information

MDPI

Article

AVD-YOLO: Active Vision-Driven Multi-Scale Feature Extraction for Enhanced Road Anomaly Detection

Minhong Jin, Zhongjie Zhu *, Renwei Tu, Ang Lv and Zhijing Yu

Key Laboratory of Industrial Vision and Industrial Intelligence, Zhejiang Wanli University, Ningbo 315100, China; 2023882054@zwu.edu.cn (M.J.); turenwei@zwu.edu.cn (R.T.); lu.ang@foxmail.com (A.L.); 2023882058@zwu.edu.cn (Z.Y.)

* Correspondence: zhongjiezhu@zwu.edu.cn

Abstract

Deficiencies in road anomaly detection systems precipitate multifaceted risks, including elevated collision probabilities from unidentified hazards, compromised traffic flow efficiency, and exponential maintenance costs. Contemporary methods struggle with complex road environments, dynamic viewing perspectives, and limited datasets. We present AVD-YOLO, an enhanced YOLO variant that synergistically integrates Active Vision-Driven (AVD) multi-scale feature extraction with Position Modulated Attention (PMA) mechanisms. PMA addresses diminished target-background discriminability under variable illumination and weather conditions by capturing long range spatial dependencies, enhancing weak-feature target detection. The AVD technique mitigates missed detections caused by real-time viewing distance variations through adaptive multi-receptive field mechanisms, maintaining conceptual target fixation while dynamically adjusting feature scales. To address data scarcity, a comprehensive Multi-Class Road Anomaly Dataset (MCRAD) comprising 14,208 annotated images across nine anomaly categories is constructed. Experiments demonstrate that AVD-YOLO improves detection accuracy, achieving a 1.6% gain in mAP@0.5 and a 2.9% improvement in F1-score over baseline. These performance gains indicate both more precise localization of abnormal objects and a better balance between precision and recall, thereby enhancing the overall robustness of the detection model.

Keywords: road anomaly detection; mobile visual target; positional encoding; multi-scale

Academic Editor: Vincenzo Moscato

Received: 15 October 2025

Revised: 18 November 2025

Accepted: 1 December 2025

Published: 3 December 2025

Citation: Jin, M.; Zhu, Z.; Tu, R.; Lv, A.; Yu, Z. AVD-YOLO: Active Vision-Driven Multi-Scale Feature Extraction for Enhanced Road Anomaly Detection. *Information* **2025**, *16*, 1064. https://doi.org/10.3390/info16121064

1. Introduction

Expanding transportation demand and unprecedented vehicle ownership growth intensify road safety concerns, accelerating the need for automated anomaly detection [1]. With United Nations projections indicating that 68% of the global population will inhabit urban areas by 2050, this rapid urbanization fundamentally amplifies traffic monitoring challenges beyond traditional management capabilities [2]. World Health Organization data reveals approximately 1.3 million annual fatalities from road traffic crashes, with road anomalies serving as primary contributors [3]. High-speed urban roadways create conditions where missing a single anomalous event can trigger cascading failures, threatening road users' safety while inducing congestion and throughput degradation [4]. Road anomalies comprise diverse categories including construction debris, foreign objects, pedestrians, fire, spills, infrastructure deterioration, and traffic incidents, collectively undermining traffic flow and safety [5]. Postponement or inability to identify such anomalies not only poses a threat to road users' safety but also triggers secondary effects, including congestion and throughput loss. Conventional manual inspection strategies inevitably fall short in terms of timeliness, coverage, and cost-effectiveness, underscoring the necessity for efficient and precise automated anomaly detection systems as a cornerstone of modern road network management [6,7].

Given these escalating challenges, current road anomaly detection methodologies fall into two primary categories: traditional manual-based approaches and multi-sensor fusion methods [8,9]. Traditional approaches encompass manual video surveillance interpretation, periodic inspections, and fixed-route examinations. These methods depend heavily on human intervention, suffering from subjective judgment, low efficiency, and insufficient capacity for modern demands. As transportation infrastructure expands, these approaches demonstrate excessive labor intensity, reduced operational efficiency, and increased susceptibility to missed detections. Such limitations render them progressively incapable of satisfying the precision and efficiency requirements of modern intelligent transportation systems [1]. Consequently, automated detection solutions based on multi-sensor integration emerge as compelling alternatives.

Among these automated solutions, multi-purpose inspection vehicles equipped with sensors such as GPS, cameras, laser profilometers, and ground-penetrating radar achieve substantial enhancements in detection accuracy and efficiency [10]. These systems facilitate data acquisition under normal traffic flow conditions, eliminating disruptions typical of conventional detection methods. During the early 21st century, numerous countries developed specialized road defect detection vehicles, with some nations deploying nocturnal inspection vehicles to expand operational coverage [11]. Nevertheless, equipment acquisition and maintenance costs constitute primary barriers to widespread adoption. Budget constraints particularly affect rural road management authorities, resulting in minimal coverage of advanced detection equipment [12]. While 3D sensor technology excels at road surface morphology reconstruction and microscopic defect identification, hardware costs and data processing complexity limit large-scale deployment feasibility. This limitation compels researchers to pursue software-based solutions leveraging existing surveillance infrastructure.

Deep learning-based computer vision algorithms achieve breakthrough progress in road anomaly detection tasks. Object detection models employ end-to-end training strategies that directly optimize the mapping relationship from input images to detection results through backpropagation algorithms, circumventing the separated feature extraction and classifier design processes inherent in traditional methods. Detection networks trained on large-scale annotated datasets attain performance metrics comparable to human annotation. DETR (DEtection TRansformer) [13] and YOLO (You Only Look Once) [14–17] series, including YOLOv5 (Version 7.0) [14], YOLOv8 (Version 8.0.0) [15], YOLOv10 [16], and YOLOv11 (Version 11.0.0) [17], serve as predominant object detection architectures widely deployed in road anomaly detection applications. DETR introduces Transformer architecture to enable global modeling yet suffers from slow training convergence and substantial computational resource consumption. YOLO series algorithms leverage single-stage detection strategies to achieve inference speed advantages, while still confronting technical challenges in weak feature target detection under complex backgrounds, multi-scale target adaptation, and precise localization.

To address the limitations of existing road anomaly detection methods in handling complex and dynamic road environments, adapting to real-time sight distance changes caused by target motion, an Active Vision-Driven Multi-scale Feature Extraction method for Enhanced Road Anomaly Detection (AVD-YOLO) is proposed. To cope with the scarcity of relevant training data, a Multi-Class Road Anomaly Dataset (MCRAD) is constructed. This research emphasizes dual objectives of hazard prevention and emergency response, enabling robust anomaly detection in dynamic traffic conditions. The experimental results verify the effectiveness of the proposed method. The contributions of this paper can be summarized as follows:

(1) Road environment complexities, including variations in light, weather, changing road conditions, and insignificant features of abnormal objects, often reduce target-background contrast, making accurate model perception and localization difficult. To overcome these limitations, a Position-Modulated Attention (PMA) module is proposed to efficiently capture long-range dependencies for precise localization, enhance scene adaptation, and improve the detection of weak-feature targets.
(2) In road anomaly detection, dynamic changes in the size of the target to be detected in the road scene due to the real-time view distance change caused by movement often lead to missed detections or reduced localization accuracy. An Active Vision Driven Multi-scale Feature Extraction (AVD) module is proposed, which performs multi-scale feature extraction from multiple receptive fields by actively adjusting the viewing distance while conceptually maintaining a fixed target position, thereby alleviating the limitations inherent in single-scale methods.
(3) Considering the difficulty of road data collection, the scarcity of public datasets, and the lack of clear anomaly classifications, MCRAD is constructed to define nine types of road anomalies and provide sufficient training data, thereby enabling robust detection within this defined scope.

