, achieving the highest AP scores for Alligator (0.145), Block (0.900), Longitudinal (0.699), and Transversal cracks (0.379). The detailed experimental results at different thresholds are shown in Table 4.

The AP and AR scores of -1.000 for small-sized damages (AP_S and AR_S) across all models indicate the absence of small damages in the dataset, rather than poor detection performance. This observation confirms that most road defects in our dataset belong to a large category. YOLOv7 achieved the highest AP of 0.127 for medium-sized damages, followed by RTMDet with 0.123. In the large damage category, where most of our dataset's instances are concentrated, our RDD4D method significantly outperformed others with an AP of 0.458, while RTMDet showed the second-best performance with 0.280.

Regarding recall performance, YOLOv6 demonstrated strong capability in detecting medium-sized damages with the highest AR of 0.460, followed by PPYOLOE with 0.388. For large damages, which constitute the majority of our

dataset, our method achieved the best AR of 0.690, with RTMDeT following at 0.623. Overall, our proposed method demonstrates superior performance with the highest AP (0.446), AP50 (0.687), and AP75 (0.451) scores across all models, showing significant improvements over the second-best performer RTMDeT (AP=0.268, AP50=0.527). These results validate the effectiveness of our approach, particularly in detecting the predominant large-sized road damages.

Regarding recall, all models struggled with small objects, scoring an AR of ~ 1.000 . This is due to the absence of small cracks in the dataset. YOLOv6 achieved the highest AR of 0.46 for medium-sized objects, demonstrating its effectiveness in consistently recognizing the presence of medium-sized cracks more accurately than its counterparts, despite its lower precision. Our model showed the highest recall for large road cracks with an AR of 0.690, confirming its robustness in terms of precision and its ability to reliably identify a higher proportion of large road cracks present in the dataset.

A comprehensive analysis of various error distributions between our model and other models is presented in Fig. 4. Detection errors in road damage assessment can be categorized into several types: localization errors (Loc) when bounding boxes are imprecisely placed, similarity-based (Sim) confusions between visually similar damage types, category confusion (Oth) where damages are misclassified, background false positives (BG) where normal road surfaces are mistakenly identified as damage, and false negatives (FN) where actual damages go undetected. Our model achieves higher confidence scores (C75=0.451, C50=0.687) compared to YOLOv6 (C75=0.095, C50=0.263), YOLOv7 (C75=0.233, C50=0.498), and YOLOv8 (C75=0.082, C50=0.299). The localization accuracy (Loc=0.864) is superior to all other models, indicating more precise bounding box predictions. Our model maintains better precision across higher recall values, with a more balanced distribution of similarity-based and category confusion errors (both Sim and Oth=0.898). While background detection errors (BG=0.995) and false negatives (FN=1.000) are present, they occupy a relatively smaller portion of the graph, particularly at high recall values. The smoother curves and more gradual degradation in precision as recall increases suggest a more robust and stable detection performance. This comprehensive improvement across all error categories demonstrates that our model achieves better absolute performance and handles various detection challenges more effectively than existing approaches.

Figure 2 presents a qualitative comparison of road damage detection across four representative test images. Each row displays the same input image processed by different methods: ground-truth annotations, our proposed approach, RTMDeT, and YOLOv8, enabling direct assessment of detection accuracy and precision. The visual comparison demonstrates that our proposed methodology achieves superior detection and classification performance compared to baseline methods. Our approach exhibits particular strength

in accurately localizing and classifying various damage types, with notably robust performance on large-scale road damage, which constitutes the majority of real-world cases. The side-by-side comparison clearly illustrates our model's enhanced capability in handling diverse damage categories and scales encountered in practical road inspection scenarios.

