

HCP-RDD: Hierarchical Caption-Prefix Framework for Road Damage Detection Based on YOLO11s

Shenghua Liu[✉], Tianci Ao^{*}, Peng Shao[✉]
School of Mathematics and Information Science,
Chaoyang Normal University, Chaoyang 122000, China

Abstract—Road damage detection identifies pavement distresses such as cracks using vision technologies to support road safety. Most existing methods rely solely on visual features and provide poor fine-grained recognition under illumination variation and shadows. Unconstrained noisy scene captions also degrade visual feature learning. This study proposes a Hierarchical Caption-Prefix framework for Road Damage Detection based on YOLO11s (HCP-RDD). Embedded between the YOLO11s backbone and neck, the Hierarchical Caption-Prefix Module (HCPM) contains a Cross-Attention Module (CAM) and a Dynamic Gate Module (DGM). It encodes disease-oriented captions generated by a large language model (LLM) using Bidirectional Encoder Representations from Transformers (BERT). Treating text semantics as a prefix, it concatenates them with scale-wise global visual descriptors to enable DGM to prioritize textual priors. CAM produces image-text interaction increments, which are adaptively weighted by scale-specific dynamic gates in DGM. Residual incremental reprojection remaps these increments to the original resolution to enhance hierarchical multi-scale features while retaining raw visual evidence. Enhanced features are fused in the neck feature pyramid, and the head performs bounding-box regression and category prediction. A bypass mechanism reverts the module to vanilla YOLO11s when valid captions are unavailable. Comparative and ablation experiments are performed on the China subset of RDD2022, with generalization tests on the Japan subset. Compared with the current state of the art, our model improves mAP@0.50 by 1.62 percentage points and achieves the highest precision and recall. The results validate its effectiveness and universality, offering insight into controllable fusion of textual scene priors and visual features for road damage detection.

Keywords—Road damage detection; YOLO11s; caption conditioning; cross-attention; multi-scale fusion; ablation study

I. INTRODUCTION

Road damage detection is a fundamental and critical task for intelligent transportation and road network operation and maintenance. It is important for automated pavement distress inspection and road-safety monitoring [1], [2], [3], [4]. With the rapid development of deep-learning-based object detection, numerous convolutional neural network (CNN) algorithms have been applied to road damage detection [5], [6], [7]. Such methods rely on visual image features to automatically identify pavement distresses including cracks and potholes, and perform damage localization and classification, which promotes the transition from manual inspection to automatic inspection. Nevertheless, under complex conditions involving illumination variation, shadow interference and pavement repair traces, existing models are vulnerable to noise and

suffer from missed and false detections. The China subset of RDD2022 contains real-world samples with complicated illumination, shadow interference and multiple co-occurring distresses, which is suitable for verifying model performance in complex cross-regional scenarios. Early road damage detection studies mainly adopted hand-crafted texture and edge features for damage identification [8], [9], [10], [11]. Later, supervised CNN-based object detection became mainstream. The complete image is fed into the network to extract visual features for pavement-damage recognition [5], [6], [7]. Previous research has demonstrated that multi-scale feature fusion is critical for detecting irregular cracks and tiny potholes [8], [9]. To reduce deployment overhead, lightweight YOLO-series models have been widely adopted for road damage detection.

Recently, multimodal large-scale models have developed rapidly. Some works introduce image captions into detection tasks, and leverage domain prior knowledge carried by text to assist pavement-distress recognition [10], [11]. Although generative scene captions can provide high-level semantic priors related to road damage, these texts inevitably suffer from missing details and description bias. If textual information is directly integrated into feature fusion without constraints, flawed text semantics will disturb visual representation learning and bring negative interference.

As shown in Fig. 1(a), when only visual image information is used, the model lacks high-level semantic priors of pavement distresses and tends to produce missed and false detections under complex interference. As illustrated in Fig. 1(b), biased or erroneous captions together with unconstrained fusion disrupt feature learning and further degrade detection performance. By contrast, high-quality captions can supply reliable domain priors and help the model make more accurate judgments, as presented in Fig. 1(c).

Therefore, a fusion mechanism is urgently required. It should introduce textual prior knowledge in a controllable manner while preserving raw visual evidence, so as to reduce the model's dependence on annotated samples and improve its robustness under complex unseen scenarios. Scene captions can deliver high-level semantic information including damage categories, morphologies and illumination conditions. Accordingly, this study takes scene captions as caption-prefixes to guide the learning of image visual features.

Motivated by the above observations, this study proposes a Hierarchical Caption-Prefix framework for Road Damage Detection (HCP-RDD) based on YOLO11s. The hierarchical caption-prefix module is embedded between the backbone and

^{*}Corresponding author.