The remainder of this paper is organized as follows. Section 2 reviews related work; Section 3 describes the proposed method in detail; Section 4 presents comprehensive experiments and result analysis; and Section 5 concludes this paper.

2. Related Work

The proliferation of deep learning in computer vision has catalyzed significant breakthroughs in automated road anomaly detection, shifting from manual inspection paradigms to intelligent visual analysis systems. Contemporary research primarily focuses on adapting general object detection frameworks to address the unique challenges of road monitoring, including variable target scales, complex backgrounds, and real-time processing requirements.

The advancement of convolutional neural networks propels progress in computer vision-based road anomaly detection technologies. Detection frameworks exemplified by DETR and YOLO series garner substantial attention due to their distinctive advantages. DETR pioneered the integration of Transformer architecture into object detection. It implements an end-to-end detection paradigm without anchor frames or non-maximum suppression, thereby improving detection scalability. To address the need for timely detection of road traffic accidents, Srinivasan et al. [18] proposed a frame-by-frame accident discrimination method based on DETR. Their approach fuses timing information through sliding windows to enable early warning. Liu et al. [19] developed the MDFD2-DETR model that enhances both efficiency and accuracy by reducing redundancy through multi-domain feature decomposition and optimizing feature interactions via hybrid positional encoding. It achieves superior performance compared to existing methods across multiple datasets. Despite these advances, DETR-based methods suffer from several limitations. These include slow convergence, high computational costs, strong data dependency, and poor small-target detection. In contrast, the lightweight, single-stage YOLO series delivers superior real-time performance for road surveillance. Liu et al. [20] combined 3D ground-penetrating radar with YOLO model to quickly identify internal defects of asphalt pavement, significantly optimizing maintenance costs and environmental impact. Similarly, Z. Yang et al. [21] proposed a lightweight PDNet by improving YOLO to achieve efficient multi-scale detection of highway pavement defects. For region-specific applications, Shuvo et al. [22] developed an exclusive traffic sign dataset for Bangladeshi roads. They also proposed a YOLO-based framework for both sign detection and recognition. Pei et al. [23] designed YOLO-RDD specifically for slender defects such as cracks. Their method employs feature fusion, dynamic convolution, and cross-layer attention mechanisms, which significantly improve recognition performance for multi-scale and morphological features. These studies collectively demonstrate the YOLO series' advantages in speed, real-time performance, and multi-scene adaptability, efficiently supporting practical road anomaly detection and traffic management. Nevertheless, YOLO methods still face challenges in detecting weak-featured targets in complex backgrounds, adapting to dynamic size changes, and achieving high-precision localization.

To better solve these limitations, researchers have suggested numerous methodological innovations in network architecture design, spatial position encoding, and multi-scale feature fusion. These innovations particularly focus on enhancing model perception abilities for weak-feature targets in cluttered environments. Excellent progress has been achieved in core technologies such as attention mechanisms, positional encoding strategies, and multi-scale representation learning. Transformer designs [24] have revolutionized spatial context modeling through self-attention mechanisms, where positional encoding techniques provide critical spatial information to compensate for the permutation-invariant nature of attention operations. However, standard Transformer implementations treat positional information as static additive inputs. This approach limits their adaptability to complex spatial dependency patterns across diverse vision tasks. Swin Transformer [25] mitigates computational complexity by employing local window-based attention mechanisms. Nevertheless, this architectural design trades off global receptive fields and may compromise the modeling of long-range dependencies. To further enhance model performance for more intricate situations like road anomaly detection, it is critical to design more flexible and effective spatial positional encoding mechanisms. These mechanisms need to dynamically capture spatial distances and correlations between features and efficiently combine global contextual information. This enables the model to accurately resolve local ambiguities and reliably identify targets with low contrast, small objects, or heavily occluded targets, hence greatly boosting model's discriminative power and adaptability in different environments.

In road anomaly detection missions, the spatial distribution and scale of targets are highly variable, requiring methods to capture spatial relationships and handle multiple scales simultaneously. Spatial Pyramid Pooling (SPP) series modules have been extensively adopted as classical solutions for multi-scale feature extraction in object detection frameworks. However, comprehensive evaluation of existing SPP variants reveals significant shortcomings in road anomaly detection applications. The initial SPP developed by He et al. [26] applies parallel pooling kernels of variable sizes to obtain multi-scale features and generate fixed-length feature vectors, effectively addressing input dimension constraints in convolutional neural networks. Nevertheless, this approach demonstrates high computational complexity and limited adaptability to dynamic scale variations due to predetermined pooling dimensions. The SPPF module introduced in YOLOv5 [14] enhances computational efficiency through serial concatenations of multiple max-pooling layers, achieving approximately twofold acceleration in processing speed. However, its exclusive reliance on max-pooling operations tends to discard critical contextual information when handling targets with ambiguous boundaries, such as road surface irregularities or scattered debris. Similarly, the SPPCSPC architecture presented by Wang et al. [27] reduces computation cost with cross-stage partial connections but restricts adaptive multi-scale processing capability with fixed feature splitting ratios.

Deep learning-based road anomaly detection methods demonstrate significant application value in enhancing road safety management, yet existing approaches encounter technical bottlenecks in weak feature target detection under complex road backgrounds, dynamic scale variation adaptation, and real-time performance requirements. Among these developments, the YOLO series has emerged as the predominant choice for road surveillance scenarios due to its efficient single-stage detection, exhibiting unique advantages in balancing speed and accuracy. This paper proposes AVD-YOLO, an algorithm optimized for road anomaly detection tasks, constructing a more efficient road anomaly detection method tailored to the practical requirements of road monitoring systems.

3. Methodology

AVD-YOLO represents an enhanced variant of the YOLO architecture specifically optimized for road anomaly detection. The innovation of AVD-YOLO lies in two modules: PMA, which enhances detection of weakly characterized targets by improving target-background discrimination and localization in complex environments, and AVD, which addresses scale variations due to varying distances between targets and the camera through adaptive multi-receptive fields, reducing missed small-target detections common in single-scale methods.

3.1. AVD-YOLO Backbone Network

The backbone network of AVD-YOLO, tasked with initial feature extraction, is engineered to augment its representational power through the integration of two pivotal modules: PMA and AVD. To accommodate diverse road monitoring scenarios, AVD-YOLO processes input images from multiple sources including fixed surveillance cameras at intersections and highways, vehicle-mounted cameras from patrol and regular vehicles, and mobile devices used for incident documentation. All input images are preprocessed to 640×640 resolution with RGB channels for consistent processing. This multi-source approach ensures comprehensive coverage of road anomalies from various perspectives and viewing angles. The overall AVD-YOLO architecture comprises three primary components: the backbone network, the neck network, and the detection head.

