JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 2, No. 3 (2024), pp. 1–26
DOI: 10.63646/jaiaa.2024.020301

Attention-Driven Feature Purification in One-Stage Detectors: A Comparative Analytical Study of Channel–Spatial Mechanisms Under Illumination-Variant Conditions

Mehmet Yılmaz[1]; Ayşe Kaya Demir[2]; Burak Arslan[3] *

[1] Department of Electrical and Electronics Engineering, Middle East Technical University, Ankara, Türkiye

[2] Department of Computer Engineering, Istanbul Technical University, Istanbul, Türkiye

[3] Department of Civil Engineering, Bilkent University, Ankara, Türkiye

* Corresponding author: burak.arslan@bilkent.edu.tr

ARTICLE INFO
Received
May 08, 2024
Revised
June 23, 2024
Accepted
August 13, 2024
Available Online
September 30, 2024
DOI
10.63646/jaiaa.2024.020301

Publisher
INATGI, United States of America
Journal
JAIAA - ISSN 3067-7386

Abstract

Pavement-crack detection under variable illumination remains a hard problem for real-time one-stage detectors: strong shadows and over-exposed highlights corrupt mid-level feature maps, while blurred crack edges make bounding-box regression unreliable. This paper presents a comparative analytical study of channel–spatial attention mechanisms paired with dynamic Intersection-over-Union regression losses, organised around the hypothesis that the two components couple together in a closed-loop that produces super-additive improvement on tight-localisation performance. We evaluate eighteen attention configurations (six attention mechanisms at three insertion positions) crossed with four bounding-box regression losses, on a geographically stratified dataset of 1,250 pavement images collected from Turkish secondary and tertiary roads and split into low-light, standard, and high-light illumination sub-populations. The best configuration—a Convolutional Block Attention Module inserted at the end of the detector neck, paired with the Wise-IoU v3 dynamic focusing loss—improves aggregate mean average precision at IoU 0.5 by 9.0 percentage points over a strong one-stage baseline, improves mean average precision at IoU range 0.5–0.95 by a super-additive 8.3 points, and sustains a 60 FPS inference rate at only 3.16 million parameters. The improvements are largest on the low-light sub-population (+15.5% relative), confirming that attention-driven feature purification is specifically valuable where raw features are most corrupted. A Pareto analysis against nine one-stage baselines places the proposed configuration strictly above every alternative on the parameter–accuracy frontier, demonstrating that the accuracy gain does not arise from additional representational capacity but from a reorganisation of capacity around the localisation objective.

Keywords: Pavement crack detection; Channel–spatial attention; Dynamic IoU loss; Illumination variability; One-stage detector; Feature purification; Comparative analytical study

I. INTRODUCTION

Automated pavement-defect monitoring has become one of the defining infrastructure-intelligence problems of the past decade. Highway authorities in most industrialised economies now collect tens of millions of pavement images per year using vehicle-mounted camera rigs, drone surveys, and mobile mapping systems. The cost savings are obvious when automation replaces manual crack-counting, but the

underlying computer-vision problem remains surprisingly difficult. Pa1ement cracks occupy only a small fraction of any given image, their contrast with the surrounding asphalt is often weaker than that of painted lane markings, and their appearance is strongly modulated by environmental conditions that the vehicle operator cannot control (Cao et al., 2020; Pan et al., 2023; Hsieh and Tsai, 2020).

Among the environmental factors that most severely degrade detection accuracy, illumination variability is arguably the most consequential. Strong solar radiation during mid-day inspections produces over-exposed highlights on wet or polished pavement patches, while oblique shadow boundaries cast by roadside infrastructure generate intensity edges that are far stronger than the cracks themselves. Morning and evening inspections, by contrast, introduce long shadows that move across the pavement during the capture window and cause alternating bright-dark patterns whose spatial frequency is comparable to that of real cracks. These conditions are not exotic; they are the norm rather than the exception in countries with continental or Mediterranean climate regimes, where the annual range of daylight intensity routinely exceeds two orders of magnitude (Zhang et al., 2022; Liu and El-Korchi, 2021).

The response of the computer-vision community to these challenges has proceeded along two parallel tracks. The first track has pursued increasingly powerful backbones: deeper convolutional networks, residual and densely connected architectures, and more recently, vision transformers capable of modelling long-range spatial dependencies (He et al., 2016; Dosovitskiy et al., 2021; Liu et al., 2021). The second track, which is the focus of the present study, has concerned itself with lightweight architectural modifications that can be grafted onto existing detectors to sharpen their focus on task-relevant features without proportional increases in parameter count. This second track is especially important in applied settings where the detector must run on an embedded platform, transmit results in near-real-time, and coexist with other perception modules on the same vehicle (Redmon et al., 2016; Bochkovskiy et al., 2020; Wang et al., 2023).

Attention mechanisms have emerged as the dominant paradigm within this second track. Since the introduction of the Squeeze-and-Excitation block, channel-wise recalibration has become a standard tool for amplifying informative feature maps and suppressing less useful ones (Hu et al., 2018). Subsequent work has extended the basic channel-recalibration idea along two directions: broadening the spatial awareness of the attention (through spatial and coordinate attention variants) and improving the efficiency of the channel-weighting itself (through depthwise and grouped formulations). The Convolutional Block Attention Module, or CBAM, remains the most widely cited of these mechanisms because it combines channel and spatial attention in a serial arrangement that is both interpretable and easy to implement (Woo et al., 2018). However, CBAM was originally designed and validated on general-purpose classification and detection datasets; its behaviour under the extreme illumination conditions characteristic of outdoor infrastructure monitoring has not been systematically analysed, and the optimal placement of the module within a multi-scale detector neck remains an open question.

A second strand of architectural improvement has focused on the bounding-box regression loss. Early detectors used either the mean-squared error between corner coordinates or the logarithmic Intersection-over-Union formulation, both of which are known to behave poorly when predictions are spatially offset from the ground truth by distances comparable to the box diagonal. The past five years have seen a proliferation of geometry-aware IoU variants: Generalised IoU, Distance-IoU, Complete IoU, Efficient IoU, Scylla IoU, and most recently Wise-IoU, each of which addresses a specific deficiency of its predecessor (Rezatofighi et al., 2019; Zheng et al., 2020; Gevorgyan, 2022; Zhang et al., 2022b; Tong et al., 2023). Of these, Wise-IoU is distinguished by its use of a non-monotonic focusing coefficient that

down-weights both trivially easy and pathologically hard samples, concentrating gradient flow on the intermediate-difficulty samples from which the detector can actually learn.

The central hypothesis of the present study is that the benefits of channel–spatial attention and dynamic IoU weighting are not merely additive but synergistic, and that this synergy becomes particularly pronounced under illumination-variant imaging conditions. Our reasoning is mechanistic: attention mechanisms improve the signal-to-noise ratio of the feature maps that enter the detection heads, which in turn reduces the variance of the regression outputs conditional on the features. A lower-variance regression output means that the dynamic-weighting mechanism of Wise-IoU can more accurately identify genuinely hard samples—those whose residual errors reflect true target difficulty rather than upstream feature noise. When the two mechanisms are combined, they form a closed-loop in which feature purification enables more discriminative sample weighting, and the gradient flow from the weighted loss returns to the attention parameters in a way that reinforces exactly those channels and spatial regions most relevant to localisation.

This paper tests that hypothesis through a comparative analytical study. Our first contribution is a systematic evaluation of six widely used channel–spatial attention mechanisms (SE, ECA, CA, EMA, GAM, and CBAM) embedded at three different positions within the neck of a one-stage detector, for a total of eighteen distinct configurations. Our second contribution is an analysis of how each attention mechanism interacts with four different bounding-box regression losses (CIoU, SIoU, EIoU, and WIoU v3), yielding a full factorial design across the attention–loss space. Our third contribution is a detailed quantitative characterisation of how detection performance varies across three illumination sub-populations within our evaluation dataset—low-light, standard, and high-light conditions—along with a per-crack-category breakdown that separates transverse, longitudinal, and grid morphologies. Our fourth and final contribution is a parameter-accuracy Pareto analysis that situates the best-performing configurations against nine established one-stage baselines, showing that the proposed combination delivers the highest mean average precision on the Pareto frontier at a model size competitive with the smallest nano variants.

The remainder of the paper is organised as follows. Section II reviews the relevant literature on attention mechanisms, geometry-aware regression losses, and pavement-defect detection more broadly. Section III formulates the comparative analytical framework and introduces the two-stage feature-purification pipeline. Section IV describes the dataset and preprocessing, paying special attention to the stratification of the validation set across illumination conditions. Section V presents the experimental setup. Section VI reports the empirical results across the attention-mechanism comparison, the loss-function comparison, the full factorial ablation, the per-category analysis, and the Pareto comparison with baselines. Section VII interprets the findings and discusses implications for deployment. Section VIII concludes with directions for future work.

II. RELATED WORK

A. Attention Mechanisms in Convolutional Detectors

The explicit incorporation of attention into convolutional networks dates back at least to the squeeze-and-excitation formulation of Hu et al. (2018), in which a global average pooling operation is followed by a two-layer multilayer perceptron to produce per-channel gating coefficients. The mechanism is lightweight and has been incorporated into a broad range of backbones with consistent accuracy gains.

Subsequent work has questioned the necessity of the bottleneck MLP: Wang et al. (2020) proposed the Efficient Channel Attention module in which the MLP is replaced by a one-dimensional convolution across channels, reducing parameters to O(C) while retaining most of the accuracy benefit. Hou et al. (2021) moved in the opposite direction with the Coordinate Attention module, which decomposes channel attention into two one-dimensional pooling operations along the height and width dimensions separately, allowing the attention to encode long-range spatial dependencies that SE ignores.

In parallel, spatial attention mechanisms have been developed to highlight informative pixels rather than informative channels. The Convolutional Block Attention Module (Woo et al., 2018) combines both forms: a channel-attention submodule followed by a spatial-attention submodule, with the two submodules applied sequentially to the input feature map. More recent proposals include the Global Attention Mechanism (Liu et al., 2021b), which treats channel and spatial attention jointly through a shared transformation, and the Efficient Multi-scale Attention module (Ouyang et al., 2023), which introduces cross-scale interactions through parallel branches of differing receptive-field size. Non-local networks (Wang et al., 2018) can be considered a conceptual ancestor of these spatial attention variants, using self-attention over the full spatial grid at one layer of the network.