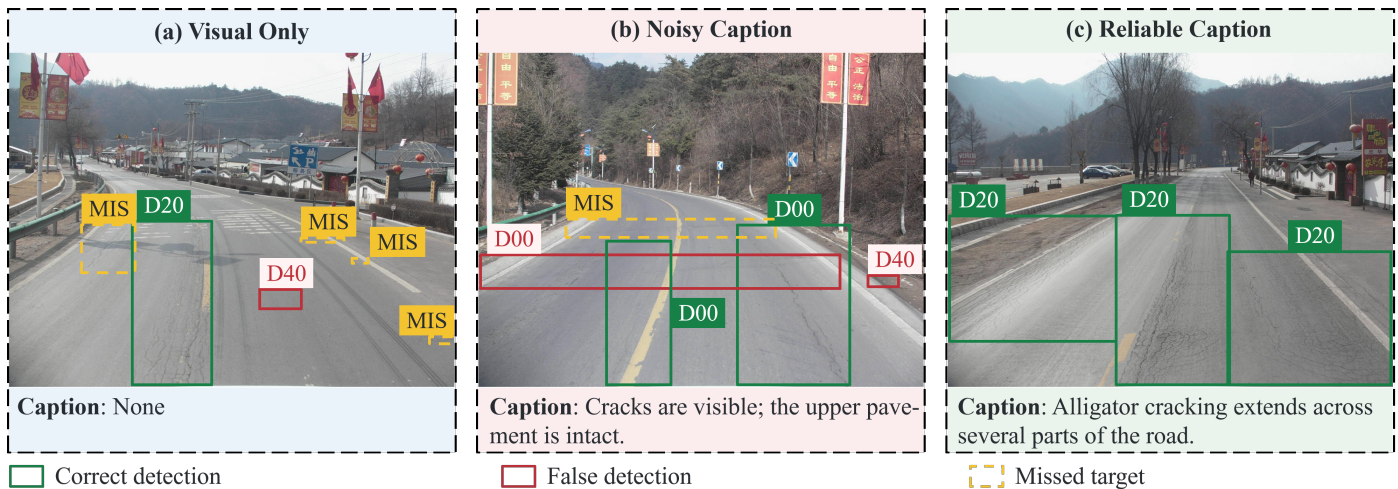


Fig. 1. Illustration of caption quality and road damage detection: (a) visual-only detection without a caption; (b) detection with a noisy caption; and (c) detection with a reliable caption. Green solid boxes indicate correct detections, red solid boxes indicate false detections, and yellow dashed boxes (MIS) indicate missed targets. D00, D20, and D40 denote longitudinal cracks, alligator cracks, and potholes, respectively.

neck of YOLO11s. Disease-oriented scene captions generated by large language models (LLMs) are encoded using Bidirectional Encoder Representations from Transformers (BERT) to obtain contextual text features. Text semantics are treated as prefixes and concatenated with scale-wise global visual descriptors, enabling the gate network to preferentially capture disease-related textual priors. Cross-attention is adopted for image-text feature interaction, and scale-specific dynamic gates adaptively adjust the contribution strength of textual information at each layer. Combined with the residual incremental reprojection operator, image-text interaction increments in the compressed space are remapped back to original feature resolution to preserve the original visual features. Equipped with a bypass mechanism, this module achieves plug-and-play deployment. When valid scene captions are available, scale-specific dynamic gates and residual incremental reprojection realize controllable guidance of text semantics. In the absence of valid text input, raw visual features output by the backbone are directly used for detection. Comparative and ablation experiments are carried out on the China subset of RDD2022, and the Japan subset is adopted for cross-regional generalization testing. Experimental results demonstrate that HCP-RDD effectively boosts detection accuracy and yields better cross-regional generalization performance than the baseline model.

In summary, the main contributions of this study are as follows:

- A hierarchical caption-prefix architecture HCP-RDD is constructed. Disease-related text semantics are concatenated as prefixes with scale-wise global visual descriptors, and cross-attention is adopted to hierarchically enhance multi-scale visual features.
- Scale-specific dynamic gates and residual incremental reprojection operators are proposed. Combined with a bypass mechanism, they can controllably modulate the contribution of textual priors and suppress negative interference brought by noisy captions.
- Comparative and ablation experiments are conducted

on the China subset of RDD2022, and the Japan subset is used to verify cross-regional generalization capability. The experimental results achieve state-of-the-art (SOTA) performance, which further validates the model's accuracy and generalization performance.

II. RELATED WORK

A. Object Detection

Object detection is an important research branch in computer vision, whose main task is to locate and recognize objects belonging to predefined categories within images [12], [13], [14]. A classical object detection framework mainly consists of three components: the backbone network, the neck network, and the detection head [15], [16], [17]. The anchor mechanism brings a critical technical breakthrough that greatly improves localization and classification performance, making anchor-based models the mainstream paradigm, which can be divided into one-stage and two-stage detection algorithms according to inference pipelines [18], [19], [20], [21], [22]. In anchor-based detection pipelines, non-maximum suppression (NMS) is an indispensable post-processing module and the most widely used bounding-box filtering algorithm [23], [24], [14]. It retains the highest-confidence bounding box within each local region and suppresses redundant candidates. Nevertheless, directly applying general object-detection techniques to pavement damage recognition has obvious limitations. Pavement diseases share highly similar visual features and produce bounding boxes with large intersection over union (IoU) values; conventional NMS cannot effectively distinguish overlapping damage targets and tends to retain redundant boxes, degrading pavement-damage detection accuracy. Accordingly, improved NMS algorithms for dense overlapping scenarios have become an important research direction [25], [26].

To eliminate inherent constraints imposed by anchors and NMS, end-to-end anchor-free detection models based on transformers have developed rapidly, and numerous schemes have been proposed to achieve a better accuracy-speed trade-off [13], [27], [28], [29], [30]. For practical road damage

detection applications, however, constrained by the computing power of edge devices, anchor-based CNN detectors are still widely adopted, and NMS remains an essential post-processing module. As a typical sub-task of object detection, pavement damage detection faces practical challenges including complex backgrounds, severe illumination variations and shadow disturbances. General object-detection algorithms usually cannot achieve satisfactory performance when directly applied to this task, which raises stricter requirements for the model's fine-grained recognition capability and detection accuracy and motivates substantial research targeting pavement distresses.

B. Road Damage Detection

Early studies mainly relied on traditional digital image-processing techniques such as edge detection and threshold segmentation for pavement defect identification [8], [31], [32], [33]. For example, Kirschke et al. [34] proposed a histogram-based method for automatic pavement crack detection; Ying and Salari [35] adopted wavelet transform to realize the detection and classification of pavement cracks. Although such methods are computationally efficient, their detection accuracy is limited against complex backgrounds with illumination variation and shadow interference.

With the continuous development of machine-learning technologies, feature-engineering-based approaches including support vector machines (SVMs) [8] and random forests [36] have been gradually applied to this field. For instance, Oliveira and Correia trained a crack-detection classifier by extracting texture and other image features [37]; Ali et al. [38] optimized the detection model with the AdaBoost algorithm; Hoang et al. [39] used SVMs to classify cracks on asphalt pavements; Kyriakou et al. [40] constructed a pavement-damage detection model based on artificial neural networks; Cord et al. [41] implemented automatic road defect detection via AdaBoost-driven texture pattern recognition; Kwon et al. [42] achieved rapid detection of various surface defects by integrating random forests with VOV features. Nevertheless, machine-learning methods heavily depend on manually designed features such as textures and thresholds, which makes it difficult to adapt to the complex and variable morphological characteristics of pavement damage.