B. COMPARISONS ON CrackTinyNet DATASET

We have not included the RDD2022 dataset [71] due to its closed test set and lack of public annotations, which excludes fair and reproducible benchmarking for new entries. Instead, we use CrackTinyNet, an open benchmark with diverse defect categories to substantiate cross-dataset generalization. Our proposed method demonstrates significant improvements over existing approaches on CrackTinyNet [72] as shown in Table 6. The dataset categorizes road surface defects into seven distinct classes: three crack types (Longitudinal-D00, Transverse-D10, and Alligator-D20), surface deformation (Pothole-D40), two types of marking deterioration (White Line-D43 and Cross Walk-D44 Blur), and infrastructure elements (Manhole Cover-D50). With a MAP@.5 score of 0.825, our model substantially outperforms the previous best performer, CrTNet [72] (0.601), on this dataset, representing a 37% improvement in detection accuracy. This enhancement is particularly noteworthy when compared to widely used object detection frameworks such as YOLOv8 [54] (0.445) and EfficientDet [73] (0.552). Our model makes fewer false positive detections, which is crucial for real-world applications where false alarms could lead to unnecessary maintenance inspections. Traditional models like YOLOv5s [74] and SSD [75] show considerably lower precision (0.27 and 0.31, respectively). Perhaps most impressively, our model achieves a recall rate of 0.98, demonstrating its ability to detect almost all instances of road damage in the dataset. This is a substantial improvement over CrTNet's 0.61 recall and far exceeds other benchmarks like Fast R-CNN [76] (0.55) and YOLOv8 [54] (0.53).

TABLE 5. Road damage categories and their descriptions with classwise Average Precision (AP) results for DRDD.

Class	Damage Type	AP@[IoU=0.50:0.95]	AP@[IoU=0.50]
D00	Longitudinal Crack	0.829	0.833
D10	Transverse Crack	0.916	0.928
D20	Alligator Crack	0.863	0.863
D40	Pothole	0.577	0.609
D43	White Line Blur	0.689	0.697
D44	Cross Walk Blur	0.944	0.945
D50	Manhole Cover	0.853	0.903

The precision-recall graph in Fig. 6 shows excellent model performance with uniform confidence scores (C75 and C50 at 0.825) across different IoU thresholds. The dominant red region representing similarity errors (Sim=0.996) indicates that most detection errors stem from confusion between similar damage types rather than localization issues (Loc=0.825) or background misidentifications (BG=0.997). The model maintains high precision across most recall values, with performance degrading only at very high recall

levels. The classwise performance analysis in Table 5 shows that our model excels at detecting Cross Walk Blur (D44) and Transverse Crack (D10), achieving AP scores above 0.90 across both IoU thresholds. The detection of Alligator Crack (D20), Manhole Cover (D50), and Longitudinal Crack (D00) also demonstrates strong performance with AP scores above 0.82. However, the model shows relatively lower accuracy in detecting White Line Blur (D43) and Potholes (D40). Potholes are the most challenging category at AP scores of 0.577 and 0.609, likely due to their irregular shapes and varying appearances.

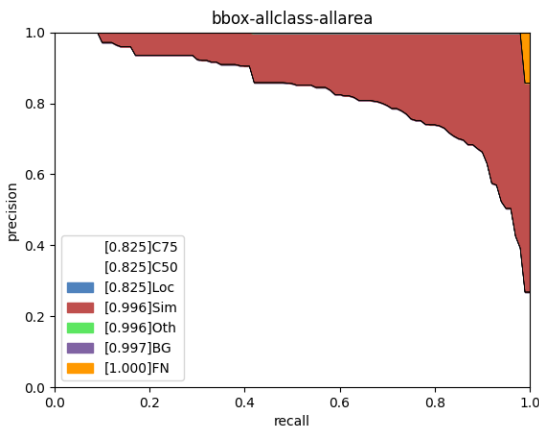


FIGURE 6. The graph demonstrates various error types affecting detection performance: localization accuracy at IoU thresholds of 0.75 (C75) and 0.50 (C50), localization errors (Loc), similarity-based false positives (Sim), other category confusions (Oth), background detections (BG), and false negatives (FN).

TABLE 6. Quantitative comparison of detection performance between our RDD4D and existing state-of-the-art methods evaluated on the CrackTinyNet [72] dataset. The best results are shown in bold and the second best underlined.