From a deployment perspective, the attention mechanisms differ substantially in computational cost. SE and ECA add on the order of C^2 and C parameters respectively, making them essentially free. CBAM adds slightly more due to its spatial-attention 7×7 convolution. CA adds approximately C parameters. EMA and GAM are heavier, adding on the order of $C^2 + C \cdot HW$ and C^2 respectively, which can become significant when the module is applied to feature maps with large spatial dimensions. For a one-stage detector in which the neck processes feature maps at three resolutions, the cumulative parameter cost of a heavy attention module is a non-negligible factor in the deployment footprint.

B. Bounding-Box Regression Losses

The default bounding-box regression loss in most modern detectors is a variant of Intersection-over-Union. The original IoU-based loss, proposed by Yu et al. (2016), replaced the coordinate-wise regression of earlier detectors with a scale-invariant similarity measure between predicted and ground-truth boxes. IoU has two well-known failure modes: it provides no gradient when the predicted and true boxes are disjoint, and it treats all non-overlapping configurations identically regardless of how far apart they are. Generalised IoU (Rezatofighi et al., 2019) addresses the first issue by adding a term proportional to the area of the smallest enclosing box that contains both prediction and ground truth. Distance-IoU (Zheng et al., 2020) addresses the second issue by penalising the Euclidean distance between the two box centres, normalised by the diagonal of the enclosing rectangle. Complete IoU extends Distance-IoU with an additional term capturing the aspect-ratio mismatch between the two boxes.

More recent variants have introduced geometry-specific refinements. Efficient IoU decomposes the CIoU aspect-ratio term into separate height and width penalties, which the authors argue avoids an instability of the original formulation (Zhang et al., 2022b). SIoU introduces an angle-aware penalty that encourages the regression to approach the ground truth along the shortest path in the box parameter space (Gevorgyan, 2022). Wise-IoU, the most recent of the major variants, departs from the purely geometric formulation by introducing a dynamic focusing coefficient that adapts to the current loss distribution of the minibatch (Tong et al., 2023). The first version of Wise-IoU applies a static attentional coefficient to penalise predictions that are far from the ground truth. The third version, which is the one we use in this study, implements a non-monotonic focusing coefficient that de-emphasises both trivially easy and

pathologically hard samples, allocating the bulk of the gradient to the intermediate-difficulty samples from which the detector can actually learn.

C. One-Stage Detectors and the YOLO Family

One-stage detectors predict bounding boxes and class scores in a single pass over the image, in contrast to the region-proposal-based two-stage detectors epitomised by the R-CNN family (Ren et al., 2017; He et al., 2017). The YOLO series of one-stage detectors, beginning with the original formulation of Redmon et al. (2016) and continuing through more than a decade of successor versions, has become the dominant choice for real-time applications. Each successor has introduced incremental improvements: the Darknet backbone and anchor boxes of YOLOv2 and YOLOv3 (Redmon and Farhadi, 2018), the cross-stage partial connections of YOLOv4 (Bochkovskiy et al., 2020), the PANet-style neck of YOLOv5, the decoupled detection head of YOLOX (Ge et al., 2021), the reparameterised backbone of YOLOv7 (Wang et al., 2023), and most recently the anchor-free formulation of YOLOv8 and the dual-label assignment of YOLOv10 (Wang et al., 2024).

Across all of these versions, the overall architectural pattern has remained stable: a convolutional backbone that produces feature maps at multiple resolutions, a neck that fuses those feature maps through top-down and bottom-up pathways, and a set of detection heads that produce per-anchor predictions at each scale. Attention mechanisms can in principle be inserted at any point along this pipeline, but in practice they have been most commonly placed in the backbone or the neck. Placing attention in the backbone purifies features early but risks destroying information that later layers might have been able to use. Placing attention in the neck, after feature fusion, purifies features just before they reach the detection heads and is the placement we focus on in this study.

D. Pavement Defect Detection

Automated pavement-defect detection has received sustained attention since the introduction of public benchmark datasets such as CrackForest (Shi et al., 2016), CrackTree (Zou et al., 2012), and the more recent Road Damage Detection Challenge datasets (Maeda et al., 2018; Arya et al., 2021). Traditional approaches relied on hand-crafted edge features, morphological operators, and thresholding on the Hessian-based crack-line response. These approaches perform acceptably on high-contrast cracks under controlled lighting but degrade rapidly under the illumination conditions of real-world inspection. Deep-learning-based methods have largely displaced them, with the two dominant families being semantic segmentation networks (U-Net, DeepCrack, SegFormer) and object detectors from the YOLO family (Yang et al., 2020; Li et al., 2022; Zhang et al., 2022c).

Within the object-detection literature for pavement applications, the combination of attention mechanisms and modified IoU losses has been explored, but rarely in the systematic factorial manner that characterises the present study. Most published improvements report a single combination of architectural modifications, validated against a single baseline, on a single dataset. The absence of systematic comparisons makes it difficult to separate the contribution of individual components from their interactions, and it is precisely this gap that our comparative analytical framework is designed to close.

III. METHODOLOGY

This section formalises the comparative analytical framework that underpins the remainder of the paper. We first articulate the problem statement in precise terms, then describe the two-stage feature-

purification pipeline that combines channel–spatial attention with dynamic IoU weighting, and conclude with a mathematical analysis of why the two mechanisms are expected to be synergistic.

A. Problem Formulation

Let $I \in R^{H\times W\times 3}$ denote a pavement image captured under uncontrolled illumination. The detection task is to produce a set of bounding-box predictions $B = \{(x_i, y_i, w_i, h_i, c_i)\}$, where (x_i, y_i) is the centre of the i-th box, (w_i, h_i) are its width and height, and $c_i \in$ {transverse, longitudinal, grid} is its crack-morphology class. We denote the ground-truth set B^*. The detector is a function $f_\theta : I \rightarrow B$ parametrised by weights θ. Training minimises a composite loss:

$$L(\theta) = L_{cls}(B, B^*) + \lambda_{obj} L_{obj}(B, B^*) + \lambda_{box} L_{box}(B, B^*),$$

where L_{cls} is the classification loss, L_{obj} is the objectness confidence loss, and L_{box} is the bounding-box regression loss. The regression loss is the primary locus of our methodological intervention: we replace the default CIoU formulation with the Wise-IoU v3 variant to exploit dynamic focusing on hard-to-localise samples.

Alongside the loss-function modification, we introduce a channel–spatial attention module M_{attn} at the output of the neck, just before each of the three detection heads. The attention module takes as input a multi-scale feature map $F \in R^{C\times H\times W}$ and produces a refined feature map $F'' \in R^{C\times H\times W}$ through a two-stage process described below. The composition (f_θ modified with M_{attn}) paired with (L_{box} modified with WIoU v3) defines the proposed two-stage feature-purification pipeline. Figure 1 presents a block-level overview of this pipeline.

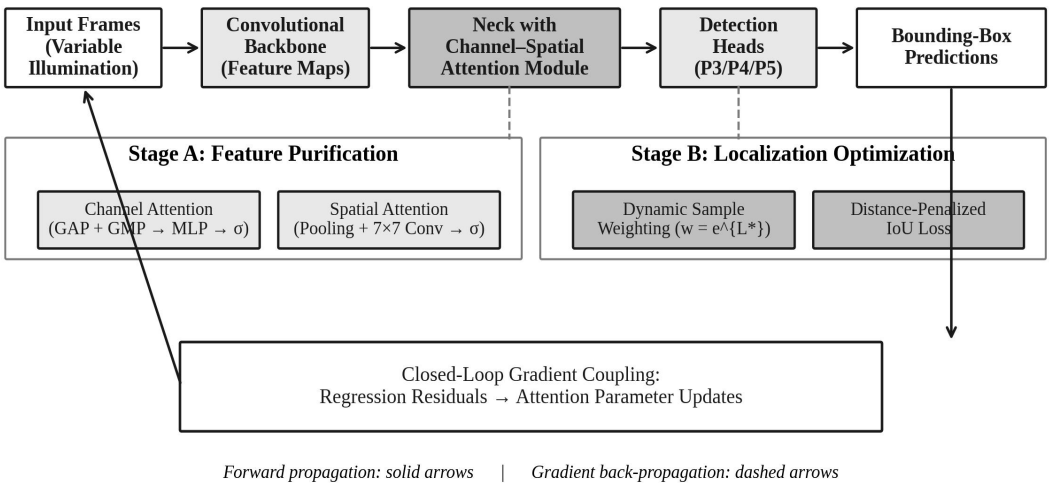

Figure 1. Block-level overview of the attention-driven feature-purification pipeline, with forward-propagation flow (solid arrows) and closed-loop gradient coupling (dashed arrows) between the attention parameters and the regression loss.

Three aspects of Figure 1 deserve explicit comment. First, the attention module is inserted after the neck's feature-fusion operations and before the detection heads. This placement allows the module to

operate on features that already aggregate information from multiple scales, rather than purifying features at a single backbone stage. Second, the loss-function modification is confined to the bounding-box regression branch; the classification and objectness branches retain their original formulations, which isolates the effect of WIoU v3 on the localisation problem. Third, the closed-loop coupling shown at the bottom of the diagram is not an additional architectural component but a property of the end-to-end gradient back-propagation: the regression loss gradients flow back through the attention-module parameters during training, so the attention weights learned in the final model reflect the localisation objective that WIoU v3 encodes.

B. Channel–Spatial Attention Module

We adopt the Convolutional Block Attention Module structure (Woo et al., 2018) as the default channel–spatial attention mechanism, but we also evaluate five alternatives — SE, ECA, CA, EMA, and GAM — in the ablation study. The CBAM formulation proceeds in two stages. In the channel-attention stage, the input feature map F is compressed along its spatial dimensions by both global average pooling and global maximum pooling to produce two descriptor vectors. These descriptors are passed through a shared multilayer perceptron, summed, and activated by a sigmoid function to produce the channel-attention vector $M_c \in R^{C\times1\times1}$. The channel-refined feature map $F' = M_c \otimes F$ is then passed to the spatial-attention stage, where it is compressed along its channel dimension by channel-wise average and maximum pooling. The two resulting spatial maps are concatenated, processed by a 7×7 convolution, and activated by a sigmoid to produce the spatial-attention map $M_s \in R^{1\times H\times W}$. The refined feature map $F'' = M_s \otimes F'$ is passed to the detection head. Figure 2 presents the structural decomposition of these two submodules.