Benefiting from powerful feature-learning capabilities, deep learning has become the dominant approach in recent years, which can be roughly divided into two categories: semantic segmentation [43], [32] and object detection [12], [13], [14]. Maeda et al. [44] first constructed the relatively systematic pavement-damage dataset RDD2018 and evaluated multiple mainstream detection models on this dataset. Since distresses such as cracks and potholes usually have fine-grained structures and small spatial scales and are easily obscured by illumination and shadows, fully convolutional networks [45] and hierarchical networks [46] are widely adopted for crack-detection tasks.

For multi-class road damage detection tasks, most existing methods improve performance by modifying mature detection models. Doshi et al. [47] employed an ensemble of YOLOv4-based detectors to identify multiple types of road damage, improving detection performance by combining predictions from different trained models. Arya et al. [48] performed

preprocessing and data augmentation on images captured by mobile phones, trained RetinaNet with multiple backbone networks, and conducted cross-region generalization tests. Liu et al. [49] proposed an efficient dense attention fusion network, which achieves high-precision pavement-damage detection by enhancing multi-scale feature interactions.

Although the above-mentioned studies have achieved promising results, most existing road-damage detection models perform defect recognition using visual features alone and cannot effectively address the challenge of fine-grained defect recognition caused by illumination variations and shadow interference. Under complex real-world working conditions including illumination variations, shadow interference and pavement repair traces, these models are prone to missed detections and false positives. Moreover, such models heavily rely on large-scale annotated samples and exhibit limited generalization capability in cross-regional scenarios. Accordingly, this study proposes a Hierarchical Caption-Prefix framework for Road Damage Detection based on YOLO11s (HCP-RDD), which introduces scene descriptions as textual priors into visual-feature learning in a controllable manner.

III. METHOD

A. Overview

The proposed Hierarchical Caption-Prefix framework for Road Damage Detection based on YOLO11s (HCP-RDD) is illustrated in Fig. 2. Existing road damage detection methods suffer from fine-grained missed and false detections under illumination variation and shadow interference. Current multimodal fusion approaches adopt unconstrained text injection, where noisy captions degrade visual representation learning. Given a road image and its LLM-generated disease-oriented caption, the image is fed into the YOLO11s backbone to generate multi-scale visual features. Meanwhile, the caption is encoded by the pre-trained BERT encoder [50] to extract high-level domain semantics covering damage categories, morphologies and illumination conditions. Embedded between the backbone and neck of YOLO11s, the Hierarchical Caption-Prefix Module (HCPM) consists of the Cross-Attention Module (CAM) and Dynamic Gate Module (DGM), using the caption prefix as a semantic prior. The CAM performs image-text semantic alignment via cross-attention to hierarchically inject textual domain priors into multi-scale visual features. The DGM adaptively weights image-text interaction increments at each scale to modulate text contribution. The residual incremental reprojection operator remaps these increments back to original feature resolution for hierarchical multi-scale feature enhancement while preserving raw visual evidence.

During training, the standard YOLO11 detection objective is computed from detection-head predictions, including decoded boxes and predicted distance distributions, and assigned ground-truth targets. The DGM adjusts caption-prefix weights to suppress negative interference from noisy LLM-generated captions. At inference, features enhanced by HCPM are passed to the vanilla YOLO11s neck and head, and final detection results are obtained after decoding and non-maximum suppression. When valid captions are unavailable, the HCPM activates the bypass mechanism and the model degenerates to vanilla YOLO11s. The framework achieves controllable hierarchical

injection of caption-prefix priors, leverages semantic guidance for fine-grained damage recognition, and improves detection robustness for cross-regional complex road scenes.

B. Representation

Road images contain rich visual cues, which play a vital role in road damage detection. We adopt the lightweight YOLO11s detector [51], as illustrated in Fig. 2. The input road image is denoted as $\mathbf{I} \in \mathbb{R}^{3 \times 640 \times 640}$ and fed into the YOLO11s backbone. The backbone consists of stacked Conv, C3K2, SPPF, and C2PSA modules. Conv implements basic convolution for preliminary feature extraction. C3K2 is the core feature-extraction block for multi-level spatial feature learning. SPPF achieves multi-scale receptive field expansion via fast spatial pooling. C2PSA introduces partial self-attention to capture long-range spatial dependencies within pavement scenes. The backbone outputs three pyramid feature maps P_3 , P_4 and P_5 with different down-sampling strides [Eq. (1) and (2)]:

$$\{P_3, P_4, P_5\} = \text{backbone}(\mathbf{I}) \quad (1)$$

$$\begin{cases} P_3 \in \mathbb{R}^{C_3 \times H_3 \times W_3}, & \text{stride} = 8 \\ P_4 \in \mathbb{R}^{C_4 \times H_4 \times W_4}, & \text{stride} = 16 \\ P_5 \in \mathbb{R}^{C_5 \times H_5 \times W_5}, & \text{stride} = 32 \end{cases} \quad (2)$$

where, C_l represents the channel dimension of feature map P_l ; H_l and W_l are the height and width of the corresponding feature map, respectively. stride is the down-sampling stride relative to the input image. $l \in \{3, 4, 5\}$ corresponds to three different pyramid scales. When the HCPM is disabled, these three feature tensors P_3 , P_4 , P_5 are directly fed into the subsequent neck module for feature-pyramid fusion.

For the text modality, the raw road image $\mathbf{I}$ is fed into the large language model LLM. Relying on its powerful vision-grounded description capability, the LLM generates a detailed disease-oriented scene caption $\mathbf{C}$. Afterwards, the obtained caption $\mathbf{C}$ is fed into the BERT [50] encoder to produce caption token features $\mathbf{C}_t$, which encapsulate high-level semantics including crack categories and illumination conditions [Eq. (3) and (4)]:

$$\mathbf{C} = \text{LLM}(\mathbf{I}) \quad (3)$$

$$\mathbf{C}_t = \text{BERT}(\mathbf{C}) \quad (4)$$

where, $\mathbf{C}_t \in \mathbb{R}^{B \times L \times d_t}$. B denotes the batch size, L denotes the token sequence length, and d_t denotes the feature dimension of BERT output.