As shown in Figure 1, AVD-YOLO extracts hierarchical features using a backbone network of CBS, C3, PMA, and AVD modules, followed by multi-scale fusion in the neck and final prediction by the detection head. PMA reweights attention with a Manhattan distance-based spatial decay matrix to improve global perception and long-range dependency modeling, enhancing detection of weak-featured targets in challenging conditions. The AVD module performs active multi-scale feature extraction across multiple receptive fields to handle scale variations due to varying distances between targets and the camera, reducing missed small target detections.

3.2. Position-Modulated Attention

Dynamic changes in road conditions, combined with low target–background contrast, are major challenges to strong perception and correct localization in detection systems. Conventional convolutional operations, limited by local receptive fields, often fail to capture critical discriminative cues dispersed in global context or distant regions. Unlike the self-attention mechanism in DETR that operates on learned object queries for end-to-end detection, PMA is specifically designed to enhance feature extraction within grid-based detection frameworks like YOLO. PMA operates on individual frames, enhancing spatial feature extraction within single images rather than processing temporal sequences or tracking objects across frames.

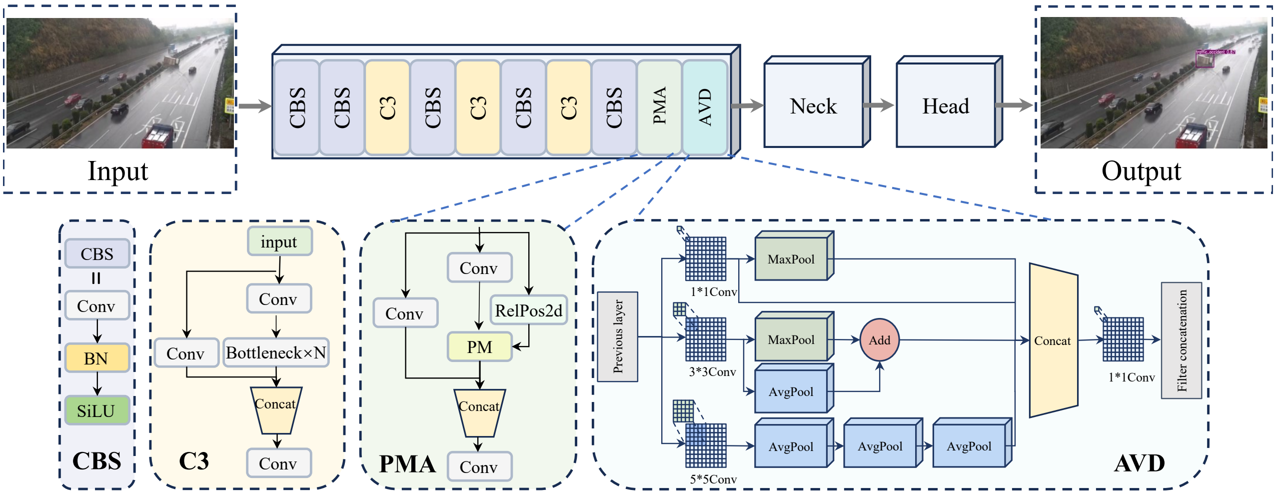

Figure 1. AVD-YOLO architecture.

The proposed PMA architecture brings forth a distance-based attention decay mechanism, extending this principle to spatial dimensions by applying Manhattan distance in the computation of decay weights. This mechanism captures planar distances within 2D images, enabling the model to handle objects at various scales and positions within a single frame. This enables each token to simultaneously consider feature similarity and relative spatial proximity as explicit priors during self-attention calculation. By leveraging a spatial decay matrix, PMA adaptively modulates attention allocation and assigns context-sensitive weights based on spatial separation, thereby enhancing the model's ability to extract spatial structures and synthesize long-range dependencies while maintaining computational efficiency. Such an approach natively integrates spatial relations into the attention mechanism, facilitating a more comprehensive representation of spatial context that is particularly instrumental for road anomaly detection tasks, in which spatial coherence and positional information are essential to precise and accurate target identification. The structural details of the PMA module are depicted in Figure 2.

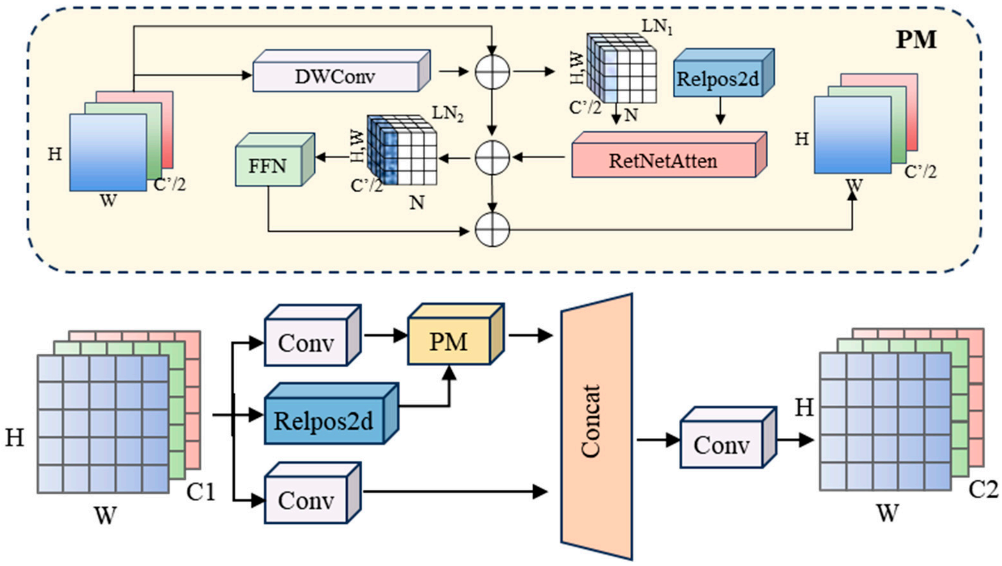

Figure 2. Position-Modulated Attention structure.

Within the overall architecture of the PMA module, the input feature map $X \in R^{C1 \times H \times W}$ is processed in multiple parallel branches, incorporating convolutional layers, Two-Dimensional Relative Position Encoding (RelPos2d) [28], and a Positional Modulator

(PM). Outputs from all branches are concatenated along the channel dimension and subsequently fused through a convolutional layer to produce a new feature map $X'' \in R^{C2\times H\times W}$, which serves as input for the succeeding layers in the backbone network.

To realize dynamic modeling of spatial dependencies, the PMA module consists mainly of two subcomponents: RelPos2d and PM. For each token in the input feature map, RelPos2d first computes its absolute coordinates in the two-dimensional space, (x_i, y_i). The pairwise Euclidean distance $d_{i,j} = \left|\left|(x_i,\ y_i) - \ (x_j,\ y_j)\right|\right|_p$ between tokens is then calculated to construct a distance matrix $D \in R^{N\ \times N}$. This matrix is subsequently mapped by a learnable embedding function f_{emb} into a parameterized relative position embedding matrix $E_{i,j} = \ f_{emb(d_{i,j})}$, which is utilized to modulate the attention weights. Moreover, a bidirectional decay mechanism is introduced, applying exponential decay to distant tokens to achieve region-weighted attention within the self-attention paradigm.