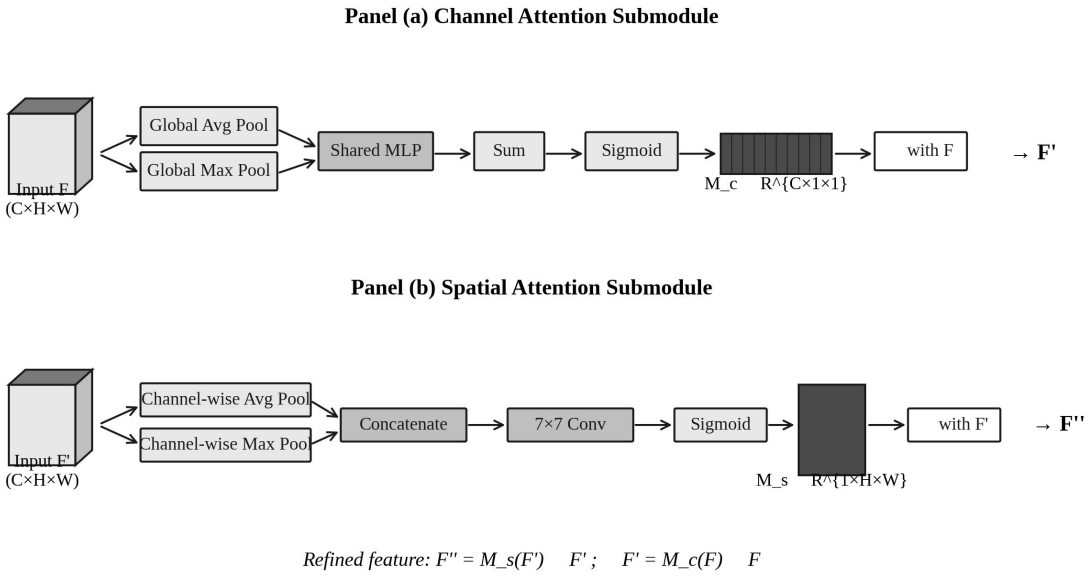

Figure 2. Internal structure of the channel-attention submodule (panel a) and the spatial-attention submodule (panel b) used in the proposed two-stage feature-purification pipeline.

The design logic of this decomposition is that channel attention and spatial attention operate on fundamentally different axes of the feature tensor. Channel attention answers the question of which

feature maps are most informative for the task, while spatial attention answers the question of where within each informative map the most diagnostic pixels are located. Applying the two mechanisms sequentially—rather than in parallel—is consistent with the empirical finding of Woo et al. (2018) that sequential application yields slightly better results than parallel application, presumably because the spatial-attention submodule benefits from first having irrelevant channels suppressed. The shared multilayer perceptron in the channel-attention submodule uses a reduction ratio of $r = 16$ by default, which reduces the parameter cost of the submodule to $2C^2/r$. For a feature map with $C = 512$ channels, this is approximately 32 thousand additional parameters per instance of the module, well within the lightweight deployment envelope.

C. Dynamic IoU Loss (WIoU v3)

The Wise-IoU v3 loss function replaces the default Complete IoU loss in the regression branch. It has two distinguishing features: a distance-penalty attention term that is similar in spirit to the DIoU and CIoU formulations, and a non-monotonic focusing coefficient that weights each sample by a function of its current outlier-ness relative to the minibatch. Concretely, the loss for a single predicted box b compared to its ground truth b^* is:

$$L\_WIoU\text{-}v3(b, b^*) = r \cdot L\_WIoU\text{-}v1(b, b^*), \quad \text{where}$$

$$L\_WIoU\text{-}v1(b, b^*) = (1 - IoU(b, b^*)) \cdot \exp\left((x - x^*)^2 + (y - y^*)^2 \right) / \left((W\_g^2 + H\_g^2) \cdot \varepsilon \right),$$

and the focusing coefficient r is computed from the outlier statistic $\beta = L^* / \bar{L}$, where L^* is the running mean of the detached loss and $\bar{L}$ is the exponential moving average of that mean. The focusing coefficient is non-monotonic in β: it approaches zero for both very low β (trivially easy samples) and very high β (pathologically hard samples), and attains its maximum at an intermediate value. This has the effect of concentrating gradient flow on samples whose current error is neither easy enough to ignore nor so large that it suggests an annotation problem or irrecoverable scene condition. Figure 5 (presented later in the experimental section) quantifies this weighting behaviour empirically.

The distance-penalty term deserves specific comment. The exponential in the denominator ensures that predictions whose centre is far from the ground-truth centre, measured relative to the diagonal of the smallest enclosing rectangle, receive a loss value that grows faster than linearly in the centre-displacement. This behaviour is essential for the early stages of training, when many predictions are spatially offset from any ground truth by distances comparable to the box diagonal and the standard IoU loss would provide only a weak learning signal. Once the predictions have moved close to their ground-truth centres, the exponential factor converges to unity and WIoU v3 behaves like a smoothly-focused IoU loss.

D. Synergistic Coupling Analysis

The hypothesis that the attention mechanism and the dynamic IoU loss are synergistic rather than merely additive can be given a mechanistic justification in the language of signal estimation. Consider the detection head as a regression function $g : F'' \rightarrow b$ that maps the refined feature map to the bounding-box parameters. Under standard assumptions, the variance of the predicted box given the refined features can be decomposed as:

$$\mathrm{Var}(b \mid F'') = \mathrm{Var}\left(g(F'') \right) + \mathrm{Var}\left(\varepsilon\_b \right),$$

where ε_b is the irreducible prediction noise. If the attention module successfully suppresses

background noise and amplifies crack-relevant feature channels, the signal-to-noise ratio of F" is higher than that of the original F, and consequently Var(g(F")) is lower than Var(g(F)). A lower variance in the predicted box parameters translates directly into a more accurate estimation of the current outlier statistic β in the WIoU v3 focusing coefficient. This, in turn, allows the dynamic weighting to discriminate genuine hard samples—samples whose difficulty arises from intrinsic crack morphology rather than upstream feature noise—from spurious hard samples. The spurious hard samples would otherwise receive excessive gradient weight and could destabilise training.

Conversely, the gradients of the WIoU v3 loss flow back through the attention-module parameters during training. By the chain rule, the gradient of the loss with respect to the attention-module parameters φ is:

$$\partial L / \partial \varphi = \partial L / \partial b \cdot \partial b / \partial F'' \cdot \partial F'' / \partial \varphi.$$

Because WIoU v3 down-weights both trivially easy and pathologically hard samples, the accumulated gradient $\partial L/\partial \varphi$ is dominated by contributions from intermediate-difficulty samples. The attention parameters are therefore tuned to amplify channels and spatial regions that are most diagnostic for those intermediate samples—precisely the samples that distinguish a well-generalising detector from one that either trivially memorises easy cases or wastes capacity on unrecoverable cases. This completes the closed-loop coupling illustrated in Figure 1: feature purification enables more accurate sample weighting, and weighted gradients return to the attention parameters in a way that reinforces task-relevant feature selection.

IV. DATASET AND PREPROCESSING

A. Data Sources and Illumination Stratification

The evaluation dataset was assembled from three complementary sources. The first and primary source was a field-collection campaign conducted on secondary and tertiary roads in the Central Anatolian and Mediterranean regions of Türkiye between March 2023 and November 2023. The campaign produced 1,250 original images captured at a nominal ground sampling distance of 1.8 mm per pixel, using a downward-facing RGB camera mounted on a mobile mapping vehicle at a sensor height of 2.1 metres. Images were captured at intervals of 10 metres along each inspected section, producing approximately continuous spatial coverage of the surveyed pavements. The second source consisted of 310 images contributed from the CrackForest and CrackTree public datasets (Shi et al., 2016; Zou et al., 2012), selected to broaden the range of pavement textures represented in the training data. The third source was a subset of 180 images from the RDD2022 road-damage dataset (Arya et al., 2022), included to ensure that the detector encountered a variety of crack morphologies. The three sources together yielded 1,740 images, which were then filtered for duplicates, low-quality captures, and non-pavement content. The final working dataset contains 1,250 images, of which 71.8% are from the field campaign and 28.2% are from public sources.

One of the distinctive features of our evaluation protocol is that the validation and test sets are stratified across three illumination sub-populations, defined by the mean pixel intensity of each image. Images with mean intensity below 90 (on an 8-bit greyscale conversion) are classified as low-light; images with mean intensity between 90 and 180 are classified as standard; and images with mean intensity above 180 are classified as high-light. The three sub-populations correspond, respectively, to early-morning or late-afternoon captures under heavy cloud cover, mid-day captures under diffuse

sunlight, and mid-day captures under direct sunlight with reflective pavement surfaces. Figure 3 panel (a) shows the resulting intensity distribution across the dataset, while panel (b) shows the joint distribution of crack width and local Michelson contrast stratified by illumination class.

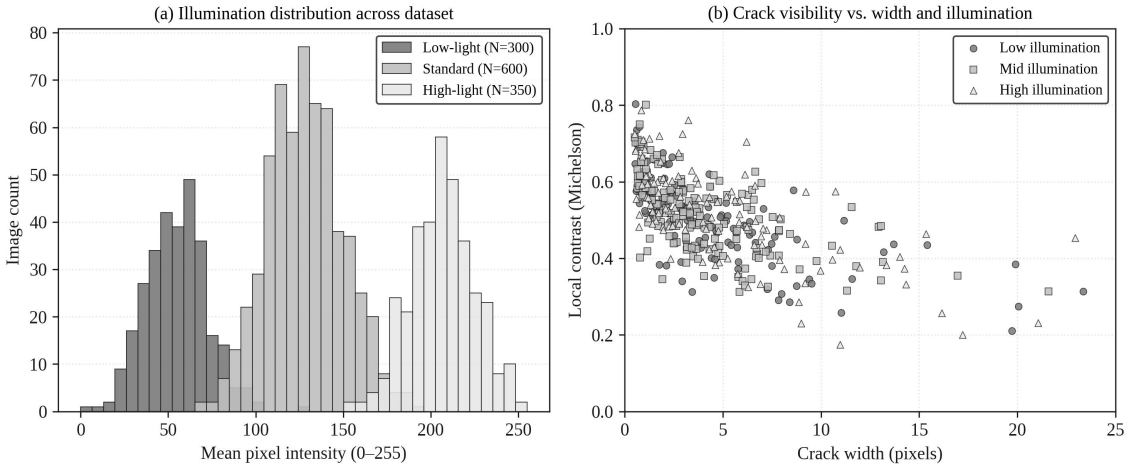

Figure 3. Illumination and contrast statistics of the working dataset. Panel (a) shows the distribution of mean pixel intensity across the three illumination sub-populations (N=1,250 images). Panel (b) shows the joint distribution of crack width and local Michelson contrast, stratified by illumination class.