In this work, Qwen3-VL-8B-Instruct is adopted as the LLM owing to its outstanding image-aware text generation performance. Moreover, supplementary experiments regarding the generation capability of different LLMs and the influence of caption token sequence length are presented in Section IV.

C. Hierarchical Caption-Prefix Module

The Hierarchical Caption-Prefix Module (HCPM) is embedded between the backbone and neck of YOLO11s and aims to inject caption-derived semantic priors into multi-scale visual features in a controllable manner at each layer. For each visual pyramid feature P_l , where $l \in \{3, 4, 5\}$, HCPM uses the Cross-Attention Module (CAM) and Dynamic Gate Module (DGM) to achieve controllable hierarchical injection of textual semantic priors, as illustrated in Fig. 2.

Specifically, each pyramid feature P_l ($l \in \{3, 4, 5\}$) flows through two independent operator branches, visual projection $\text{M1}(\cdot)$ and average pooling $\text{M2}(\cdot)$ as marked in Fig. 2. Via the visual projection operator $\text{M1}(\cdot)$, each raw pyramid feature P_l is projected into the joint multimodal feature space and yields projected visual feature F_l [Eq. (5)]:

$$F_l = \text{M1}(P_l), \quad l \in \{3, 4, 5\} \quad (5)$$

where, the resulting F_3, F_4, F_5 are prepared for subsequent cross-modal interaction within CAM together with the caption token feature $\mathbf{C}_t$.

Meanwhile, the average pooling operator $\text{M2}(\cdot)$ is executed on each P_l to extract a layer-specific global visual descriptor V_{gl} . Apart from visual features, the caption token feature $\mathbf{C}_t$ is also fed into the same average pooling operator $\text{M2}(\cdot)$ to generate a compact global caption descriptor $\mathbf{C}_g$ [Eq. (6) and (7)]:

$$V_{gl} = \text{M2}(P_l), \quad l \in \{3, 4, 5\} \quad (6)$$

$$\mathbf{C}_g = \text{M2}(\mathbf{C}_t) \quad (7)$$

For each feature scale l , the layer-specific global visual descriptor V_{gl} is fused with the global caption descriptor $\mathbf{C}_g$ via the caption-prefix concatenation operator $\oplus$. In this operation, $\mathbf{C}_g$ is treated as the semantic prefix, placed first, and then concatenated with V_{gl} to construct the joint global descriptor CV_l [Eq. (8)]:

$$CV_l = \mathbf{C}_g \oplus V_{gl}, \quad l \in \{3, 4, 5\} \quad (8)$$

where, $\oplus$ is redefined as the caption-prefix concatenation operator. Rather than element-wise addition, this operator performs vector concatenation, where the left-hand tensor $\mathbf{C}_g$ is treated as the semantic prefix and placed at the beginning of the fused vector.

Such caption-prefix concatenation treats textual semantic information as the leading prior. It enables the gate multi-layer perceptron (MLP) to first perceive high-level pavement-crack semantics delivered by captions before processing visual features. This prefix-oriented fusion pattern strengthens the semantic guidance of text priors, and enables the MLP to establish an explicit mapping between caption semantics and visual feature response, which alleviates the risk that textual semantics are submerged by high-dimensional visual signals. The fused descriptors CV_3, CV_4, CV_5 serve as inputs to the subsequent DGM branch to generate scale-adaptive gate coefficients.

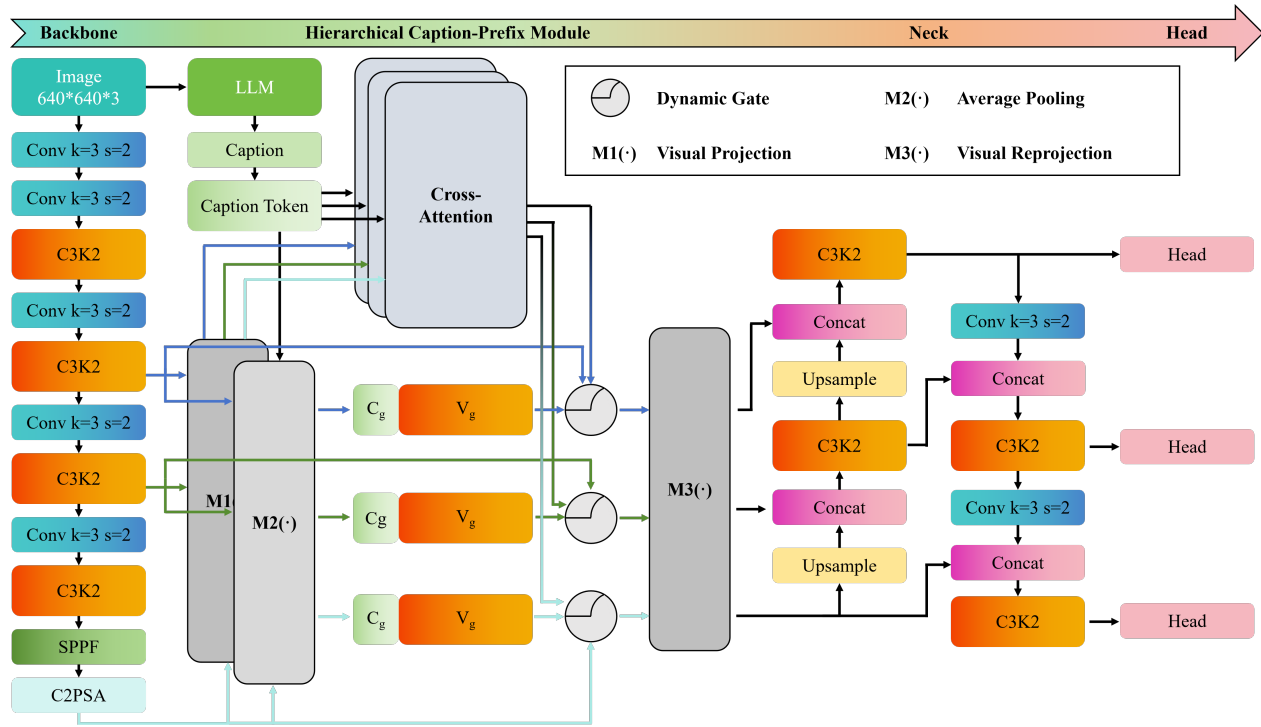


Fig. 2. The framework of HCP-RDD. The Hierarchical Caption-Prefix Module (HCPM) is embedded between the YOLO11s backbone and Neck to enhance three scale-wise visual features with disease-oriented caption prefixes. M1 and M3 denote visual projection operations, and M2 denotes average pooling.