The PM leverages the decay embedding matrix produced by RelPos2d to dynamically modulate attention weights as a function of token-to-token spatial distance. This enables the model to effectively capture local spatial features while preserving global contextual information. Specifically, for tokens n and m with coordinates (x_n, y_n) and (x_m, y_m) respectively, the one-dimensional spatial decay matrices based on Manhattan distance, denoted as D^H_{nm} and D^W_{nm}, are defined as follows:

$$D^H_{nm} = \gamma^{|y_n - y_m|} \tag{1}$$

$$D^W_{nm} = \gamma^{|x_n - x_m|} \tag{2}$$

where γ is a learnable decay base (0 < γ < 1) that controls the decay rate. Different attention heads correspond to different values of γ to realize multi-scale information capture. Then the query (Q), key (K) and value (V) matrices are computed as

$$Q = X'W_Q, K = X'W_K, V = X'W_V \tag{3}$$

where X' denotes the input X that has been processed by Layer Normalization (LN), X' = LN(X), and W_Q, W_K, and W_V are learnable linear projection matrices.

Next, the spatial attention in the decomposed form is computed by calculating the attention scores along the width $Attn_W$ and the height $Attn_H$, respectively, and applying the attenuation, which is denoted by $Attn_W$ and $Attn_H$:

$$Attn_W = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \odot D^W \tag{4}$$

$$Attn_H = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \odot D^H \tag{5}$$

where d_k is the dimension of the key vector and $\odot$ denotes the element-by-element multiplication (Hadamard product).

$$V_W = Attn_W V \tag{6}$$

$$PMAttn(X) = Attn_H V_W \tag{7}$$

The final decomposed attention mechanism models 2D spatial interactions through a sequential two-stage feature aggregation process. It initially performs a weighted summation along the width dimension, utilizing width-dependent spatial decay attention weights, followed by an additional weighting of the intermediate output along the height dimension, guided by height-dependent spatial decay attention weights. This strategy effectively reduces computational complexity while maintaining critical spatial relationships. To further strengthen local feature expressiveness, we introduce Local Enhancement Positional Encoding (*LEPE*), which applies a 5×5 depth-wise separable convolution [29] (*DWConv*) to the linearly transformed value matrix V. The larger convolutional kernel facilitates the extraction of richer local contextual information, enabling the module to attend to fine-grained details while retaining global context. Denote *LEPE* as the local enhancement positional encoding operator. The locally enhanced representation of V is formulated as

$$LEPE(V) = DWConv_{5\times5}(V) \tag{8}$$

where V represents the input value matrix and $DWConv_{5\times5}$ denotes the depthwise separable convolution operation with a 5×5 kernel.

Within the PMA architecture, the attention output is combined with LEPE through residual connections and a Feed-Forward Network (FFN). The input features X_{in} to the PM component undergo depth-wise separable convolution with positional encoding $DWConv_{pos}()$, which preserves information integrity while facilitating identity mapping to enhance training stability. W_O represents the final learnable output projection matrix of the attention module, and X_{final} denotes the output of this component within PM. The complete PMA module produces the final output $X'' \in R^{C2\times H\times W}$ through the following formulation:

$$X_{pos} = X_{in} + DWConv_{pos}(X_{in}) \tag{9}$$

$$X_{out} = X_{pos} + DropPath\big(W_O\big(PMAttn\big(LN_1\big(X_{pos}\big)\big) + LEPE(V)\big)\big) \tag{10}$$

$$X_{final} = X_{out} + DropPath(FFN(LN_2(X_{out}))) \tag{11}$$

FFN consists of two linear layers and activation functions, while LN and DropPath denote layer normalization and regularization. In PMA, the C3 module's two-branch structure is preserved with the main branch Bottleneck replaced by PM, thereby retaining CSP's efficient feature extraction and introducing long-range dependency modeling to enhance feature representation in complex spatial scenarios.

3.3. Active Vision Driven Multi-Scale Feature Extraction

To facilitate effective multi-scale feature extraction, this work formulates an AVD module that dynamically shapes the perceptual field through multi receptive field processing for comprehensive feature representation. From an input feature map $X \in R^{C\times H\times W}$, where C denotes the number of input channels, and H, W represent spatial dimensions, the AVD module processes features through three parallel pathways. The first branch employs 1×1 convolution to maintain the original feature dimensions while preserving spatial details. The second branch integrates 3×3 convolution with dual pooling mechanisms to balance feature saliency and spatial consistency within medium receptive fields. The third branch utilizes 5×5 convolution followed by three sequential AvgPooling operations to progressively expand the receptive field and aggregate large-scale contextual information. This multi-branch architecture enhances feature representation richness through parallel processing and improves model robustness by leveraging complementary effects of different kernel sizes and heterogeneous pooling strategies, thereby achieving more comprehensive multi-scale feature capture.

Figure 3 presents the structure of Active Vision Driven Multi-scale Feature Extraction. During feature fusion, multi-scale feature maps are concatenated along the channel dimension and subsequently processed by a 1×1 convolution, enabling adaptive integration of cross-channel information through learnable weights at each spatial location. The resulting feature map, $Y \in R^{C'\times H\times W}$, where C' denotes the number of output channels, encapsulates both fine-grained local details and extensive contextual information, thus providing a robust foundation for downstream detection tasks.

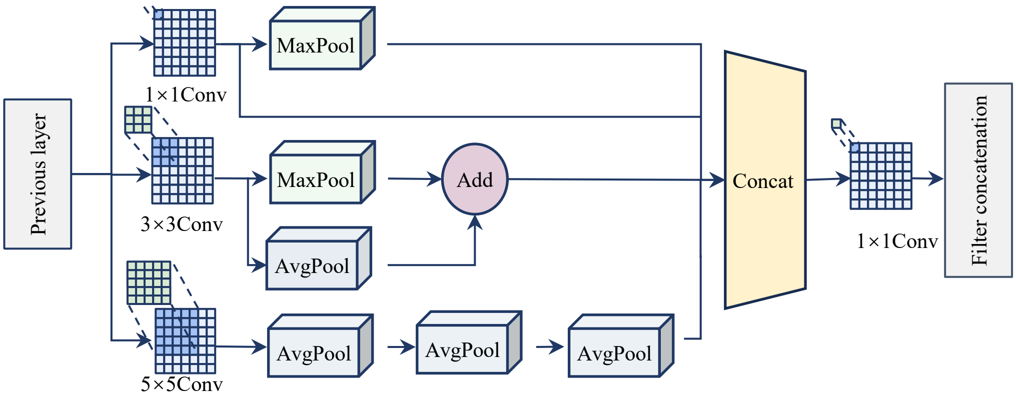

Figure 3. Active Vision Driven Multi-scale Feature Extraction structure.