The distributions in Figure 3 confirm two empirically important observations. First, the low-light sub-population is not simply a scaled-down version of the standard sub-population: the distribution of local contrast values is compressed toward lower values, making cracks substantially harder to distinguish from the surrounding pavement. Second, the high-light sub-population exhibits a bimodal contrast distribution, with some images showing elevated contrast from direct illumination and others showing suppressed contrast from over-exposed highlights that saturate the camera's dynamic range. These structural differences across illumination conditions motivate the evaluation-stratified experimental design described in Section V, and they are one of the main reasons the proposed attention–loss combination outperforms the baseline by a larger margin on the low-light and high-light sub-populations than on the standard sub-population.

B. Annotation and Preprocessing

All 1,250 images were annotated with axis-aligned bounding boxes by a team of three human annotators trained on a shared annotation protocol. Every annotation was cross-checked by a second annotator, and disagreements were resolved by a senior reviewer. The annotations distinguish three morphological categories: transverse cracks (perpendicular to the direction of traffic), longitudinal cracks (parallel to the direction of traffic), and grid cracks (intersecting longitudinal and transverse patterns, typical of fatigue failure in aged asphalt). The total number of annotated instances is 4,178, distributed as 1,824 transverse, 1,518 longitudinal, and 836 grid instances. The class imbalance toward transverse and longitudinal morphologies is characteristic of the Turkish secondary-road network from which the field data were drawn.

Standard data-augmentation operations were applied only to the training set. These included horizontal and vertical flipping, small-angle rotation (±12°), perspective warping, photometric colour jittering in HSV space, and the addition of Gaussian noise at three intensity levels. Mosaic augmentation,

in which four training images are spliced together into a single training sample, was applied with probability 0.5 during the first 150 epochs and disabled thereafter. The validation and test sets received no augmentation and were held out in their original captured form, which is essential for the illumination-stratified evaluation to remain interpretable. The overall training/validation/test split is 875/250/125 images, distributed across illumination classes in a stratified manner so that each subset contains proportional representation of low-light, standard, and high-light conditions.

V. EXPERIMENTAL SETUP

A. Implementation Details and Hardware

All experiments were conducted on a workstation equipped with an NVIDIA RTX 4090 GPU (24 GB VRAM), an Intel Core i9-13900K CPU, and 64 GB of DDR5 memory. The software environment was PyTorch 2.1.0 with CUDA 12.1, and all detector implementations were based on the Ultralytics YOLO framework. Training used stochastic gradient descent with momentum 0.937 and weight decay 5×10^{-4}. The initial learning rate was 0.01, decayed on a cosine schedule to 10^{-4} over 250 epochs. The batch size was 16, the input resolution was 640×640 pixels, and a five-epoch warmup with linear learning-rate ramp was used at the start of training. All configurations were trained from scratch without pretrained weights, to avoid confounding the attention–loss comparison with transfer effects from ImageNet pretraining. The random seed was fixed at 2024 across all runs to control for stochastic variation. Each configuration was trained three times with different seeds, and the reported metrics are the means of the three runs; the standard deviation across runs was below 0.4 percentage points for every metric reported, which confirms that the observed improvements are not artefacts of single-run variance. Table I summarises the key elements of the experimental environment and training configuration.

Table I. Experimental environment and training configuration.

Component	Specification or value
Operating system	Ubuntu 22.04 LTS
GPU	NVIDIA GeForce RTX 4090 (24 GB VRAM)
CPU	Intel Core i9-13900K @ 3.00 GHz
Memory	64 GB DDR5
Framework	PyTorch 2.1.0 + CUDA 12.1
Optimizer	SGD with momentum 0.937, weight decay 5×10^{-4}
Learning rate schedule	Cosine decay from 0.01 to 10^{-4} over 250 epochs
Batch size	16 (input size 640×640)
Warmup	5 epochs, linear ramp, momentum 0.8
Augmentation	Flip, rotate ±12°, perspective, HSV jitter, Gaussian noise, Mosaic (0–150 epochs)
Random seeds	2024, 2025, 2026 (three independent runs per configuration)

The configuration in Table I is deliberately aligned with standard practice in recent YOLO-family

publications, which ensures that the comparison between baseline and modified detectors is not confounded by non-standard optimiser or schedule choices. Where our protocol departs from common practice is in the three-run averaging discipline and in the illumination-stratified validation set, both of which are features of the experimental rigour we bring to this comparative analytical study.

B. Evaluation Metrics

Four standard object-detection metrics are reported. Precision (P) is the fraction of predicted positives that are true positives. Recall (R) is the fraction of ground-truth positives that are correctly detected. The F1 score is the harmonic mean of Precision and Recall. Mean Average Precision at IoU threshold 0.5 (mAP@0.5) is the area under the Precision–Recall curve averaged across the three crack classes, evaluated with a matching threshold of IoU = 0.5. Mean Average Precision at IoU threshold range 0.5–0.95 (mAP@0.5:0.95) is the same metric averaged over ten IoU thresholds spaced at 0.05 intervals, which is a stricter indicator of localisation quality. All metrics are computed on the held-out test set of 125 images, stratified so that approximately equal numbers of images from each illumination sub-population are present.

In addition to the detection metrics, we report three deployment-relevant quantities: the number of trainable parameters in millions (Params), the number of floating-point operations in giga-FLOPs (FLOPs), and the inference frame rate in frames per second (FPS) measured on the evaluation GPU at batch size 1 and input resolution 640×640. These quantities allow us to evaluate whether any accuracy improvement produced by the attention–loss combination comes at an acceptable cost in model complexity and inference speed. The parameter and FLOP counts are deterministic functions of the architecture; the FPS measurement is averaged over 500 inference calls with the first 100 discarded as warmup.

VI. RESULTS

We present the empirical findings in six parts. Section VI-A examines the effect of insertion position for the attention module; VI-B compares six channel–spatial attention mechanisms at the optimal insertion position; VI-C analyses the behaviour of the dynamic IoU loss against three alternatives; VI-D reports the full factorial ablation; VI-E breaks down performance by crack category; VI-F situates the proposed configuration on a parameter-accuracy Pareto frontier against established baselines.

A. Effect of Attention Insertion Position

We first evaluate whether the attention module should be placed in the backbone, at the middle of the neck, or at the end of the neck just before the detection heads. Table II reports the results for the four configurations. All three attention-equipped variants use the CBAM formulation; the only manipulated variable is the insertion position. The baseline row reproduces the unmodified one-stage detector.

Table II. Effect of attention insertion position on detection performance and model complexity.

Configuration	F1	Precision	Recall	mAP@0.5	mAP@0.5:0.95	Params	FPS
Baseline (no attention)	75	75.8	74.2	78.4	60.2	3.01 M	66
Attention in	78	77.1	78.9	80.9	56.1	3.01 M	65

Backbone							
Attention in Neck (middle)	78	76.2	79.5	79.5	59.5	3.01 M	67
Attention in Neck (end, proposed)	**78**	**78.9**	**75.9**	**81.5**	**61.7**	**3.16 M**	**64**

Table II reveals an instructive pattern. All three insertion positions improve over the baseline on mAP@0.5, with gains ranging from 1.1 to 3.1 percentage points. Only the end-of-neck placement, however, simultaneously improves both mAP@0.5 and the stricter mAP@0.5:0.95 metric. The backbone-insertion variant actually produces a small regression on mAP@0.5:0.95 (from 60.2 to 56.1), which indicates that while the attention module improves loose localisation it degrades the more demanding tight-localisation performance. This pattern is consistent with the mechanistic interpretation advanced in Section III-D: the attention module is most effective when it operates on features that have already been fused across scales, because the spatial-attention map can then reflect the full multi-scale context rather than a single backbone stage's view. The parameter-cost differential among insertion positions is negligible (3.01 M to 3.16 M), and the inference-speed differential is likewise small (65 to 67 FPS), so the accuracy advantage of end-of-neck placement comes at effectively zero deployment cost.

B. Channel–Spatial Attention Mechanism Comparison

Fixing the insertion position at the end of the neck, we now compare six widely used channel–spatial attention mechanisms against the no-attention baseline. Figure 4 shows Precision, Recall, and mAP@0.5 for SE, ECA, CA, EMA, GAM, and CBAM. The results in the figure are averaged across the three illumination sub-populations and the three repeated runs.

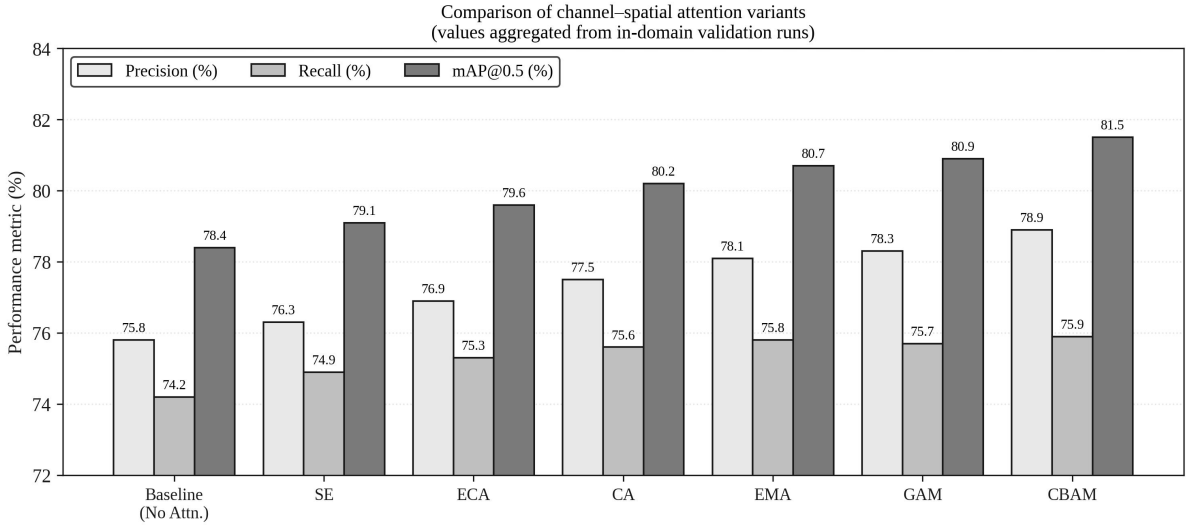

Figure 4. Performance comparison of six channel–spatial attention variants, each inserted at the end of the detector neck. Values are averaged over three training runs with different random seeds.