1) *Cross-Attention Module (CAM)*: The CAM conducts cross-modal semantic alignment between the M1-projected visual feature F_l and caption token feature C_t . Note that F_l ($l \in \{3, 4, 5\}$) is already the visual feature output by visual projection operator $M1(\cdot)$ in the multimodal joint space. No additional reshaping operation is performed, and F_l is directly adopted as the query of cross-attention. We set the query $Q = F_l$, the key $K = C_t$, and the value $V = C_t$ for cross-attention computation [Eq. (9)]:

$$F_l^{\text{enh}} = \text{CrossAttn}(Q = F_l, K = C_t, V = C_t) \quad (9)$$

The cross-modal incremental term Δ_l denotes the semantic correction brought by caption semantics for the l -th feature scale, which is obtained by subtracting the original projected visual feature from the cross-attention-enhanced visual feature [Eq. (10)]:

$$\Delta_l = F_l^{\text{enh}} - F_l, \quad l \in \{3, 4, 5\} \quad (10)$$

2) *Dynamic Gate Module (DGM)*: The DGM aims to produce a scale-specific gate coefficient $G_l \in [0, 1]$ to adaptively adjust the contribution strength of caption-prefix semantic priors. Unlike naive concatenation of globally pooled features, DGM directly takes the prefix-aware joint global descriptor CV_l as input. Each CV_l is fed into a scale-specific gate MLP, followed by the sigmoid activation σ to constrain gate output within $[0, 1]$ [Eq. (11)]:

$$G_l = \sigma(\text{MLP}(CV_l)) \in \mathbb{R}^{B \times 1}, \quad l \in \{3, 4, 5\} \quad (11)$$

where, the gate coefficient G_l adaptively controls the contribution weight of cross-modal increment Δ_l for each feature scale. When the caption contains noise or semantic bias, G_l will automatically decrease to suppress negative interference from text priors.

After obtaining the scale-specific gate coefficient G_l , gate-weighted residual fusion is executed at the projected-feature level. The cross-modal increment Δ_l is weighted by G_l and added back to the original projected visual feature F_l [Eq. (12)]:

$$F_l^{\text{out}} = F_l + G_l \cdot \Delta_l \quad (12)$$

Subsequently, the enhanced projected feature F_l^{out} is transformed back to a caption-enhanced pyramid feature map via visual reproduction operator $M3(\cdot)$ shown in Fig. 2 [Eq. (13)]:

$$\tilde{P}_l = M3(F_l^{\text{out}}), \quad l \in \{3, 4, 5\} \quad (13)$$

where, $\tilde{P}_3, \tilde{P}_4, \tilde{P}_5$ denote the final caption-enhanced multi-scale pyramid features.

The above computation pipeline runs independently for P_3, P_4 and P_5 . The enhanced features $\tilde{P}_3, \tilde{P}_4, \tilde{P}_5$ are then delivered to the subsequent neck module for feature-pyramid fusion. If a valid caption is unavailable, HCPM is directly bypassed, and the original P_3, P_4, P_5 are forwarded to the neck without modification.

D. Loss Function

HCP-RDD adopts the native YOLO11s detection objective without additional multimodal or text-auxiliary loss terms. HCPM only modifies the intermediate multi-scale features supplied to the YOLO11s neck and head. Optimization uses detection-head predictions and assigned ground-truth targets.

The objective combines bounding-box regression, classification, and distribution focal loss (DFL) [Eq. (14)]:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{box}}\mathcal{L}_{\text{box}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{dff}}\mathcal{L}_{\text{dff}} \quad (14)$$

where, the coefficients are the native YOLO11s loss gains. The expression omits the common batch-size scaling applied during training.

Classification uses BCEWithLogitsLoss between category logits O_{ic} and target scores S_{ic} produced by TaskAlignedAssigner [Eq. (15)]:

$$\mathcal{L}_{\text{cls}} = \frac{1}{Z} \sum_{i,c} \text{BCEWithLogits}(O_{ic}, S_{ic}) \quad (15)$$

Here, i indexes prediction locations across the batch and scales, c indexes the four damage categories, and $Z = \max(\sum_{i,c} S_{ic}, 1)$. The assigned scores encode both category and assignment quality; background targets are zero.

For assigned positive locations $\mathcal{P}$, the box loss compares differentially decoded boxes B_i with their matched ground-truth boxes B_i^{gt} using complete intersection over union (CIoU) [Eq. (16)]:

$$\mathcal{L}_{\text{box}} = \frac{1}{Z} \sum_{i \in \mathcal{P}} w_i [1 - \text{CIoU}(B_i, B_i^{\text{gt}})] \quad (16)$$

where, $w_i = \sum_c S_{ic}$ weights each positive prediction by its assigned target score.

The head models each of the four boundary distances from a reference point as a discrete distribution with 16 bins ($\text{reg_max} = 16$). DFL supervises these distributions using the assigned box distances [Eq. (17)]:

$$\mathcal{L}_{\text{dff}} = \frac{1}{4Z} \sum_{i \in \mathcal{P}} w_i \sum_{k=1}^4 \text{DFL}(p_{ik}, d_{ik}) \quad (17)$$

Here, p_{ik} is the predicted distance distribution and d_{ik} is the target distance in feature-map units, clipped to the supported range. DFL interpolates the negative log-probabilities of the two neighboring bins around d_{ik} . Box regression and DFL share the target-score weighting and normalization. NMS is applied only during inference post-processing and does not enter the training loss.

In the training phase, the pre-trained BERT encoder is fully frozen to avoid overfitting to domain-specific road scene captions. The backbone, neck, and head of YOLO11s are initialized with official pre-trained weights. Parameters inside

HCPM, including projection operators M1, M3, the cross-attention layer in CAM and the gate MLP in DGM, are randomly initialized and optimized together with a subset of the detection-head parameters.