Name	MAP@.5	Precision	Recall
YOLOv5s	0.381	0.27	<u>0.77</u>
YOLOv5x	0.404	0.45	0.51
SSD	0.406	0.31	0.58
Fast R-CNN	0.443	0.42	0.55
YOLOv8	0.445	0.43	0.53
EfficientDet	0.552	0.38	0.69
CrTNet	<u>0.601</u>	<u>0.62</u>	0.61
Ours	0.825	0.81	0.98

C. ABLATION STUDIES

The ablation study on Attention4D block placement strategies, as shown in Table 7, reveals that utilizing blocks in the top-down path yields superior performance (mAP=0.446, AP50=0.687, AP75=0.451) compared to the bottom-up path (mAP=0.412) or single block placement (mAP=0.398). While implementing blocks in both paths marginally improves performance (mAP=0.455), it comes at the cost of reduced inference speed (25.1 FPS vs 26.8 FPS) and increased parameter count (40.2M vs 38.5M). This suggests that the top-down path placement provides the optimal balance between detection accuracy and computational efficiency. Further analysis of the number of Attention4D blocks

in Table 8 demonstrates a clear performance trend. While a single block shows modest results (mAP=0.425), implementing two blocks achieves significantly better performance (mAP=0.446) with a reasonable parameter increase of 1.3M. Although adding three or four blocks marginally improves accuracy (mAP=0.449 and 0.451 respectively), the computational overhead becomes substantial, reducing the FPS to 25.3 and 23.7 respectively, while increasing the parameter count to 39.8M and 41.1M. These results validate our choice of using two Attention4D blocks in the top-down path as the optimal configuration, providing the best trade-off between detection accuracy and computational efficiency. Our 2-block configuration requires 81.2 GFLOPs, representing only a 3.6% increase over the single-block baseline (78.4 GFLOPs), while achieving significant mAP improvement from 0.425 to 0.446. This finding is consistent with recent work on efficient attention mechanisms for consumer imaging devices [4], where layer-frozen assisted dual attention achieves competitive performance while minimizing computational overhead.

Our comprehensive evaluation provides critical insights into automated road damage detection systems. Precision recall analysis reveals that similarity-based confusions between damage types, rather than localization errors, constitute the primary detection challenge, suggesting that future research should prioritize enhanced discriminative features for fine-grained classification. Our scale-specific analysis confirms that large-scale damage detection remains the predominant real-world challenge, with our model achieving superior performance (APL=0.458). The substantial improvement on CrackTinyNet (MAP@.5 of 0.825 vs. 0.601 for previous best) validates the generalizability of our approach across different datasets. Finally, our ablation studies demonstrate that two Attention4D blocks in the top-down path provide the optimal balance between accuracy (mAP=0.446) and real-time performance (26.8 FPS), offering practical guidance for deploying automated inspection systems on resource-constrained edge devices.

TABLE 7. Performance comparison of different placement strategies in RDD4D. The best results are provided when both paths are used.

Placement Strategy	mAP	AP50	AP75	FPS	Params(M)	GFLOPs
Only Top-down Path	0.446	0.687	0.451	26.8	38.5	81.2
Only Bottom-up Path	0.412	0.654	0.423	26.5	38.5	81.2
Single Block at End	0.398	0.642	0.412	27.2	37.8	78.4
Both Paths	0.455	0.693	0.460	25.1	40.2	84.8

TABLE 8. Impact of number of Attention4D blocks on performance of RDD4D. Using more blocks increases performance.

Blocks (Attention4D)	mAP	AP50	AP75	FPS	Params(M)	GFLOPs
1	0.425	0.668	0.432	27.5	37.2	78.4
2	0.446	0.687	0.451	26.8	38.5	81.2
3	0.449	0.689	0.453	25.3	39.8	84.1
4	0.451	0.690	0.455	23.7	41.1	87.3

VII. CONCLUSION

This paper addressed the challenge of multi-type road damage detection by introducing RDD4D, an enhanced detection

architecture with an integrated Attention4D mechanism. We also collected a novel dataset that captures diverse road damage types, with most images containing different damage types, serving as a valuable resource for the research community.

While our work demonstrates promising results, it has several limitations. Our current model architecture presents challenges for edge deployment due to its computational requirements, particularly as the Attention4D module increases the parameter count when used in both paths of the architecture. Environmental factors, such as shadows, pose challenges by obscuring features, creating false edges, and reducing contrast, which complicates detection. These shadows vary throughout the day based on sun position and are affected by surrounding infrastructure such as buildings, trees, and overpasses. It creates dynamic challenges for consistent detection. Furthermore, dataset size constraints, lighting variations, weather conditions, difficulties in detecting certain damage types or handling occluded damage, and limited geographic diversity represent areas requiring further investigation.