4. Experimental Results and Analysis

The effectiveness of AVD-YOLO for road anomaly detection is evaluated on the MCRAD. Experimental settings and evaluation metrics are presented, with quantitative comparison to mainstream algorithms and qualitative analysis via visualization. Ablation studies further assess the impact of key modules on performance.

4.1. Dataset

A major challenge in road anomaly detection is the scarcity of real-world data, as anomalous events are rare and unpredictable, making large-scale collection difficult. Manual labeling is costly, and privacy concerns often restrict data release. As a result, existing datasets are small, lack scenario diversity, and have category imbalance.

Although some studies utilize domain-specific datasets for road anomaly detection, publicly available collections remain notably limited. For instance, the Road Anomaly dataset [30] offers an image compilation focused on hazardous road scenarios, comprising 60 uniformly sized images that depict various perils vehicles might encounter, such as animals, rocks, lost tires, trash cans, and construction equipment. Focusing specifically on pavement integrity, RDD-2020 [31] constitutes a large-scale, heterogeneous dataset of road surface defects, containing 26,336 images that document over 31,000 instances of pavement damage, classified into four principal categories: longitudinal cracks, transverse cracks, alligator cracks, and potholes. However, current datasets target single anomaly types, leaving a comprehensive road anomaly dataset unavailable.

The main challenges in road anomaly detection include the ambiguous definition of anomaly types and the absence of unified standards. These issues impede standardized dataset construction and hinder fair comparison of detection methods. Based on systematic analysis of road safety risks, traffic management needs, and relevant research, nine major road anomaly categories are defined in this study, with their sample distribution and comprehensive descriptions detailed in Table 1.

This research establishes a specialized road anomaly detection dataset comprising 16,244 annotated images across nine distinct categories: construct, matter, person, fire, spill, bad state, illicit vehicle, animal (e.g., pigs, sheep, cows, etc.) and traffic accident. The dataset incorporates images from heterogeneous sources to reflect authentic road monitoring conditions, including stationary traffic cameras, onboard vehicular recording systems, and handheld devices used by traffic personnel. This diversity in image acquisition sources provides varying perspectives, resolutions, and environmental conditions, ensuring the dataset encompasses the spectrum of real-world road scenarios, ranging from wide-angle highway views to detailed close-up captures of specific incidents. All images are meticulously annotated with bounding boxes using the open-source LabelImg tool in accordance with object detection requirements, with each anomaly category manually delineated and subjected to dual review by Professionals. For irregular or diffusive anomalies such as spill and fire, bounding boxes are appropriately expanded to encompass ambiguous regions, thereby enhancing the model's capacity to detect amorphous targets. For small and easily occluded objects such as person and illicit vehicles, each instance is individually annotated to strictly prevent omissions, ensuring comprehensive and precise target labeling. The dataset is partitioned into training and validation sets with an 80:20 split to ensure rigorous experimental standards and robust generalization assessment. The annotation format adheres to the conventional YOLO-series specification, where each target instance is encoded as "class id, center x, center y, width, height" with normalized coordinates relative to image dimensions.

Table 1. Number of samples and detailed definitions for each category in the dataset.

Style	Sample Size	Definition and Typical Scenarios
Construct	572	Temporary construction zones including warning signs, and workers in safety gear (reflective vests, helmets)
Matter	484	Foreign objects on road surface such as fallen trees, rocks, and other obstacles that pose collision risks
Person	1081	pedestrians in unauthorized road areas where their presence poses safety risks
Fire	1173	Fire-related incidents including vehicle combustion, roadside fires, and smoke that affect visibility
Spill	1471	Scattered items and materials on road surface including plastic bags, tires, packaging materials, and other dispersed objects that affect driving safety
Bad state	1140	Road infrastructure damage such as potholes, cracks, and collapsed sections
Illicit vehicles	1283	Non-motorized and unauthorized vehicles in dangerous or restricted road areas, particularly tricycles and bicycles
Animal	4349	Animals on roadway, particularly those frequently appearing on highways including pigs, cattle, sheep, horses, dogs, and other livestock or domestic animals that create collision hazards
Traffic accident	2655	Vehicle collision incidents including crashes and overturned vehicles
Total	14,208	

As demonstrated in Figure 4, which presents representative annotation samples from all nine categories, the center coordinates (x, y) and dimensional parameters (width, height) of each bounding box are normalized to the [0, 1] range relative to the image dimensions. Each image corresponds to an annotation file containing line-by-line records of all bounding boxes. As illustrated in Figure 4, we present representative annotation examples encompassing all nine categories. Specifically, (a) displays bounding boxes for construct and person; (b) includes matter, person, and illicit vehicles; (c) shows annotated regions for fire; (d) contains annotations for traffic accident and person; (e) shows traffic accident and person; (f) incorporates traffic accident, person, illicit vehicles, and spill; (g) highlights bad state; (h) involves person and illicit vehicles; and (i) presents annotations for animal and traffic accident. These examples clearly demonstrate the diversity and precision of the annotation strategy adopted in this work. These annotations comprehensively encompass all nine major anomaly categories within the dataset, effectively demonstrating the clarity and accuracy of bounding box delineation across diverse classes and complex scenarios. The standardized annotation protocol provides robust data support for precise anomaly detection and reliable differentiation among multiple categories.

Figure 4. Annotation examples for nine categories. (**a**) construct and person; (**b**) matter, person, and illicit vehicles; (**c**) fire; (**d**) traffic accident and person; (**e**) traffic accident and person; (**f**) traffic accident, person, illicit vehicles, and spill; (**g**) bad state; (**h**) person and illicit vehicles; (**i**) animal and traffic accident.

4.2. Experimental Environment and Evaluating Indicator

All the experiments are conducted on a computing server equipped with an Intel Xeon Gold 5218 CPU (2.30GHz) and an NVIDIA A100 GPU (40GB graphics memory), with the operating system Ubuntu 18.04.5 LTS. The training parameter configuration is presented in Table 2, where SGD (Stochastic Gradient Descent) is employed as the optimizer.

Table 2. Training parameter settings.

Parameter	Value
epochs	200
batch size	4
imgsz	640×640
weight decay	0.0005
learning rate	0.01
momentum	0.937
weight decay	0.0005
warmup epochs	3.0

MCRAD is constructed to serve as the empirical foundation for the investigations presented. Standard object detection metrics are used to evaluate model performance, including Precision (P), Recall (R), F1-Score (F_1), and mean Average Precision (mAP). P reflects prediction accuracy, R reflects detection completeness, and F_1 provides a measure of the balance between the two. The mAP calculated under the commonly used IoU threshold of 0.5 is mainly reported as the core performance metric. Comprehensive analysis of these metrics provides an in-depth and objective evaluation of model effectiveness for road anomaly detection. The formulas for Precision, Recall, and F_1 are as follows:

$$P = \frac{TP}{TP + FP} \tag{12}$$

$$R = \frac{TP}{TP + FN} \tag{13}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{14}$$

where TP, FP, and FN represent the number of True Positives, False Positives, and False Negatives, respectively. AP is defined as the Average Precision for a single class. mAP is calculated by averaging the AP over all C classes, with C being the total number of classes. The specific formula is provided as follows.