When caption input is missing or invalid, HCPM is directly bypassed, and the model degenerates to vanilla YOLO11s with exactly the same loss formula. For low-quality or biased LLM-generated captions, the scale-specific gate coefficient G_l produced by DGM will automatically decrease in forward propagation to suppress negative semantic interference. This suppression mechanism operates at the feature level during the forward pass and does not require any modification to the original loss objective.

IV. EXPERIMENTS

In this section, we conduct comprehensive experiments on the RDD2022-China dataset. Our overarching goal is to answer the following research questions:

- RQ1: How does our model compare with the state-of-the-art approaches in terms of quantitative metrics?
- RQ2: How much do various components of our model contribute to its overall performance?
- RQ3: How do different key factors affect the detection performance of the proposed HCP-RDD framework?
- RQ4: How does the proposed model compare to the state-of-the-art methods on visual qualitative results?

A. Datasets

RDD2022 is a public large-scale benchmark dataset for automatic road damage detection, which originates from the Crowdsensing-based Road Damage Detection Challenge (CRDDC'2022) [52]. The dataset contains 47420 road images collected from six countries including China, Japan, India, the Czech Republic, Norway and the United States, with more than 55000 annotated road-damage instances. In this study, the China subset of RDD2022 (RDD2022-China) is selected as the experimental dataset. RDD2022-China contains China_Drone and China_MotorBike, and this study uses China_MotorBike for training and evaluation. This subset features heterogeneous and complex image acquisition conditions with diverse variations in viewpoint, object scale and background environment, which can effectively verify the detection performance of road-damage detectors under complex working conditions.

The official CRDDC benchmark defines four core road-damage categories: longitudinal crack (D00), transverse crack (D10), alligator crack (D20) and pothole (D40). Only samples of the above four damage categories are adopted in our experiments, and the Repair category used in some related studies is excluded. All damage samples are annotated with bounding boxes, and the original annotation files are XML files in Pascal-VOC format. To adapt to the training of YOLO-series models, the XML annotations are converted into YOLO-format label files in the local training pipeline, while the category mapping of D00/D10/D20/D40 is fully preserved.

In this study, the RDD2022-China_MotorBike dataset is divided into training, validation and test sets at a ratio of

TABLE I. TRAINING SETTINGS OF HCP-RDD

Item	Setting
CPU	Intel Core i9-13900K
GPU	NVIDIA A40
RAM	256 GB
Operating System	Ubuntu 20.04 LTS
Programming Language	Python 3.10.11
Deep Learning Framework	PyTorch 2.1
Input / training schedule	640 × 640; 300 epochs; batch size 32
Optimizer	AdamW; trainable modules LR 2×10^{-4}
Regularization / LR schedule	Weight decay 5×10^{-4} ; cosine LR decay; 3-epoch warm-up
Text Encoder	bert-base-uncased
Caption Token Length	80
Text / reproducibility	Qwen3-VL-8B-Instruct; seeds 0, 1, 2; AMP enabled

8:1:1. All comparative models use exactly the same dataset split to ensure a consistent data protocol for model training and evaluation. The test set is kept isolated throughout the whole process and is only used for the final performance evaluation of pre-specified model configurations.

B. Implementation Details

Table I summarizes the training configurations of HCP-RDD. All experiments are conducted on a computing server equipped with an Intel Core i9-13900K CPU, an NVIDIA A40 GPU and 256 GB of RAM. The software environment consists of Ubuntu 20.04 LTS, Python 3.10.11 and PyTorch 2.1.

All models are trained for 300 epochs with an input resolution of 640×640 and a batch size of 32. The AdamW optimizer is used, with the initial learning rate of all trainable modules is set to 2×10^{-4} and the weight decay is 5×10^{-4} . A three-epoch warm-up is performed during the initial training stage, followed by cosine learning-rate decay. Automatic mixed precision (AMP) is enabled to improve training efficiency. The pretrained weights of bert-base-uncased are used for the BERT text encoder.

Captions are pre-generated offline by the LLM from raw images, without accessing ground-truth bounding boxes or category labels. Qwen3-VL-8B-Instruct is adopted as the caption generator in our final configuration, and the maximum token length of scene captions is limited to 80. In this study, mAP@0.50, mAP@0.50:0.95, F1-score, precision, recall, and GFLOPs are selected as comprehensive evaluation metrics.

C. Baselines

To provide a comprehensive comparison, we select representative two-stage, anchor-based, and anchor-free object detectors as baselines.

- Faster R-CNN [53] represents the classical two-stage detection paradigm and has been widely applied to road damage detection.
- YOLOv5s [54] is a lightweight one-stage detector widely adopted as a baseline for real-time object detection.
- SSD [55] performs object classification and localization within a unified single-stage detection framework.

- YOLOv3 [56] is a classical real-time detector and represents an earlier generation of the YOLO family.
- YOLOv7 [57] is an efficient YOLO-based detector designed to achieve a strong balance between detection accuracy and inference efficiency.
- BL-YOLOv8 [58] is an improved YOLOv8-based model developed specifically for road defect detection.
- YOLOv6 [59] is an efficient one-stage detector designed for practical object detection applications.
- YOLOX [60] is an anchor-free one-stage detector that removes the dependence on predefined anchor boxes.
- YOLO11s [51] is adopted as a recent lightweight YOLO baseline and serves as the principal reference model for comparison with our proposed framework.

D. Comparative Experiments (RQ1)

To comprehensively evaluate the detection performance of the proposed HCP-RDD model, this study compares it with representative road damage detection methods to verify the effectiveness of the proposed approach. As shown in Table II, HCP-RDD achieves the best overall detection performance among all compared methods: the recall is 89.98%, the precision is 94.27%, mAP@0.50 reaches 92.36%, mAP@0.50:0.95 is 79.29%, and the F1-score is 92.08%. Compared with BL-YOLOv8, which has the highest mAP@0.50 among the other methods in Table II, HCP-RDD improves mAP@0.50 by 1.62 percentage points (90.74% to 92.36%) and achieves higher precision and recall.