Future work will focus on three key directions. First, we will optimize RDD4D for edge deployment by applying model compression techniques, such as knowledge distillation and quantization, to reduce model size while maintaining detection accuracy. To facilitate practical deployment, we will convert our trained model to hardware-optimized formats, including ONNX for cross-platform compatibility, TensorRT for NVIDIA Jetson devices, NCNN for ARM-based platforms, and TensorFlow Lite for Edge TPU deployment. Second, we will expand the DRDD dataset across additional geographic regions and integrate multimodal sensors, including depth and thermal imaging, to address environmental challenges. To handle real-world noise factors, we will implement illumination normalization using Retinex-based preprocessing to reduce lighting artifacts, introduce ‘patch repair’ as an explicit class in our training dataset, and develop temporal aggregation across video frames to handle occlusion. We will implement semantically-aware augmentation strategies: (1) copy-paste augmentation to extract small/medium damage instances and composite them onto different road backgrounds while preserving orientation and class labels, (2) scale augmentation through multi-scale training with random scale jittering ($0.5\text{--}1.5\times$) to synthetically create smaller instances, (3) focal loss with class-aware weighting, and (4) color jittering and brightness adjustments that preserve spatial structure while increasing sample diversity. Third, we will develop severity-assessment capabilities for maintenance prioritization, integrate them with asset management systems, and conduct temporal analysis to track damage progression. Finally, we will provide practical insights for decision-makers, including cost-benefit analysis frameworks, deployment considerations for agencies of varying sizes and budgets, training requirements, and integration pathways with existing inspection processes

ACKNOWLEDGMENT

The authors are grateful to Prof. Rabie Ramadhan and Prof. Ahmad Salah for their valuable feedback and comments that improved this manuscript.