$$AP = \int_0^1 P(R)dR \tag{15}$$

$$mAP = \sum_{i=1}^{C} \frac{AP_i}{C} \tag{16}$$

4.3. Comparisons with Representative Methods

To demonstrate the effectiveness of the proposed method, it is evaluated against mainstream detection models using the constructed MCRAD. Detection performance is assessed to compare the methods. The baseline models involved in the comparison include YOLOv5 [14], YOLOv8 [15], YOLOv10 [16], RT-DETR [18], and YOLO11 [17]. All experiments follow a unified training strategy and evaluation criteria, using AP mean, precision, recall, F1 score, and frames per second (FPS), which measures the number of images the model can process per second, as metrics. As shown in Table 3, AVD-YOLO achieves the best performance across all key metrics. Specifically, AVD-YOLO's mAP@0.5 reaches 98.2%, which is an improvement of 1.6% compared to the optimal results among all baseline models. Its precision rate reaches 96.2%, which is 2.8% higher than the highest precision rate in the baseline model. In road safety applications, high recall is particularly crucial as it indicates a lower miss rate, which is vital for preventing traffic accidents. AVD-YOLO achieves a recall rate of 96.5%, which is 2.9% higher than the best baseline model, thereby substantially reducing the risk of missing critical road anomalies. The comprehensive evaluation metric F_1 score also reaches an optimal 96.3%, which is 2.9% higher than the highest F_1 score in the baseline model. Experimental comparisons demonstrate that while AVD-YOLO significantly outperforms representative methods in detection performance, it involves moderately increased computational requirements. AVD-YOLO contains 108.8 M parameters and requires 65.1 GFLOPs for inference. Despite this increased model complexity, the method maintains real-time capability with 54.14 FPS, meeting practical deployment requirements [32].

Table 3. Comparison of examination results of each model.

Methods	P/%	R/%	mAP@0.5/%	F_1/%	FPS
Yolov5 [14]	93.4%	93.6%	96.6%	93.4%	117.08
Yolov8 [15]	91.8%	88.6%	93.5%	90.2%	90.61
Yolov10 [16]	93.4%	90.6%	95.8%	92.0%	62.25
RT-DETR [18]	90.6%	91.1%	92.9%	90.8%	46.32
Yolo11 [17]	90.1%	87.8%	93.3%	88.9%	78.68
Proposed	96.2%	96.5%	98.2%	96.3%	54.14

Notably, compared with the Transformer-based representative model RT-DETR, the mAP @ 0.5 of AVD-YOLO is increased by 5.3%, showing a significant advantage in the current dataset and tasks. These quantitative results clearly show that on the MCRAD containing nine classes of complex road anomaly events, the AVD-YOLO model has significant advantages in terms of detection accuracy, target detection completeness, and the balance between the two compared to the existing mainstream methods, which verifies that the proposed method can handle the anomaly detection task in complex road scenarios more effectively. Notably, mAP@0.5 for AVD-YOLO is 5.3% higher than for the Transformer-based RT-DETR, demonstrating a clear advantage on the current dataset and tasks. These results show that on MCRAD with nine classes of complex anomalies, AVD-YOLO achieves superior accuracy, detection completeness, and balance compared to mainstream methods, confirming its effectiveness for complex road anomaly detection.

4.4. Comparison and Analysis of Visualization Results

AVD-YOLO is qualitatively compared with leading methods on the MCRAD, as shown in Figure 5, using representative scenarios such as small-target foreign object detection, pedestrian detection in complex environments, traffic accident detection, and multi-target hybrid cases.

The performance comparison of different methods in detecting pedestrians, construction signs, and road surface foreign objects within multi-object coexistence scenarios is presented in Figure 5(a1–a4). In such scenarios, the targets are usually presented in small and medium sizes and often have similar color characteristics with the roadside environment, which increases the difficulty of detection, and the comparative analysis shows that AVD-YOLO significantly outperforms the comparative methods in multi-category detection. Figure 5(b1–b4) present the performance differences in the methods in detecting larger foreign objects such as fallen trees on the road. In this case, despite the large size of the foreign object, its irregular shape and high degree of integration with the natural environment pose a challenge for accurate recognition. The results show that AVD-YOLO accurately detects both tree trunks and pedestrians, demonstrating superior scene understanding. In contrast, YOLOv5 and YOLOv11 show significant instability in foreign object detection, while YOLOv10 can detect foreign objects but with significantly low confidence. Figure 5(c1–c4) demonstrate the detection results of pedestrian interaction scenarios in an ordinary road environment. Such scenes are mainly characterized by a large number of targets and partial occlusion, which puts higher requirements on the robustness of the detector. Figure 5(d1–d4) illustrate the performance differences in various methods in detecting roadside explosions. By analyzing the detection and identification results of complex targets such as the explosion core and flame obtained by different methods, the robustness of the model in such extreme, highly dynamic, and strong visual interference scenarios can be evaluated. The anomaly detection performance of the methods in the tunnel environment is demonstrated in Figure 5(e1–e4). The main challenges of the tunnel scenario are uneven lighting conditions and low contrast, which tend to affect the recognition accuracy of small targets. Figure 5(f1–f4) present scenes with severely damaged road surfaces, and the performance of different models varies in such large-area, texture-feature-based anomaly detection tasks. Notably, AVD-YOLO achieves the highest detection confidence and more accurately localizes major damaged areas compared to other methods. The challenging nature of detecting violating vehicles in scenes with similar target-background color characteristics and partial occlusion is demonstrated in Figure 5(g1–g4), which show the detection results for such a composite scene. Comparative analysis demonstrates AVD-YOLO's capability to effectively detect offending vehicles while concurrently yielding superior bounding box localization and enhanced class confidence scores. Figure 5(h1–h4) demonstrate a complex scene where a traffic accident occurs, containing multiple abnormal targets with diverse target categories and complex factors such as large size change and partial occlusion.

4.5. Ablation Study

To evaluate the individual and combined contributions of the proposed PMA and AVD modules, a series of ablation experiments are conducted. Using YOLO as the baseline, PMA and AVD are sequentially integrated while maintaining consistent training settings. Table 4 compares detailed experimental configurations and performance metrics.

Table 4. Ablation analysis of AVD-YOLO.

Methods	**PMA**	**AVD**	**P/%**	**R/%**	**mAP50/%**	$\mathbf{F_1}$
Yolov5			93.4%	93.6%	96.6%	93.4%
A	√		95.4%	96.4%	97.9%	95.9%
B		√	94.8%	94.8%	97.5%	94.8%
C	√	√	96.2%	96.5%	98.2%	96.3%

The symbol "√" indicates that the corresponding module is included.

The experimental results clearly demonstrate the effectiveness of each module. With the PMA module introduced alone, mAP@0.5 increases by 1.3 percentage points to 97.9%, and precision, recall, and F1 score are also improved. This indicates that PMA enhances the model's ability to discriminate targets in complex contexts by improving long-range dependency capture and global context awareness, which is particularly beneficial for detecting small or weakly characterized targets. Secondly, integrating the AVD module alone brings improvements in all metrics compared to the baseline model, mAP@0.5 growing from 96.6% to 97.5%, with improvements in precision, recall and F1. This study preliminarily confirms that AVD improves the extraction and fusion of multi-scale features by fixing the target location and adaptively adjusting the receptive field.