Specifically, the classical two-stage detector Faster R-CNN imposes substantially higher computational overhead than other methods. Meanwhile, SSD and early YOLO variants yield inferior detection metrics compared with modern one-stage detectors. BL-YOLOv8 and YOLOX serve as strong competitive baselines, which reflect the advantages of state-of-the-art one-stage detectors in terms of accuracy–efficiency trade-off.

Compared with the YOLO11s baseline, HCP-RDD increases recall by 3.11 percentage points, precision by 3.23 percentage points, mAP@0.50 by 4.14 percentage points, mAP@0.50:0.95 by 2.81 percentage points, and F1-score by 3.17 percentage points. Meanwhile, the computational cost in GFLOPs rises from 21.6 to 37.2, demonstrating that the performance gain comes with extra computational overhead.

In summary, the hierarchical caption-prefix mechanism introducing scene textual priors can effectively compensate for the limitations of pure visual features. With gate-regulated cross-attention, textual priors are utilized to hierarchically enhance multi-scale visual features, which alleviates the challenge of fine-grained road damage recognition caused by illumination variations and shadow interference. Despite the increase in computational cost, the model obtains prominent accuracy improvements. The results demonstrate that integrating scene textual priors into the YOLO11s visual detection network is feasible and superior for the road damage detection task.

TABLE II. QUANTITATIVE COMPARISON WITH REPRESENTATIVE ROAD DAMAGE DETECTION METHODS

Model	Recall (%)	Precision (%)	mAP@0.50 (%)	mAP@0.50:0.95 (%)	F1-score (%)	GFLOPs
Faster R-CNN [53]	67.41	73.27	73.21	61.53	70.22	941.0
YOLOv5s [54]	86.23	89.31	85.24	73.66	87.74	15.8
SSD [55]	68.31	74.46	74.68	63.45	71.25	62.7
YOLOv3 [56]	82.72	85.45	82.31	73.43	84.06	13.0
YOLOv7 [57]	86.67	89.84	85.88	75.45	88.23	13.1
BL-YOLOv8 [58]	89.32	91.81	90.74	78.93	90.55	25.5
YOLOv6 [59]	86.74	89.32	85.56	74.94	88.01	44.0
YOLOX [60]	87.97	91.01	89.07	77.29	89.46	15.4
YOLO11s [51]	86.87	91.04	88.22	76.48	88.91	21.6
HCP-RDD	89.98	94.27	92.36	79.29	92.08	37.2

Bold values indicate the best result per metric column (minimum for GFLOPs).

TABLE III. ABLATION RESULTS OF CAM AND DGM

CAM	DGM	Recall (%)	Precision (%)	mAP@0.50 (%)	mAP@0.50:0.95 (%)	F1-score (%)	GFLOPs
		86.87	91.04	88.22	76.48	88.91	21.6
✓		88.91	92.15	90.05	77.50	90.50	29.6
	✓	89.65	92.98	91.26	78.64	91.28	30.3
✓	✓	89.98	94.27	92.36	79.29	92.08	37.2

Bold values indicate the best result per metric column (minimum for GFLOPs).

E. Ablation Experiments (RQ2)

The Cross-Attention Module (CAM) and Dynamic Gate Module (DGM) are embedded between the backbone and neck of YOLO11s to construct the complete HCP-RDD model. To clarify the individual contribution of the two sub-modules to detection performance, three groups of ablation experiments are conducted: only CAM, only DGM, and the combination of CAM and DGM. The experimental results are shown in Table III.

Both CAM and DGM can improve the detection performance of the YOLO11s baseline when used independently. With only CAM enabled, mAP@0.50 and F1-score reach 90.05% and 90.50%, respectively. When only DGM is activated, mAP@0.50 increases to 91.26% and the F1-score achieves 91.28%. When CAM and DGM work cooperatively, the full HCP-RDD model obtains the best overall performance, with mAP@0.50 up to 92.36% and F1-score reaching 92.08%.

By comparing the two single-module experimental groups, it can be observed that DGM yields larger performance gains than CAM, which verifies the effectiveness of the textual-prefix mechanism designed in this study. DGM enables priority perception of disease-oriented textual priors and performs adaptive weighting for image-text interaction increments, playing a vital role in suppressing interference from noisy captions and calibrating damage features. Meanwhile, CAM is also indispensable. It generates fine-grained image-text interaction increments and provides fundamental cross-modal information for feature enhancement. These two modules are not mutually exclusive but complementary to each other: CAM produces valid image-text interaction information, while DGM implements screening and regulation over such information. Only through their joint cooperation can the benefits of textual priors be fully exploited for road damage detection.

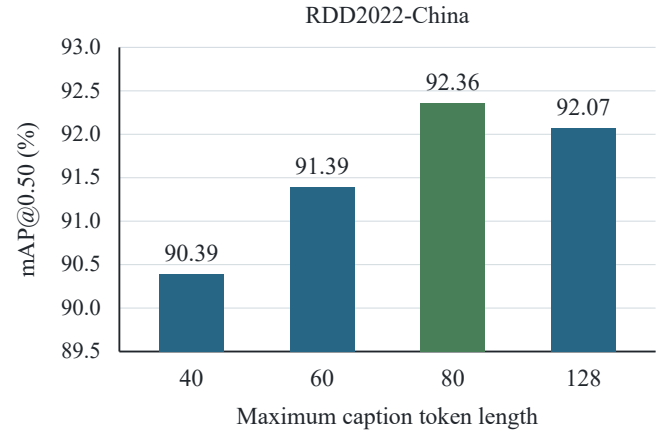


Fig. 3. Effect of maximum caption token length on HCP-RDD on RDD2022-China. Bars and numerical labels show mAP@0.50 (%); the green bar marks the selected 80-token configuration. The vertical axis starts at 89.5%.

F. Additional Experimental Analysis (RQ3)

1) *Scalability across YOLO11 model scales*: To verify that the HCPM is not only applicable to a single detector scale, we embed it into YOLO11n, YOLO11s, YOLO11m, YOLO11l and YOLO11x, respectively. For these five models with different sizes, the HCPM is inserted at the interface between the backbone network and the neck. Except for the size of the base YOLO11 model, the training settings, input resolution, data partition, caption configuration and evaluation metrics are kept consistent.