REFERENCES

- [1] D. Arya, H. Maeda, S. K. Ghosh, D. Toshniwal, A. Mraz, T. Kashiyama, and Y. Sekimoto, “Deep learning-based road damage detection and classification for multiple countries,” *Autom. Construction*, vol. 132, Dec. 2021, Art. no. 103935.
- [2] A. Alfarrarjeh, D. Trivedi, S. H. Kim, and C. Shahabi, “A deep learning approach for road damage detection from smartphone images,” in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 5201–5204.
- [3] M. Carlier, “Highway mileage—United States 1990–2020,” Statista, Apr. 2022.
- [4] *Study: Pothole Damage Costs U.S. Drivers \$3B a Year*, Wells Media Group, Inc., Mar. 2016. Accessed: Dec. 15, 2022. [Online]. Available: <https://www.insurancejournal.com/news/national/2026/02/10/857605.htm>
- [5] G. Malkoc, “The importance of road maintenance,” *World Highways* (Route One Publishing Ltd), Jul. 2015.
- [6] Q. Li, M. Yao, X. Yao, and B. Xu, “A real-time 3D scanning system for pavement distortion inspection,” *Meas. Sci. Technol.*, vol. 21, no. 1, Jan. 2010, Art. no. 015702.
- [7] X. Yu and X. Yu, “Vibration-based system for pavement condition evaluation,” in *Applications of Advanced Technology in Transportation*. American Society of Civil Engineers, 2006, pp. 183–189.
- [8] A. Zhang, K. C. P. Wang, Y. Fei, Y. Liu, C. Chen, G. Yang, J. Q. Li, E. Yang, and S. Qiu, “Automated pixel-level pavement crack detection on 3D asphalt surfaces with a recurrent neural network,” *Comput.-Aided Civil Infrastructure Eng.*, vol. 34, no. 3, pp. 213–229, Mar. 2019.
- [9] H. Maeda, Y. Sekimoto, T. Seto, T. Kashiyama, and H. Omata, “Road damage detection and classification using deep neural networks with smartphone images,” *Comput.-Aided Civil Infrastruct. Eng.*, vol. 33, no. 12, pp. 1127–1141, Dec. 2018.
- [10] L. Zhang, F. Yang, Y. D. Zhang, and Y. J. Zhu, “Road crack detection using deep convolutional neural network,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3708–3712.
- [11] H. S. Lee and K. Kim, “Simultaneous traffic sign detection and boundary estimation using convolutional neural network,” *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 5, pp. 1652–1663, May 2018.
- [12] K. Castañeda, O. Sánchez, R. F. Herrera, E. Pellicer, and H. Porras, “BIM-based traffic analysis and simulation at road intersection design,” *Autom. Construction*, vol. 131, Nov. 2021, Art. no. 103911.
- [13] M. Tahir and S. Anwar, “Transformers in pedestrian image retrieval and person re-identification in a multi-camera surveillance system,” *Appl. Sci.*, vol. 11, no. 19, p. 9197, Oct. 2021.
- [14] K. J. Joseph, S. Khan, F. S. Khan, and V. N. Balasubramanian, “Towards open world object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5826–5836.
- [15] P. Marcelino, M. D. Lurdes Antunes, and E. Fortunato, “Comprehensive performance indicators for road pavement condition assessment,” *Struct. Infrastruct. Eng.*, vol. 14, no. 11, pp. 1433–1445, Nov. 2018.
- [16] S. C. Radopoulou and I. Brilakis, “Detection of multiple road defects for pavement condition assessment,” in *Proc. 22nd Int. Workshop Intell. Comput. Eng. (EG-ICE)*. Eindhoven University of Technology, 2015.
- [17] B. Akarsu, M. Karaköse, K. Parlak, E. Akin, and A. Sarimaden, “A fast and adaptive road defect detection approach using computer vision with real time implementation,” *Int. J. Appl. Math., Electron. Comput.*, vol. 2016, no. 1, pp. 290–295, Dec. 2016.
- [18] A. Akagic, E. Buza, S. Omanovic, and A. Karabegovic, “Pavement crack detection using Otsu thresholding for image segmentation,” in *Proc. 41st Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2018, pp. 1092–1097.
- [19] Y. Sari, P. B. Prakoso, and A. R. Baskara, “Road crack detection using support vector machine (SVM) and OTSU algorithm,” in *Proc. 6th Int. Conf. Electric Veh. Technol. (ICEVT)*, Nov. 2019, pp. 349–354.
- [20] Y. Maode, B. Shaobo, X. Kun, and H. Yuyao, “Pavement crack detection and analysis for high-grade highway,” in *Proc. 8th Int. Conf. Electron. Meas. Instrum.*, Aug. 2007, pp. 1–14.
- [21] N.-D. Hoang, “An artificial intelligence method for asphalt pavement pothole detection using least squares support vector machine and neural network with steerable filter-based feature extraction,” *Adv. Civil Eng.*, vol. 2018, no. 1, Jan. 2018, Art. no. 7419058.