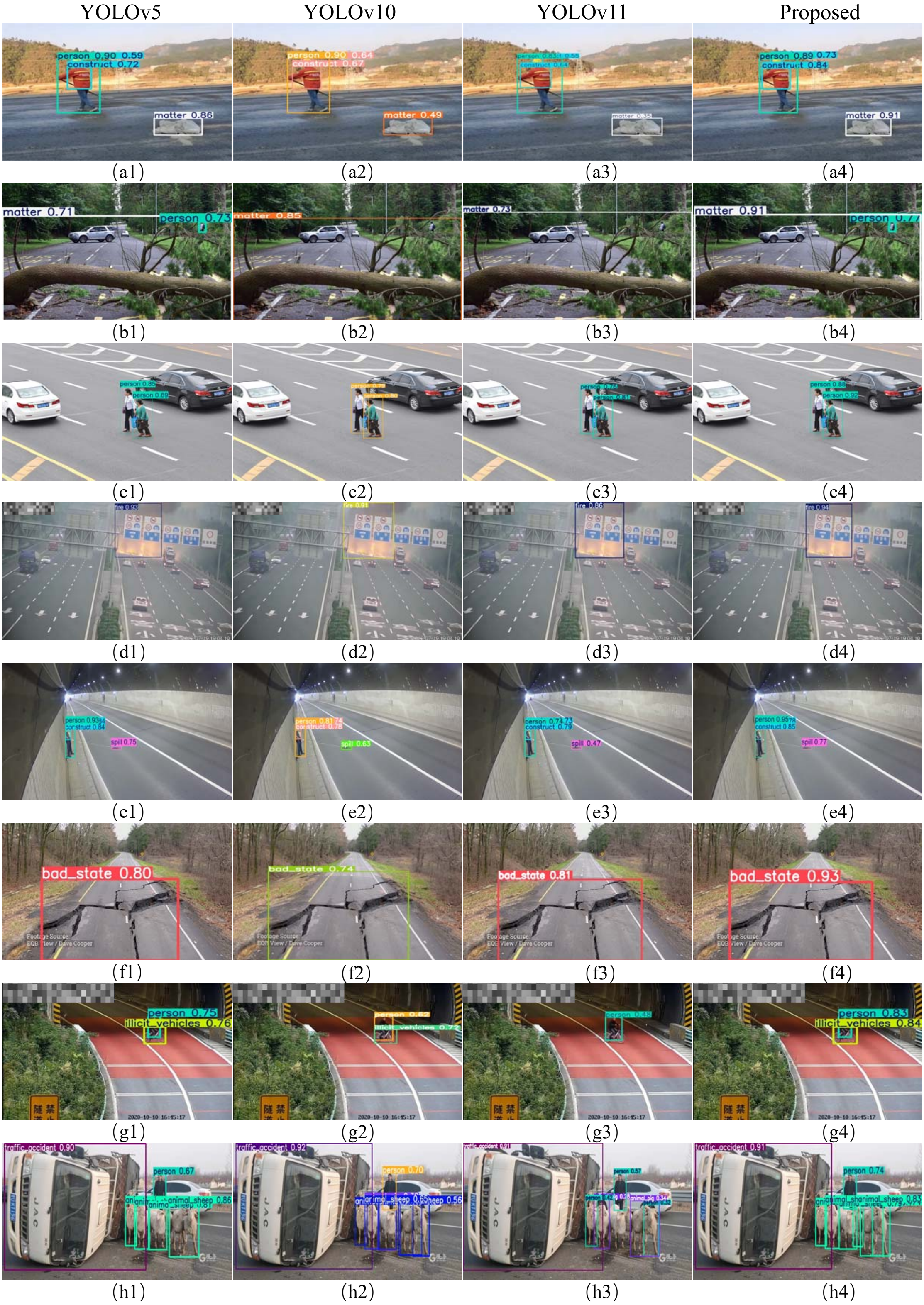

Figure 5. Visualization comparison with other mainstream methods. (**a1**–**a4**) multi-object coexistence scenarios with pedestrians, construction signs, and road surface foreign objects; (**b1**–**b4**) detection of large foreign objects; (**c1**–**c4**) pedestrian interaction scenarios with multiple targets and partial occlusion; (**d1**–**d4**) roadside explosion detection; (**e1**–**e4**) anomaly detection in tunnel environments; (**f1**–**f4**) severely damaged road surface detection; (**g1**–**g4**) violating vehicles detection; (**h1**–**h4**) complex traffic accident scene with multiple anomaly targets.

5. Conclusions

To address challenges posed by complex dynamic scenes, variable viewing distances, and limited data, an Active Vision Driven Multi-scale Feature Extraction for Enhanced Road Anomaly Detection method (AVD-YOLO) is proposed, and a Multi-Class Road Anomaly Dataset (MCRAD) is constructed. A Position-Modulated Attention (PMA) module is developed to enhance detection of weak-feature targets in complex backgrounds, while an Active Vision Driven Multi-scale Feature Extraction (AVD) module is introduced to mitigate scale variations due to varying distances between targets and the camera by adaptively adjusting receptive fields, thus reducing missed detections and localization errors. Experimental results show that a mean Average Precision (mAP) of 98.2% is achieved on the MCRAD across nine anomaly categories, surpassing most existing methods.

Although the proposed method demonstrates superior performance, the improvement in detection accuracy comes at the cost of increased computational complexity. In future work, lightweight model architectures and pruning techniques will be explored to enhance deployment flexibility in resource-constrained scenarios.

Author Contributions: Conceptualization, M.J. and Z.Z.; methodology, M.J. and R.T.; validation, M.J. and A.L.; resources, Z.Z.; data curation, M.J.; writing—original draft preparation, M.J.; writing—review and editing, M.J., Z.Z., R.T., A.L. and Z.Y.; visualization, M.J.; supervision, Z.Z. and R.T.; project administration, Z.Z.; funding acquisition, Z.Z. All authors have read and agreed to the published version of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Ningbo Municipal Major Project of Science and Technology Innovation 2025 (2022Z076); the Zhejiang Provincial Natural Science Foundation of China (LZ24F010004); the Yongjiang Sci-Tech Innovation 2035 (2024Z023, 2024Z122, 2024Z125, 2024Z295, 2025Z040); the National Natural Science Foundation of China (61671412); and the Basic Public Welfare Research Project of Zhejiang Provincial (LGN22F010002).

Data Availability Statement: The results/data/figures in this manuscript have not been published elsewhere, nor are they under consideration (from you or one of your Contributing Authors) by another publisher. The original contributions of this study are presented within the article. Further inquiries can be directed to 2023882054@zwu.edu.cn.