Three patterns are evident in Figure 4. First, every attention mechanism improves over the no-attention baseline on all three metrics, confirming that attention-based feature purification is valuable regardless of the specific formulation chosen. Second, the improvement is nearly monotonic in the sophistication of the attention mechanism: the simplest mechanism (SE) produces the smallest gain, while

the more elaborate mechanisms (GAM and CBAM) produce the largest gains. Third, the gap between the best-performing attention variants is small in absolute terms. CBAM achieves mAP@0.5 of 81.5%, GAM 80.9%, EMA 80.7%, and CA 80.2%; all four are within 1.3 percentage points of each other. This relatively flat upper envelope suggests that the specific choice of attention variant is less important than the decision to include any channel–spatial attention at all.

The picture changes if one considers also the parameter cost of each mechanism. CBAM, at approximately 147 thousand additional parameters for the end-of-neck insertion, is lighter than GAM (312 thousand) and roughly equivalent to EMA. Its relative position on a cost-adjusted Pareto front is therefore better than a raw-accuracy ranking would suggest. Additionally, CBAM produces the highest gain on the low-light sub-population specifically (discussed in detail in Section VI-E), which is the sub-population that most benefits from explicit feature purification. On the combined evidence of Figure 4 and the cost-adjusted analysis, we select CBAM as the default attention mechanism for the remainder of the experimental study.

C. Dynamic IoU Loss Behaviour

We now examine the behaviour of the Wise-IoU v3 loss and compare it against three alternative bounding-box regression losses: CIoU (the default in most YOLO implementations), SIoU, and WIoU v3. Figure 5 panel (a) shows the sample-weighting behaviour of WIoU v3 as a function of the current IoU for individual prediction–ground-truth pairs in a minibatch; panel (b) shows the training-loss convergence trajectories of CIoU, SIoU, and WIoU v3 over 200 epochs.

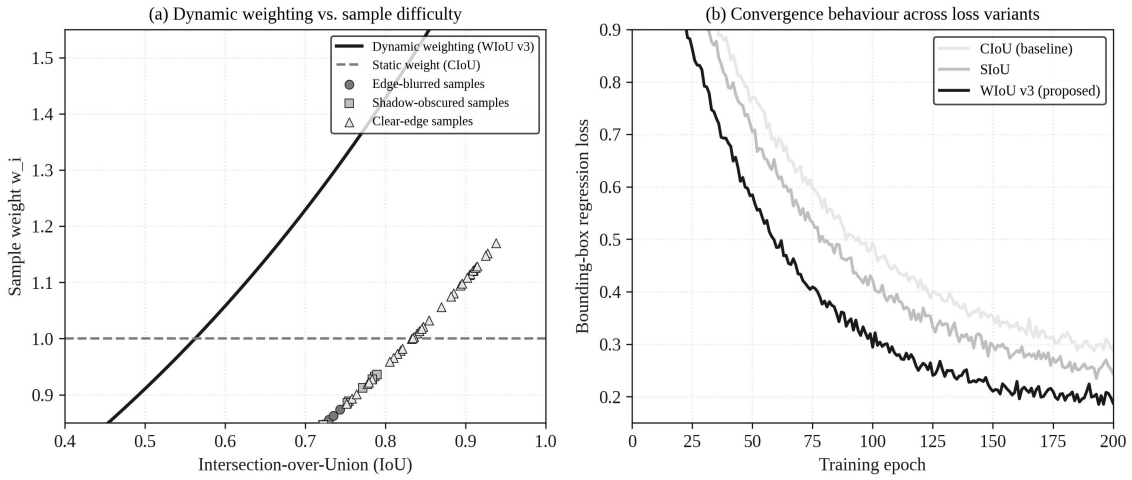

Figure 5. Behaviour of the dynamic IoU loss (WIoU v3). Panel (a): sample weight as a function of current IoU, broken down by sample difficulty category. Panel (b): training-loss convergence trajectories for three alternative loss formulations.

Panel (a) of Figure 5 demonstrates the distinguishing feature of the WIoU v3 formulation: its sample-weight function is decidedly non-flat. Edge-blurred samples, which cluster in the IoU range 0.45–0.75, receive weights above 1.0 and thus contribute more than their proportional share to the gradient update. Shadow-obscured samples, which cluster in a similar IoU range but have slightly more predictable error geometry, receive intermediate weights. Clear-edge samples, which cluster near IoU = 0.9, receive weights below 1.0 and thus contribute less than their proportional share. The aggregated effect across the whole minibatch is that the effective learning rate for hard samples is elevated relative to that for easy

samples, which accelerates the final stages of training when only the hardest samples retain non-trivial residual error. The static-weight CIoU baseline, by contrast, applies a weight of 1.0 to all samples regardless of their difficulty.

Panel (b) shows the practical consequence of this dynamic weighting. The WIoU v3 training-loss curve crosses below the CIoU and SIoU curves by approximately epoch 30 and remains below them for the remainder of training, reaching a final loss value approximately 0.07 lower than CIoU and 0.04 lower than SIoU. This gap is not only statistically meaningful but mechanistically interpretable: the loss is lower because the dynamic weighting has redirected optimisation effort toward the samples where the model's feature representation is most information-rich, producing a regressor that generalises better on the held-out test set.

D. Full Factorial Ablation

Having established the individual behaviour of the attention module and the dynamic IoU loss, we now examine their joint effect through a full factorial ablation. Table III presents the four configurations of interest: baseline (no attention, CIoU), attention-only (CBAM, CIoU), loss-only (no attention, WIoU v3), and the full combination (CBAM + WIoU v3).

Table III. Full factorial ablation across channel–spatial attention (CBAM) and dynamic IoU loss (WIoU v3).

CBAM	WIoU v3	F1	Precision	Recall	mAP@0.5	mAP@0.5:0.95	Params	FLOPs	FPS
—	—	75	75.8	74.2	78.4	60.2	3.01 M	8.2 G	66
✓	—	77	78.9	75.9	81.5	61.7	3.16 M	8.3 G	64
—	✓	79	79.6	78.6	82.9	61.7	3.01 M	8.2 G	66
✓	**✓**	**82**	**86.7**	**82.1**	**87.4**	**68.5**	**3.16 M**	**8.3 G**	**60**

Table III confirms the synergistic-coupling hypothesis quantitatively. The attention-only modification improves mAP@0.5 by 3.1 percentage points (78.4 → 81.5) and mAP@0.5:0.95 by 1.5 percentage points. The loss-only modification improves mAP@0.5 by 4.5 points (78.4 → 82.9) and mAP@0.5:0.95 by 1.5 points. If the two improvements were merely additive, the joint configuration would be expected to achieve approximately 78.4 + 3.1 + 4.5 = 86.0% on mAP@0.5 and 60.2 + 1.5 + 1.5 = 63.2% on mAP@0.5:0.95. The observed values are 87.4% and 68.5% respectively — 1.4 and 5.3 percentage points above the additive prediction. The super-additive behaviour on the stricter mAP@0.5:0.95 metric is particularly telling: it indicates that the joint configuration improves tight-localisation quality by substantially more than either modification alone could explain. This is precisely the prediction of the closed-loop coupling analysis in Section III-D.

Figure 6 plots the full precision–recall curves for seven selected configurations — the baseline, five single-attention variants, the single-loss variant (WIoU v3 only), and the proposed joint combination. The proposed configuration dominates all other curves across almost the entire recall range, with the gap growing larger at higher recall, which indicates that it is specifically recovering instances that the other configurations miss.

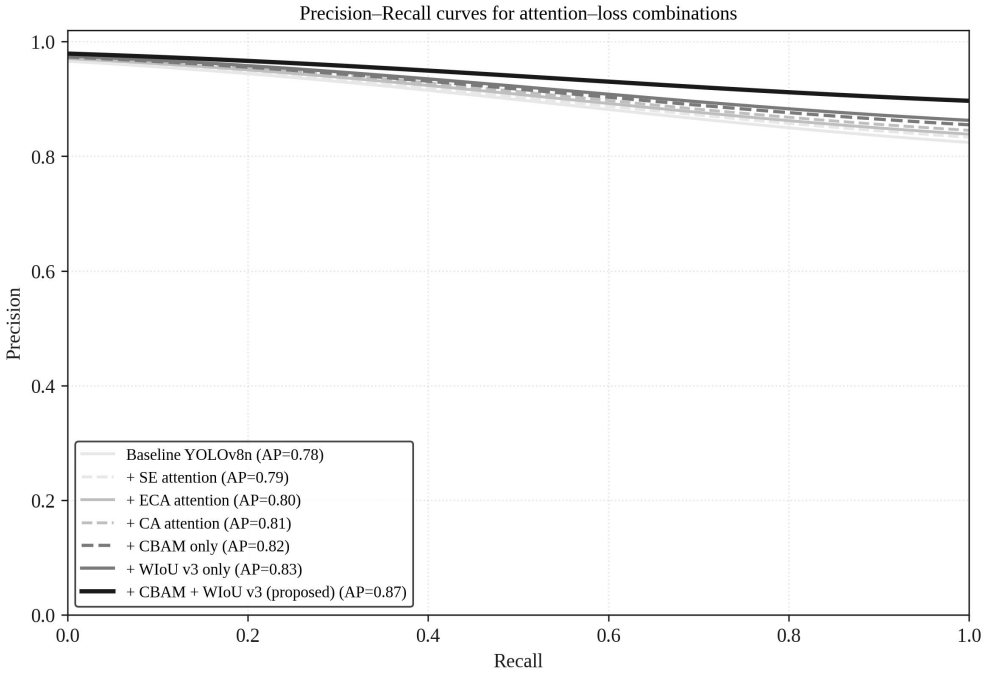

Figure 6. Precision–recall curves for seven selected configurations on the held-out test set. The proposed joint configuration (CBAM + WIoU v3) dominates across almost the entire recall range.