- [22] M. Gao, X. Wang, S. Zhu, and P. Guan, "Detection and segmentation of cement concrete pavement pothole based on image processing technology," *Math. Problems Eng.*, vol. 2020, pp. 1–13, Jan. 2020.
- [23] A. Ayenu-Prah and N. Attoh-Okine, "Evaluating pavement cracks with bidimensional empirical mode decomposition," *EURASIP J. Adv. Signal Process.*, vol. 2008, no. 1, pp. 1–7, Dec. 2008.
- [24] A. Zhang, K. C. P. Wang, B. Li, E. Yang, X. Dai, Y. Peng, Y. Fei, Y. Liu, J. Q. Li, and C. Chen, "Automated pixel-level pavement crack detection on 3D asphalt surfaces using a deep-learning network," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 32, no. 10, pp. 805–819, Aug. 2017.
- [25] S. Biçici and M. Zeybek, "An approach for the automated extraction of road surface distress from a UAV-derived point cloud," *Autom. Construction*, vol. 122, Feb. 2021, Art. no. 103475.
- [26] W. R. L. D. Silva and D. S. D. Lucena, "Concrete cracks detection based on deep learning image classification," in *Proc. 18th Int. Conf. Experim. Mech.*, Jun. 2018, p. 489.
- [27] S. Anand, S. Gupta, V. Darbari, and S. Kohli, "Crack-Pot: Autonomous road crack and pothole detection," in *Proc. Digit. Image Comput., Techn. Appl. (DICTA)*, Dec. 2018, pp. 1–6.
- [28] Z. Fan, Y. Wu, J. Lu, and W. Li, "Automatic pavement crack detection based on structured prediction with the convolutional neural network," 2018, *arXiv:1802.02208*.
- [29] H. Nhat-Duc, Q.-L. Nguyen, and V.-D. Tran, "Automatic recognition of asphalt pavement cracks using metaheuristic optimized edge detection algorithms and convolution neural network," *Autom. Construct.*, vol. 94, pp. 203–213, Oct. 2018.
- [30] E. Karaaslan, U. Bagci, and F. N. Catbas, "Attention-guided analysis of infrastructure damage with semi-supervised deep learning," *Autom. Construction*, vol. 125, May 2021, Art. no. 103634.
- [31] J. Guan, X. Yang, L. Ding, X. Cheng, V. C. S. Lee, and C. Jin, "Automated pixel-level pavement distress detection based on stereo vision and deep learning," *Autom. Construct.*, vol. 129, Sep. 2021, Art. no. 103788.
- [32] V. M. A. Souza, R. Giusti, and A. J. L. Batista, "Asfalt: A low-cost system to evaluate pavement conditions in real-time using smartphones and machine learning," *Pervas. Mobile Comput.*, vol. 51, pp. 121–137, Dec. 2018.
- [33] G. Singh, D. Bansal, S. Sofat, and N. Aggarwal, "Smart patrolling: An efficient road surface monitoring using smartphone sensors and crowdsourcing," *Pervas. Mobile Comput.*, vol. 40, pp. 71–88, Sep. 2017.
- [34] C. Mertz, S. Varadharajan, S. Jose, K. Sharma, L. Wander, and J. Wang, "City-wide road distress monitoring with smartphones," in *Proc. ITS World Congr.*, May 2014, pp. 1–9.
- [35] D. A. Casas Avellaneda and J. F. López-Parra, "Detection and localization of potholes in roadways using smartphones," *DYNA*, vol. 83, no. 195, pp. 156–162, Feb. 2016.
- [36] A. Angulo, J. A. Vega-Fernández, L. M. Aguilar-Lobo, S. Natraj, and G. Ochoa-Ruiz, *Road Damage Detection Acquisition System Based on Deep Neural Networks for Physical Asset Management*. Springer, 2019, pp. 3–14.
- [37] R. Roberts, G. Giancontieri, L. Inzerillo, and G. Di Mino, "Towards low-cost pavement condition health monitoring and analysis using deep learning," *Appl. Sci.*, vol. 10, no. 1, p. 319, Jan. 2020.
- [38] Y. Du, N. Pan, Z. Xu, F. Deng, Y. Shen, and H. Kang, "Pavement distress detection and classification based on YOLO network," *Int. J. Pavement Eng.*, vol. 22, no. 13, pp. 1659–1672, Jan. 2020.
- [39] H. Majidifard, P. Jin, Y. Adu-Gyamfi, and W. G. Buttler, "Pavement image datasets: A new benchmark dataset to classify and densify pavement distresses," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2674, no. 2, pp. 328–339, Feb. 2020.
- [40] S. Patra, A. I. Middya, and S. Roy, "PotSpot: Participatory sensing based monitoring system for pothole detection using deep learning," *Multimedia Tools Appl.*, vol. 80, no. 16, pp. 25171–25195, Apr. 2021.
- [41] Y. O. Ouma and M. Hahn, "Pothole detection on asphalt pavements from 2D-colour pothole images using fuzzy C-means clustering and morphological reconstruction," *Autom. Construction*, vol. 83, pp. 196–211, Nov. 2017.
- [42] H. Oliveira and P. L. Correia, "CrackIT—An image processing toolbox for crack detection and characterization," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 798–802.
- [43] Y. Shi, L. Cui, Z. Qi, F. Meng, and Z. Chen, "Automatic road crack detection using random structured forests," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 12, pp. 3434–3445, Dec. 2016.
- [44] Q. Zou, Y. Cao, Q. Li, Q. Mao, and S. Wang, "CrackTree: Automatic crack detection from pavement images," *Pattern Recognit. Lett.*, vol. 33, no. 3, pp. 227–238, Feb. 2012.
- [45] X. Weng, Y. Huang, and W. Wang, "Segment-based pavement crack quantification," *Autom. Construction*, vol. 105, Sep. 2019, Art. no. 102819.
- [46] S. Dorafshan, R. J. Thomas, and M. Maguire, "SDNET2018: An annotated image dataset for non-contact concrete crack detection using deep convolutional neural networks," *Data Brief*, vol. 21, pp. 1664–1668, Dec. 2018.
- [47] L. Li, L. Sun, G. Ning, and S. Tan, "Automatic pavement crack recognition based on BP neural network," *PROMET-Traffic Transp.*, vol. 26, no. 1, pp. 11–22, Jan. 1970.
- [48] F. Yang, L. Zhang, S. Yu, D. Prokhorov, X. Mei, and H. Ling, "Feature pyramid and hierarchical boosting network for pavement crack detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1525–1535, Apr. 2020.
- [49] Q. Mei and M. Gül, "A cost effective solution for pavement crack inspection using cameras and deep neural networks," *Construction Building Mater.*, vol. 256, Sep. 2020, Art. no. 119397.
- [50] M. Eisenbach, R. Stricker, D. Seichter, K. Amende, K. Debes, M. Sesselmann, D. Ebersbach, U. Stoeckert, and H.-M. Gross, "How to get pavement distress detection ready for deep learning? A systematic approach," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2039–2047.
- [51] R. Stricker, M. Eisenbach, M. Sesselmann, K. Debes, and H.-M. Gross, "Improving visual road condition assessment by extensive experiments on the extended GAPs dataset," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.
- [52] S. Xu, X. Wang, W. Lv, Q. Chang, C. Cui, K. Deng, G. Wang, Q. Dang, S. Wei, Y. Du, and B. Lai, "PP-YOLOE: An evolved version of YOLO," 2022, *arXiv:2203.16250*.
- [53] C. Lyu, W. Zhang, H. Huang, Y. Zhou, Y. Wang, Y. Liu, S. Zhang, and K. Chen, "RTMDet: An empirical study of designing real-time object detectors," 2022.
- [54] G. Jocher, A. Chaurasia, and J. Qiu. (2023). *Ultralytics YOLOv8*. [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [55] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
- [56] Z. Guo, C. Liu, X. Zhang, J. Jiao, X. Ji, and Q. Ye, "Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8788–8797.
- [57] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
- [58] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.
- [59] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *Proc. 16th Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 282–298.
- [60] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "TOOD: Task-aligned one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3490–3499.
- [61] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2017, *arXiv:1708.02002*.
- [62] Z. Tian, X. Chu, X. Wang, X. Wei, and C. Shen, "Fully convolutional one-stage 3D object detection on LiDAR range images," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 34899–34911.
- [63] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.
- [64] Z. Ge, S. Liu, Z. Li, O. Yoshie, and J. Sun, "OTA: Optimal transport assignment for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 303–312.
- [65] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.
- [66] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.
- [67] X. Li, C. Lv, W. Wang, G. Li, L. Yang, and J. Yang, "Generalized focal loss: Towards efficient representation learning for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3139–3153, Mar. 2023.