Conflicts of Interest: All authors declare that they have no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AVD-YOLO	Active Vision Driven Multi-scale Feature Extraction for Enhanced Road Anomaly Detection
PMA	Position-Modulated Attention
AVD	Active Vision Driven Multi-scale Feature Extraction
DETR	DEtection TRansformer
YOLO	You Only Look Once
MCRAD	Multi-Class Road Anomaly Dataset
RelPos2d	Two-Dimensional Relative Position Encoding
PM	Positional Modulator
LEPE	Local Enhancement Positional Encoding
DWConv	depthwise separable convolution
FPS	Frames Per Second

References

1. Santhosh, K.K.; Dogra, D.P.; Roy, P.P. Anomaly detection in road traffic using visual surveillance: A survey. *Acm. Com.-Puting Surv. (CSUR)* **2020**, *53*, 1–26. [CrossRef]
2. United Nations Department of Economic and Social Affairs. *World Urbanization Prospects: The 2018 Revision*; United Nations: New York, NY, USA, 2019.
3. World Health Organization. Road Traffic Injuries. 2023. Available online: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries (accessed on 28 February 2023).
4. Moraga, Á.; de Curtò, J.; de Zarzà, I.; Calafate, C.T. AI-Driven UAV and IoT Traffic Optimization: Large Language Models for Con-gestion and Emission Reduction in Smart Cities. *Drones* **2025**, *9*, 248. [CrossRef]
5. Cao, J.; Liu, W.; Xing, W. Dynamic Spatial-Temporal Perception Graph Convolutional Networks for Traffic Flow Forecasting. In *Pattern Recognition and Computer Vision, Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV) Urumqi, China, 18–20 October 2024*; Springer: Singapore, 2024.
6. Ma, Y.; Xu, J.; Gao, C.; Mu, M.; E, G.; Gu, C. Review of research on road traffic operation risk prevention and control. *Int. J. Environ. Res. Public Health* **2022**, *19*, 12115. [CrossRef] [PubMed]
7. Gowthami, C.; Kavitha, S. Comprehensive approach to predictive analysis and anomaly detection for road crash fatalities. *AIP Adv.* **2025**, *15*, 015022. [CrossRef]
8. Zhao, C.; Chang, X.; Xie, T.; Fujita, H.; Wu, J. Unsupervised anomaly detection based method of risk evaluation for road traffic accident. *Appl. Intell.* **2023**, *53*, 369–384. [CrossRef]
9. Zhang, Y.; Lin, L.; Huang, Y.; Wang, X.; Hsieh, S.Y.; Gadekallu, T.R.; Piran, J. A cooperative vehicle-road system for anomaly detection on vehicle tracks with augmented intelligence of things. *IEEE Internet Things J.* **2024**, *11*, 35975–35988. [CrossRef]
10. Ma, J.; Zhao, X.; He, S.; Song, H.; Zhao, Y.; Song, H.; Cheng, L.; Wang, J.; Yuan, Z.; Huang, F.; et al. Review of pavement detection technology. *J. Traffic Transp. Eng.* **2017**, *17*, 121–137.
11. Fan, L.; Zhao, H.; Li, Y. RAO-UNet: A residual attention and octave UNet for road crac k detection via balance loss. *IET Intell. Transp. Syst.* **2022**, *16*, tdac026. [CrossRef]
12. Ma, N.; Fan, J.; Wang, W.; Wu, J.; Jiang, Y.; Xie, L.; Fan, R. Computer vision for road imaging and pothole detection: A state-of-the-art review of systems and algorithms. *Transp. Saf. Environ.* **2022**, *4*, 1–16. [CrossRef]
13. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Detrs beat yolos on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024.
14. Jocher, G. *Ultralytics YOLOv5 (Version 7.0)*; GitHub: San Francisco, CA, USA, 2020. Available online: https://github.com/ultralytics/yolov5s (accessed on 15 November 2023).
15. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLO (Version 8.0.0). 2023. Available online: https://docs.ultralytics.com/zh/models/yolov8/ (accessed on 23 December 2023).
16. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J. Yolov10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2025**, *37*, 107984–108011.
17. Jocher, G.; Qiu, J. *Ultralytics YOLO11 (Version 11.0.0)*; GitHub: San Francisco, CA, USA, 2024. Available online: https://github.com/ultralytics/ultralytics (accessed on 13 February 2025).
18. Srinivasan, A.; Srikanth, A.; Indrajit, H.; Narasimhan, V. A novel approach for road accident detection using DETR algorithm. In *Intelligent Data Science Technologies and Applications, Proceedings of the 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA 2020), Valencia, Spain, 19–22 October 2020*; IEEE: New York, NY, USA, 2020.
19. Liu, J.W.; Yang, D.; Feng, T.W.; Fu, J.J. MDFD2-DETR: A Real-Time Complex Road Object Detection Model Based on Multi-Domain Feature Decomposition and De-Redundancy. *IEEE Trans. Intell. Veh.* **2024**, *10*, 4343–4359. [CrossRef]
20. Liu, Z.; Wu, W.; Gu, X.; Li, S.; Wang, L.; Zhang, T. Application of combining YOLO models and 3D GPR images in road detection and maintenance. *Remote Sens.* **2021**, *13*, 1081. [CrossRef]
21. Yang, Z.; Li, L.; Luo, W. PDNet: Improved YOLOv5 nondeformable disease detection network for asphalt pavement. *Comput. Intell. Neurosci.* **2022**, *2022*, 5133543. [CrossRef] [PubMed]
22. Shuvo, M.M.R.; Dey, A.; Rahman, M.O. A YOLO-Based Framework for Road Sign Detection and Recognition in the Context of Bangladesh. In *2024 IEEE International Conference on Computing, Applications and Systems (COMPAS 2024), Proceedings of the 2024 IEEE International Conference on Computing, Applications and Systems (COMPAS), Cox's Bazar, Bangladesh, 25–26 September 2024*; IEEE: New York, NY, USA, 2024; pp. 1–6.
23. Pei, J.; Wu, X.; Liu, X. YOLO-RDD: A road defect detection algorithm based on YOLO. In *Computer Supported Cooperative Work in Design. International Conference, Proceedings of the 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Tianjin, China, 8–10 May 2024*; IEEE: New York, NY, USA, 2024; pp. 1695–1703.
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing System 30: Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.

25. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
27. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 7464–7475.
28. Li, J.N.; Guan, J.; Wu, W.; Yu, Z.; Yan, R. 2d-tpe: Two-dimensional positional encoding enhances table understanding for large language models. In Proceedings of the ACM on Web Conference 2025, Sydney, Australia, 28 April–2 May 2025; pp. 2450–2463.
29. Zhang, R.; Zhu, F.; Liu, J.; Liu, G. Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis. *IEEE Trans. Inf. Forensics Secur.* **2019**, *15*, 1138–1150. [CrossRef]
30. Lis, K.; Nakka, K.; Fua, P.; Salzmann, M. Detecting the unexpected via image resynthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2152–2161.
31. Arya, D.; Maeda, H.; Ghosh, S.K.; Toshniwal, D.; Sekimoto, Y. RDD2020: An annotated image dataset for automatic road damage detection using deep learning. *Data Brief.* **2021**, *36*, 107133. [CrossRef] [PubMed]
32. Mohankumar, C.E.; Manikandan, A. Decentralized traffic management with Federated Edge AI: A reinforced transnet model for real-time vehicle object detection and collaborative route optimization. *Discov. Appl. Sci.* **2025**, *7*, 729. [CrossRef]