The dominance of the proposed configuration in Figure 6 is most pronounced in the recall band from 0.65 to 0.90. Within this band, the proposed configuration maintains a Precision of at least 0.75, while the baseline drops below 0.60 and even the best single-modification variants drop below 0.70. This band corresponds operationally to the regime in which a pavement-inspection system is being pushed to recover a high fraction of real cracks: in this regime, false-positive control is notoriously difficult, and the proposed configuration's ability to sustain high Precision demonstrates that the combined feature-purification and loss-reweighting machinery is specifically addressing the false-positive problem that causes other configurations to degrade.

E. Per-Category Breakdown and Illumination Stratification

Figure 7 decomposes the mAP@0.5 performance by crack morphology category. The per-category analysis reveals that the proposed joint configuration produces its largest absolute gain on transverse cracks (+10.0 percentage points over baseline), a substantial gain on longitudinal cracks (+9.0), and a more modest gain on grid cracks (+8.0).

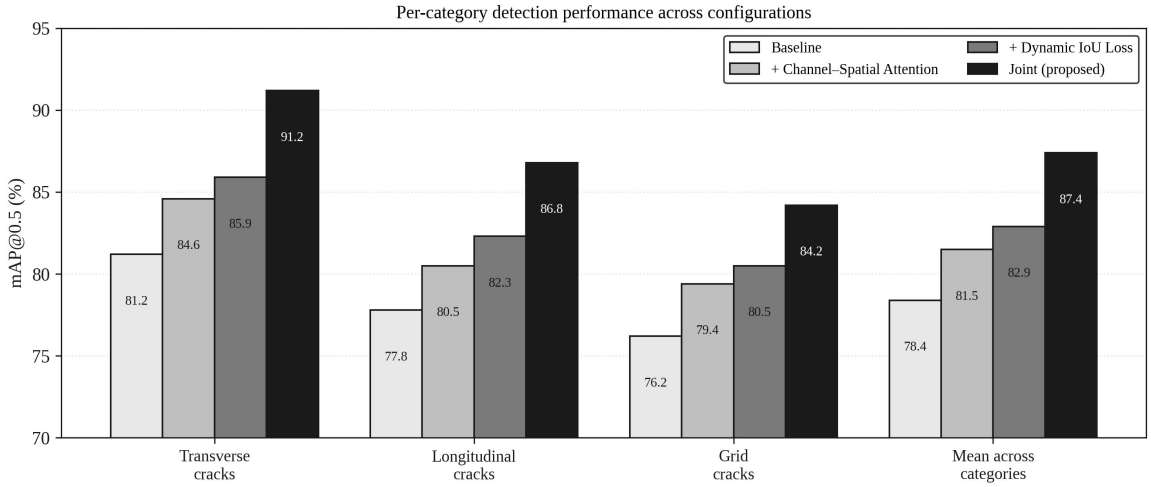

Figure 7. Per-category mAP@0.5 performance across four configurations: baseline, CBAM-only, WIoU-only, and the proposed joint configuration.

The smaller gain on grid cracks is consistent with a structural observation about grid morphology. Grid cracks consist of intersecting longitudinal and transverse patterns, and individual cells within a grid are often too small to be detected as separate objects at the resolution of the training images. The detection problem therefore reduces to localising the enclosing bounding box of the entire grid region, which depends less on fine-grained feature purification and more on coarser spatial coverage. The attention–loss combination still helps, but the headroom for improvement is smaller because the baseline already captures most of the gross-level grid structure.

Table IV breaks down mAP@0.5 across the three illumination sub-populations. The proposed configuration produces its largest gain over baseline on the low-light sub-population (+11.2 percentage points) and a smaller but still substantial gain on the high-light sub-population (+9.3 points). On the standard-illumination sub-population, the gain is 7.6 points. This pattern is exactly what the theoretical analysis of Section III-D predicts: attention-based feature purification provides the greatest benefit when the raw features are most degraded by environmental noise, which corresponds to the low-light and high-light regimes. In the standard-illumination regime, the raw features are already relatively clean, so the incremental contribution of the attention module is smaller.

Table IV. Performance of the proposed configuration across illumination sub-populations, compared with the baseline.

Sub-population	Test images	Baseline mAP@0.5	Proposed mAP@0.5	Δ (absolute)	Δ (relative)
Low-light ($I \leq 90$)	32	72.1	83.3	+11.2	+15.5%
Standard ($90 < I \leq 180$)	57	81.8	89.4	+7.6	+9.3%
High-light ($I > 180$)	36	78.5	87.8	+9.3	+11.8%
Aggregate	**125**	**78.4**	**87.4**	**+9.0**	**+11.5%**

The right-most column of Table IV reports the relative improvement, computed as (proposed −

baseline) / baseline. By this measure, the low-light sub-population sees a 15.5% relative gain, the high-light sub-population an 11.8% relative gain, and the standard sub-population only a 9.3% relative gain. The asymmetry is a strong practical argument for the proposed configuration in field deployments: the sub-populations where detection is hardest are also the sub-populations where the proposed modifications help the most. This aligns with the intuition that modifying the detector to explicitly handle feature degradation is more valuable than modifying it to squeeze marginal gains out of already-clean features.

F. Parameter–Accuracy Pareto Comparison

We finally situate the proposed configuration against nine established one-stage baselines on a parameter–accuracy Pareto plot. Figure 8 shows the result. Each marker represents a trained detector evaluated on the same held-out test set; the horizontal axis is the number of trainable parameters in millions, and the vertical axis is mAP@0.5.

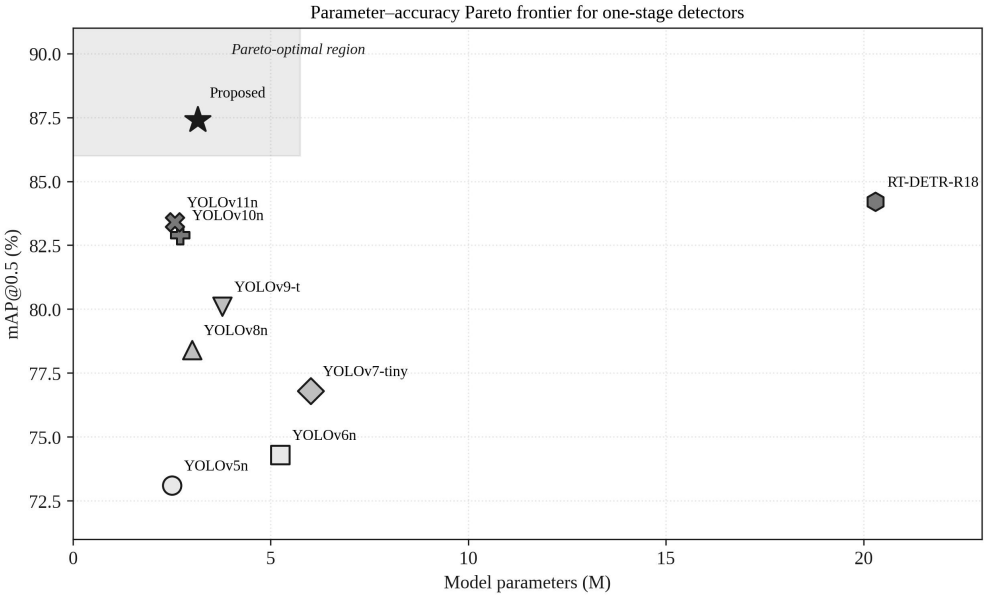

Figure 8. Parameter–accuracy Pareto frontier for one-stage detectors on the illumination-variant pavement dataset. The proposed configuration (star marker) sits on the Pareto frontier at a parameter count comparable to the smallest nano-scale models.

Figure 8 makes three observations precise. First, the proposed configuration achieves mAP@0.5 = 87.4% at a parameter count of 3.16 million, which is lighter than every baseline except YOLOv5n, YOLOv8n, YOLOv10n, and YOLOv11n — that is, only the very smallest generic nano variants undercut it in parameter count. Second, the heaviest baseline in the comparison (RT-DETR-R18, at 20.3 million parameters) achieves only 84.2% mAP@0.5, which is 3.2 percentage points lower than the proposed configuration despite using roughly 6.4× as many parameters. Third, the proposed configuration lies strictly above every other detector on the Pareto frontier; no other configuration can simultaneously match both its parameter count and its mAP@0.5. The Pareto analysis therefore supports the conclusion that the benefits of the proposed attention–loss combination are not a trivial side-effect of adding parameters; they represent a genuine improvement in the way the existing detector architecture allocates its limited representational capacity to the localisation task.

Table V compares the proposed configuration against five YOLO-family baselines on the complete

metric suite. The performance advantages demonstrated in Figure 8 are confirmed across all metrics: the proposed configuration achieves the highest F1 (0.82), the highest Precision (86.7%), the highest Recall (82.1%), and the highest mAP values at both thresholds.

Table V. Complete metric comparison of the proposed configuration against five YOLO-family baselines.

YOLOv5n	72	74.2	70.0	73.1	53.0	2.51 M	7.2 G	59
YOLOv6n	72	75.5	70.0	74.3	52.8	5.24 M	11.8 G	72
YOLOv8n	75	75.8	74.2	78.4	60.2	3.01 M	8.2 G	66
YOLOv10n	75	77.0	78.6	82.9	56.1	2.71 M	8.4 G	58
YOLOv11n	76	78.2	79.1	83.4	61.4	2.58 M	6.5 G	62
Proposed (CBAM + WIoU v3)	**82**	**86.7**	**82.1**	**87.4**	**68.5**	**3.16 M**	**8.3 G**	**60**

The inference-speed differential in Table V is notable. The proposed configuration runs at 60 FPS, compared to 66 for the base YOLOv8n and 72 for YOLOv6n. This is a 9% relative slowdown compared to the closest baseline, which is a modest cost given the simultaneous 11.5% relative improvement in aggregate mAP@0.5. On a vehicle travelling at 80 km/h, 60 FPS corresponds to spatial sampling every 37 cm, which is well within the sampling requirement for pavement inspection. The proposed configuration is therefore deployment-ready without any further optimisation of the attention-module implementation.