- [68] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.
- [69] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9756–9765.
- [70] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, "YOLOv6: A single-stage object detection framework for industrial applications," 2022, *arXiv:2209.02976*.
- [71] D. Arya, H. Maeda, S. K. Ghosh, D. Toshniwal, H. Omata, T. Kashiyama, and Y. Sekimoto, "Crowdsensing-based road damage detection challenge (CRDDC'2022)," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2022, pp. 6378–6386.
- [72] H. Li, T. Peng, N. Qiao, Z. Guan, X. Feng, P. Guo, T. Duan, and J. Gong, "CrackTinyNet: A novel deep learning model specifically designed for superior performance in tiny road surface crack detection," *IET Intell. Transp. Syst.*, vol. 18, no. 12, pp. 2693–2712, Dec. 2024.
- [73] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10778–10787.
- [74] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, K. Michael, J. Fang, C. Wong, Z. Yifu, A. V. D. Montes, Z. Wang, C. Fati, J. Nadar, P. Skalski, A. Hogan, M. Strobel, M. Jain, and L. Mammana, "Ultralytics/YOLOv5: V6.2–YOLOv5 classification models, Apple M1, reproducibility, ClearML and Deci.AI integrations," Aug. 2022.
- [75] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. 14th Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, Oct. 2016, pp. 21–37.
- [76] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.



ASMA MUSABAH ALKALBANI received the B.Sc. degree in computer engineering from National University, Oman, in 2005, the M.Sc. degree in information technology from La Trobe University, Melbourne, Australia, in 2010, and the Ph.D. degree in computer science from the University of Technology Sydney, Australia, in 2018.

She is currently an Assistant Professor with the Department of Information Technology, University of Technology and Applied Sciences, Ibri,

Oman. She is also seconded to Oman Vision 2040 Unit, where she works with the Digital Transformation and Emerging Technologies team on national-level initiatives. Previously, she was a Visiting Scholar with the School of Computing, Macquarie University, Sydney, Australia. Her research interests include cloud computing, data analytics, and applied artificial intelligence research. She received Oman National Research Award, in 2021.



MUHAMMAD SAQIB received the Ph.D. degree in analytics from the University of Technology Sydney, Australia, in 2019.

He is currently a Research Scientist with the Engineering and Technology Division, National Collection and Marine Infrastructure, CSIRO, Australia. Additionally, he holds an honorary position of Research Fellow with the University of Technology Sydney (UTS). His research interests include deep learning, data analytics, and computer vision.



AHMED SALIM ALRAWAHI received the B.Sc. degree in information systems from the University of Nizwa, Nizwa, Oman, in 2009, the M.Sc. degree in network management and security from Murdoch University, Perth, Australia, in 2012, and the Ph.D. degree in computer science from Nottingham Trent University, Nottingham, U.K., in 2020. He is currently an Assistant Professor with the Department of Information Technology, University of Technology and Applied Sciences,

Ibri, Oman, and a Visiting Scholar with the AI Applications Research Chair, University of Nizwa. His research interests include the IoT, cloud computing, optimization approaches for combinatorial exchange problems, and AI applied research.



ABBAS ANWAR received the bachelor's degree in computer science from the University of Abdul Wali Khan, Mardan, Pakistan. He is currently an Associate Researcher with the National University of Computer and Emerging Sciences (NUCES), Pakistan. He is a Machine Learning Engineer with FutureDataMinds. He is also affiliated with FutureDataMinds, Australia. He has published in top-tier conferences and journals, such as CVPR, IEEE ACCESS, *Signals*, and *Electronics*.

His research interests include computer vision, machine learning, and their applications across various domains.



CHANDRANATH ADAK (Senior Member, IEEE) received the Ph.D. degree in analytics from the University of Technology Sydney, Australia, in 2019.

He is currently an Assistant Professor with the Department of CSE, Indian Institute of Technology Patna, India. His research interests include deep learning, data analytics, and computer vision.



SAEED ANWAR is currently a Senior Lecturer of computer vision, machine learning, AI, and data science with the Department of Computer Science and Software Engineering, University of Western Australia. He regularly reviews submissions to top venues in his field, including TPAMI, TIP, CVPR, ICCV, ECCV, ACM MM, and many others. He has served as a team lead on multiple commercial projects. He has also supervised Ph.D. and master's students through to completion.

He also holds honorary affiliations at the Australian National University and CSIRO. He has received best paper awards at ICPR'24 and RS'22, and has been nominated for best paper awards at CVPR'20 and PR'19. Moreover, he is a Clarivate Highly Cited Researcher for 2025 and holds a Senior Fellow of the Higher Education Academy (SFHEA) recognition. His editorial contributions include guest editorials for the IEEE's JOE, *Sensors* (MDPI), and *Frontiers in Robotics and AI*. He has served as the Area Chair for ECCV'26, CVPR'26, WACV'26, ICCV'25, IJCAI'25, and ACCV'24. He serves as an Associate Editor for IEEE JOURNAL OF OCEANIC ENGINEERING (JOE), and previously served as an Associate Editor for *Neurocomputing*.

...