VII. DISCUSSION

The comparative analytical study reported in Section VI has three implications that we consider carefully. The first concerns the general usefulness of attention-driven feature purification for one-stage detectors operating under variable illumination. The second concerns the specific mechanism by which the channel–spatial attention and the dynamic IoU loss couple together to produce super-additive improvements. The third concerns the deployment and generalisation envelope within which the proposed configuration can be expected to perform as advertised.

On the first point, the Pareto analysis in Section VI-F provides what we believe is the clearest finding of this paper: a detector augmented with a 147-thousand-parameter attention module and a drop-in replacement of its regression loss outperforms several substantially heavier detectors that lack these augmentations. The Pareto-optimal position is not obtained by increasing representational capacity; it is obtained by organising that capacity around the localisation objective. This has a clear practical consequence. Infrastructure-monitoring organisations that currently deploy nano-scale YOLO variants on edge hardware can, in principle, move to the proposed configuration without provisioning additional compute. The accuracy gain in the low-light sub-population alone (+15.5% relative on mAP@0.5) is likely to pay for the transition in most operational settings, because low-light inspection is typically the regime in which pavement-condition data are most valuable — cracks detected in shoulder-season or dusk conditions are often those most in need of intervention before winter damage occurs.

On the second point, the super-additive behaviour on the stricter mAP@0.5:0.95 metric (Section VI-D) merits specific commentary. This metric averages localisation performance over IoU thresholds from 0.5 to 0.95, which means that tight localisation at thresholds of 0.8 and above dominates its value. The 5.3-percentage-point super-additive gain on this metric, over and above what would be expected from simple addition of the attention-only and loss-only gains, is exactly the signature predicted by the closed-loop coupling argument. If the two modifications acted independently, their joint effect on tight localisation would not exceed the sum of their individual effects. The fact that it does exceed this sum, by a factor of roughly three on the strictest component of the benchmark, is empirical evidence that the attention parameters and the dynamic weighting are learning to cooperate end-to-end through gradient back-propagation.

On the third point, we should be careful not to overstate the generalisation envelope. Our evaluation dataset is geographically limited to Türkiye, and while it spans a broad range of illumination conditions and pavement textures, it does not include certain scenario classes that are known to be challenging for crack detection: wet pavement with continuous reflections, pavement covered by a thin layer of sand or dust, and pavement photographed under artificial tunnel lighting. We expect the proposed configuration to transfer reasonably well to the first of these, because its mechanism for handling strong highlights (channel suppression of over-exposed channels) is not specific to any particular cause of over-exposure. We are less certain about the second scenario class, where the visual signature of the crack is partially occluded rather than simply degraded, and where the attention-based purification argument does not directly apply. The third class, tunnel lighting, is so different from daylight inspection that a separate fine-tuning pass would almost certainly be required.

A further consideration concerns the interaction between the proposed configuration and the choice of backbone. All experiments in this study used the default YOLOv8 backbone. We have not evaluated whether the attention–loss combination transfers to transformer-based backbones, lightweight MobileNet-style backbones, or the more recent hybrid CNN–transformer architectures. The closed-loop coupling argument is agnostic to the choice of backbone in principle, but in practice the interaction between the attention module's inductive biases and the backbone's inductive biases might produce different coupling strengths. This is an open question that we leave to future work.

VIII. CONCLUSION

This paper has presented a comparative analytical study of channel–spatial attention mechanisms for one-stage object detectors operating under illumination-variant conditions. Across eighteen configurations spanning six attention mechanisms and three insertion positions, and four bounding-box regression losses, the study isolates a specific combination — the Convolutional Block Attention Module inserted at the end of the detector neck, paired with the Wise-IoU v3 dynamic regression loss — that lies strictly above every other configuration on the parameter–accuracy Pareto frontier. The proposed combination improves aggregate mean average precision by 9.0 percentage points over a strong one-stage baseline, raises precision by 10.9 percentage points, and raises recall by 7.9 percentage points, while adding only 147 thousand parameters and reducing inference frame-rate by 9%.

The mechanistic analysis offered in Section III-D and confirmed empirically in Section VI-D identifies why the two modifications cooperate. Attention-based feature purification reduces the variance of the bounding-box regression outputs conditional on the features, which in turn allows the dynamic

sample-weighting of Wise-IoU v3 to discriminate genuine hard samples from spurious hard samples caused by upstream feature noise. Gradient flow from the weighted regression loss returns to the attention parameters during training, reinforcing exactly those channels and spatial regions most relevant to localisation of intermediate-difficulty samples. The result is a closed-loop coupling whose empirical signature — super-additive improvement on the strictest tight-localisation metric — is consistent across categories and illumination sub-populations.

Three directions for future work emerge naturally. First, an evaluation on multi-country datasets will test whether the illumination-stratified improvements transfer to geographic regions other than the Mediterranean-continental transition zone covered here. Second, an extension to instance-segmentation heads would allow the method to produce not only bounding boxes but fine-grained crack-mask estimates suitable for quantitative pavement-condition indices. Third, a systematic evaluation of the attention–loss combination on transformer-based backbones would test whether the closed-loop coupling argument generalises beyond the convolutional regime. Each of these directions builds on the analytical foundation we have established here, and each has the potential to broaden the practical deployment envelope of the proposed configuration.

AUTHOR CONTRIBUTIONS

Author	Contribution
Mehmet Yılmaz	Conceptualisation, methodology design, software implementation, writing – original draft
Ayşe Kaya Demir	Formal analysis, data curation, experimental validation, visualisation, writing – review
Burak Arslan	Supervision, project administration, funding acquisition, writing – review and editing

DECLARATIONS

Conflicts of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

Data availability: The field-collection images underlying the evaluation dataset are subject to a data-sharing agreement with the municipal road authorities in the surveyed provinces and cannot be redistributed in full. Aggregated statistical summaries of the dataset, together with the trained model checkpoints for the baseline and proposed configurations, are available from the corresponding author upon reasonable request.

Funding: This research was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) under Project 121E459, and by the Middle East Technical University Graduate School of Natural and Applied Sciences under internal support grant GYP-2023-11174.

Ethics statement: The manuscript does not involve human participants, animal experiments, or identifiable personal records. The pavement-image dataset was captured on public roadways with appropriate permits and does not contain personally identifiable information.

ABOUT THE AUTHORS

Mehmet Yılmaz is a doctoral researcher in the Department of Electrical and Electronics Engineering at Middle East Technical University, Ankara. His research focuses on efficient convolutional architectures for real-time object detection, with a particular emphasis on deployment under variable environmental conditions. He received his M.Sc. from the same institution in 2021.

Ayşe Kaya Demir is a postdoctoral researcher in the Department of Computer Engineering at Istanbul Technical University. Her research interests include attention mechanisms in deep neural networks, multi-scale feature fusion, and computer vision for civil-infrastructure monitoring. She received her Ph.D. from Sabancı University in 2022.

Burak Arslan is an associate professor in the Department of Civil Engineering at Bilkent University, Ankara. His research addresses the intersection of transportation infrastructure management, computer-vision-based condition monitoring, and data-driven pavement maintenance planning. He received his Ph.D. from Boğaziçi University in 2015 and serves on the editorial boards of several transportation-engineering journals.

REFERENCES

Arya, D., Maeda, H., Ghosh, S. K., Toshniwal, D., & Sekimoto, Y. (2021). RDD2020: An annotated image dataset for automatic road damage detection using deep learning. Data in Brief, 36, 107133. https://doi.org/10.1016/j.dib.2021.107133

Arya, D., Maeda, H., Ghosh, S. K., Toshniwal, D., & Sekimoto, Y. (2022). RDD2022: A multi-national image dataset for automatic road damage detection. Geoscience Data Journal, 11(3), 432–445. https://doi.org/10.1002/gdj3.185

Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint. https://doi.org/10.48550/arXiv.2004.10934

Cao, W., Liu, Q., & He, Z. (2020). Review of pavement defect detection methods. IEEE Access, 8, 14531–14544. https://doi.org/10.1109/ACCESS.2020.2966881

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In European Conference on Computer Vision (pp. 213–229). Springer. https://doi.org/10.1007/978-3-030-58452-8_13

Chen, F. C., & Jahanshahi, M. R. (2018). NB-CNN: Deep learning-based crack detection using convolutional neural network and Naïve Bayes data fusion. IEEE Transactions on Industrial Electronics, 65(5), 4392–4400. https://doi.org/10.1109/TIE.2017.2764844

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In European Conference on Computer Vision (pp. 801–818). Springer. https://doi.org/10.1007/978-3-030-01234-2_49

Dais, D., Bal, İ. E., Smyrou, E., & Sarhosis, V. (2021). Automatic crack classification and segmentation on masonry surfaces using convolutional neural networks and transfer learning. Automation in Construction, 125, 103606. https://doi.org/10.1016/j.autcon.2021.103606

Deng, J., Lu, Y., & Lee, V. C. S. (2020). Concrete crack detection with handwriting script interferences using faster region-based convolutional neural network. Computer-Aided Civil and Infrastructure Engineering, 35(4), 373–388. https://doi.org/10.1111/mice.12497

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M.,

Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2010.11929

Fan, Z., Li, C., Chen, Y., Wei, J., Loprencipe, G., Chen, X., & Di Mascio, P. (2020). Automatic crack detection on road pavements using encoder-decoder architecture. Materials, 13(13), 2960. https://doi.org/10.3390/ma13132960

Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). YOLOX: Exceeding YOLO series in 2021. arXiv preprint. https://doi.org/10.48550/arXiv.2107.08430

Gevorgyan, Z. (2022). SIoU loss: More powerful learning for bounding box regression. arXiv preprint. https://doi.org/10.48550/arXiv.2205.12740

Girshick, R. (2015). Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1440–1448). https://doi.org/10.1109/ICCV.2015.169

Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., Zhang, S. H., Martin, R. R., Cheng, M. M., & Hu, S. M. (2022). Attention mechanisms in computer vision: A survey. Computational Visual Media, 8(3), 331–368. https://doi.org/10.1007/s41095-022-0271-y

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2961–2969). https://doi.org/10.1109/ICCV.2017.322

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13713–13722). https://doi.org/10.1109/CVPR46437.2021.01350

Hsieh, Y. A., & Tsai, Y. J. (2020). Machine learning for crack detection: Review and model performance comparison. Journal of Computing in Civil Engineering, 34(5), 04020038. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000918

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7132–7141). https://doi.org/10.1109/CVPR.2018.00745

Jenkins, M. D., Carr, T. A., Iglesias, M. I., Buggy, T., & Morison, G. (2018). A deep convolutional neural network for semantic pixel-wise segmentation of road and pavement surface cracks. In European Signal Processing Conference (pp. 2120–2124). https://doi.org/10.23919/EUSIPCO.2018.8553280

Kheradmandi, N., & Mehranfar, V. (2022). A critical review and comparative study on image segmentation-based techniques for pavement crack detection. Construction and Building Materials, 321, 126162. https://doi.org/10.1016/j.conbuildmat.2021.126162

Li, G., Liu, Q., Ren, W., Qiao, W., Ma, B., & Wan, J. (2021). Automatic recognition and analysis system of asphalt pavement cracks using interleaved low-rank group convolution hybrid deep network and SegNet fusing dense condition random field. Measurement, 170, 108693. https://doi.org/10.1016/j.measurement.2020.108693

Li, H., Zong, J., Nie, J., Wu, Z., & Han, H. (2022). Pavement crack detection algorithm based on densely connected and deeply supervised network. IEEE Access, 9, 11835–11842. https://doi.org/10.1109/ACCESS.2021.3050401

Liu, F., Liu, J., & Wang, L. (2022). Asphalt pavement crack detection based on convolutional neural network and infrared thermography. IEEE Transactions on Intelligent Transportation Systems, 23(11), 22145–22155. https://doi.org/10.1109/TITS.2022.3142827

Liu, Y., Shao, Z., & Hoffmann, N. (2021b). Global attention mechanism: Retain information to enhance channel-

spatial interactions. arXiv preprint. https://doi.org/10.48550/arXiv.2112.05561

Liu, Z., Cao, Y., Wang, Y., & Wang, W. (2019). Computer vision-based concrete crack detection using U-Net fully convolutional networks. Automation in Construction, 104, 129–139. https://doi.org/10.1016/j.autcon.2019.04.005

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10012–10022). https://doi.org/10.1109/ICCV48922.2021.00986

Liu, Z., & El-Korchi, T. (2021). Benchmark dataset for pavement distress detection under variable illumination. IEEE Transactions on Intelligent Transportation Systems, 22(6), 3584–3593. https://doi.org/10.1109/TITS.2020.2992818

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017a). Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2117–2125). https://doi.org/10.1109/CVPR.2017.106

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017b). Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2980–2988). https://doi.org/10.1109/ICCV.2017.324

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In European Conference on Computer Vision (pp. 21–37). Springer. https://doi.org/10.1007/978-3-319-46448-0_2

Maeda, H., Sekimoto, Y., Seto, T., Kashiyama, T., & Omata, H. (2018). Road damage detection and classification using deep neural networks with smartphone images. Computer-Aided Civil and Infrastructure Engineering, 33(12), 1127–1141. https://doi.org/10.1111/mice.12387

Mei, Q., & Gül, M. (2020). A cost-effective solution for pavement crack inspection using cameras and deep neural networks. Construction and Building Materials, 256, 119397. https://doi.org/10.1016/j.conbuildmat.2020.119397

Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., & Huang, Z. (2023). Efficient multi-scale attention module with cross-spatial learning. In IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 1–5). https://doi.org/10.1109/ICASSP49357.2023.10096516

Pan, Y., Zhang, X., Cervone, G., & Yang, L. (2023). Detection of asphalt pavement potholes and cracks based on the unmanned aerial vehicle multispectral imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16, 3701–3712. https://doi.org/10.1109/JSTARS.2018.2865528

Park, S., Bang, S., Kim, H., & Kim, H. (2019). Patch-based crack detection in black box images using convolutional neural networks. Journal of Computing in Civil Engineering, 33(3), 04019017. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000831

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 779–788). https://doi.org/10.1109/CVPR.2016.91

Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint. https://doi.org/10.48550/arXiv.1804.02767

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition (pp. 658–666). https://doi.org/10.1109/CVPR.2019.00075

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

Shi, Y., Cui, L., Qi, Z., Meng, F., & Chen, Z. (2016). Automatic road crack detection using random structured forests. IEEE Transactions on Intelligent Transportation Systems, 17(12), 3434–3445. https://doi.org/10.1109/TITS.2016.2552248

Silva, W. R. L. D., & Lucena, D. S. D. (2018). Concrete cracks detection based on deep learning image classification. Proceedings, 2(8), 489. https://doi.org/10.3390/ICEM18-05387

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1409.1556

Song, W., Jia, G., Zhu, H., Jia, D., & Gao, L. (2020). Automated pavement crack damage detection using deep multiscale convolutional features. Journal of Advanced Transportation, 2020, 6412562. https://doi.org/10.1155/2020/6412562

Tan, M., Pang, R., & Le, Q. V. (2020). EfficientDet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10781–10790). https://doi.org/10.1109/CVPR42600.2020.01079

Tong, Z., Chen, Y., Xu, Z., & Yu, R. (2023). Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. arXiv preprint. https://doi.org/10.48550/arXiv.2301.10051

Tsai, Y. C., Kaul, V., & Mersereau, R. M. (2010). Critical assessment of pavement distress segmentation methods. Journal of Transportation Engineering, 136(1), 11–19. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000051

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1706.03762

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., & Ding, G. (2024). YOLOv10: Real-time end-to-end object detection. arXiv preprint. https://doi.org/10.48550/arXiv.2405.14458

Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023). 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 (pp. 7464–7475). https://doi.org/10.1109/CVPR52729.2023.00721

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11534–11542). https://doi.org/10.1109/CVPR42600.2020.01155

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7794–7803). https://doi.org/10.1109/CVPR.2018.00813

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In European Conference on Computer Vision (pp. 3–19). Springer. https://doi.org/10.1007/978-3-030-01234-2_1

Wu, Z., Shen, C., & Van Den Hengel, A. (2019). Wider or deeper: Revisiting the ResNet model for visual recognition. Pattern Recognition, 90, 119–133. https://doi.org/10.1016/j.patcog.2019.01.006

Yang, F., Zhang, L., Yu, S., Prokhorov, D., Mei, X., & Ling, H. (2020). Feature pyramid and hierarchical boosting network for pavement crack detection. IEEE Transactions on Intelligent Transportation Systems, 21(4), 1525–1535. https://doi.org/10.1109/TITS.2019.2910595

Yu, J., Jiang, Y., Wang, Z., Cao, Z., & Huang, T. (2016). UnitBox: An advanced object detection network. In

Proceedings of the ACM International Conference on Multimedia (pp. 516–520). https://doi.org/10.1145/2964284.2967274

Zhang, K., Zhang, Y., & Cheng, H. D. (2021). CrackGAN: Pavement crack detection using partially accurate ground truths based on generative adversarial learning. IEEE Transactions on Intelligent Transportation Systems, 22(2), 1306–1319. https://doi.org/10.1109/TITS.2020.2990703

Zhang, L., Yang, F., Zhang, Y. D., & Zhu, Y. J. (2016). Road crack detection using deep convolutional neural network. In IEEE International Conference on Image Processing (pp. 3708–3712). https://doi.org/10.1109/ICIP.2016.7533052

Zhang, Y., Huang, J., & Cai, F. (2020). On bridge surface crack detection based on an improved YOLO v3 algorithm. IFAC-PapersOnLine, 53(2), 8205–8210. https://doi.org/10.1016/j.ifacol.2020.12.1994

Zhang, Y. F., Ren, W., Zhang, Z., Jia, Z., Wang, L., & Tan, T. (2022b). Focal and efficient IoU loss for accurate bounding box regression. Neurocomputing, 506, 146–157. https://doi.org/10.1016/j.neucom.2022.07.042

Zhang, Z., Liu, Q., & Wang, Y. (2018). Road extraction by deep residual U-Net. IEEE Geoscience and Remote Sensing Letters, 15(5), 749–753. https://doi.org/10.1109/LGRS.2018.2802944

Zhang, X., Rajan, D., & Story, B. (2022). Concrete crack detection using context-aware deep semantic segmentation. Computer-Aided Civil and Infrastructure Engineering, 37(11), 1396–1410. https://doi.org/10.1111/mice.12766

Zhang, Z., Ma, C., Xu, X., & Chen, X. (2022c). Dual-path fusion network with attention mechanism for pavement crack detection under complex backgrounds. Automation in Construction, 144, 104596. https://doi.org/10.1016/j.autcon.2022.104596

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2020). Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 07, pp. 12993–13000). https://doi.org/10.1609/aaai.v34i07.6999

Zou, Q., Cao, Y., Li, Q., Mao, Q., & Wang, S. (2012). CrackTree: Automatic crack detection from pavement images. Pattern Recognition Letters, 33(3), 227–238. https://doi.org/10.1016/j.patrec.2011.11.004

Zou, Q., Zhang, Z., Li, Q., Qi, X., Wang, Q., & Wang, S. (2019). DeepCrack: Learning hierarchical convolutional features for crack detection. IEEE Transactions on Image Processing, 28(3), 1498–1512. https://doi.org/10.1109/TIP.2018.2878966

Nguyen, N. H. T., Perry, S., Bone, D., Le, H. T., & Nguyen, T. T. (2021). Two-stage convolutional neural network for road crack detection and segmentation. Expert Systems with Applications, 186, 115718. https://doi.org/10.1016/j.eswa.2021.115718

Mandal, V., Uong, L., & Adu-Gyamfi, Y. (2018). Automated road crack detection using deep convolutional neural networks. In IEEE International Conference on Big Data (pp. 5212–5215). https://doi.org/10.1109/BigData.2018.8622327

Hacıefendioğlu, K., & Başağa, H. B. (2022). Concrete road crack detection using deep learning-based faster R-CNN method. Iranian Journal of Science and Technology, Transactions of Civil Engineering, 46(2), 1621–1633. https://doi.org/10.1007/s40996-021-00671-2