The experimental results are listed in Table IV. After introducing the HCPM, recall, precision, mAP and F1-score are improved across all five YOLO11 model scales. For

TABLE IV. SCALABILITY OF HCPM ACROSS DIFFERENT YOLO11 MODEL SCALES

Model	Recall (%)	Precision (%)	mAP@0.50 (%)	mAP@0.50:0.95 (%)	F1-score (%)	GFLOPs
YOLO11n	86.51	90.31	86.32	73.47	88.37	10.2
YOLO11n-HCPM	87.94	91.72	87.95	75.78	89.79	21.4
YOLO11s	86.87	91.04	88.22	76.48	88.91	21.6
YOLO11s-HCPM	89.98	94.27	92.36	79.29	92.08	37.2
YOLO11m	88.14	91.90	89.98	78.96	89.98	68.7
YOLO11m-HCPM	89.83	93.61	92.20	79.90	91.68	95.3
YOLO11l	89.07	92.78	90.77	78.95	90.89	87.1
YOLO11l-HCPM	89.93	93.72	92.12	80.11	91.79	122.9
YOLO11x	89.51	93.39	91.30	79.54	91.41	195.3
YOLO11x-HCPM	90.35	94.95	92.61	80.68	92.59	274.6

Bold values indicate the best result per metric column (minimum for GFLOPs).

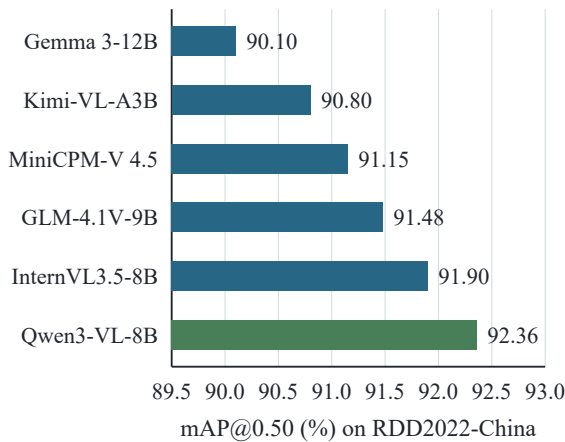


Fig. 4. Effect of caption generator on HCP-RDD on RDD2022-China. Bars and numerical labels show downstream detection mAP@0.50 (%); the green bar marks Qwen3-VL-8B-Instruct. Short model names are used for readability; full names are given in the text. The horizontal axis starts at 89.5%.

YOLO11n, mAP@0.50 and mAP@0.50:0.95 are increased by 1.63 and 2.31 percentage points, respectively. The corresponding improvements for YOLO11s are 4.14 and 2.81 percentage points. YOLO11m obtains gains of 2.22 and 0.94 percentage points, YOLO11l gains of 1.35 and 1.16 percentage points, and YOLO11x gains of 1.31 and 1.14 percentage points. The consistent performance gains at multiple scales sufficiently demonstrate that the HCPM possesses good generality and portability. Without relying on a specific network size, it can adapt to various YOLO11 detectors ranging from lightweight to large-scale ones, showing broad potential for deployment.

From YOLO11m to YOLO11x, the mAP@0.50 gain brought by the HCPM decreases from 2.22 to 1.35 and 1.31 percentage points. This trend is specific to mAP@0.50: the corresponding mAP@0.50:0.95 gains are 0.94, 1.16, and 1.14 percentage points, respectively. The accuracy improvement is accompanied by an increase in GFLOPs at every model scale. Therefore, embedding the HCPM into different YOLO11 variants represents an accuracy-complexity trade-off rather than cost-free performance improvement.

2) *Sensitivity to caption token length*: Four maximum caption token lengths, namely 40, 60, 80 and 128, are set to test the performance of HCP-RDD. Fig. 3 illustrates the mAP@0.50 on the RDD2022-China dataset under different token configurations. The experimental results show that mAP@0.50 rises from 90.39% at 40 tokens to 91.39% at 60 tokens, reaches the peak value of 92.36% at 80 tokens, and slightly drops to 92.07% at 128 tokens. This phenomenon indicates that when the caption token length is too short, the text cannot carry sufficient disease-related semantic knowledge, failing to provide adequate valid scene priors for the network. By contrast, excessive token length may introduce redundant text irrelevant to road damage, which interferes with image-text cross-modal interaction and degrades detection performance. Considering the above experimental observations, 80 is selected as the maximum caption token length for the proposed HCP-RDD model.

3) *Evaluation of LLM caption generators*: To explore the influence of disease-oriented captions generated by different open-source LLMs on the downstream road damage detection task, six representative open-source LLMs are selected as caption generators: Gemma 3-12B-IT [61], Kimi-VL-A3B-Thinking-2506 [62], MiniCPM-V 4.5 [63], GLM-4.1V-9B-Thinking [64], InternVL3.5-8B [65], and Qwen3-VL-8B-Instruct [66].

The maximum caption token length is uniformly set to 80 for all generators. The detector, dataset partition, training strategy, optimizer parameters and other hyperparameters are kept fixed. Only the caption generation model is replaced, and comparative tests are carried out on the China subset of RDD2022.

As shown in Fig. 4, Qwen3-VL-8B-Instruct achieves the best downstream detection performance when serving as the caption generator, with mAP@0.50 reaching 92.36% on the China subset. Compared with InternVL3.5-8B (mAP@0.50 = 91.90%), the detection metric is improved by 0.46 percentage points. The mAP@0.50 results of each model on the China subset are as follows: Gemma 3-12B-IT is 90.10%, Kimi-VL-A3B-Thinking-2506 is 90.80%, MiniCPM-V 4.5 is 91.15%, GLM-4.1V-9B-Thinking is 91.48%, InternVL3.5-8B is 91.90%, and Qwen3-VL-8B-Instruct is 92.36%.

It can be observed that Qwen3-VL-8B-Instruct and InternVL3.5-8B form the top-performing group among the evaluated generators. On the one hand, these models have

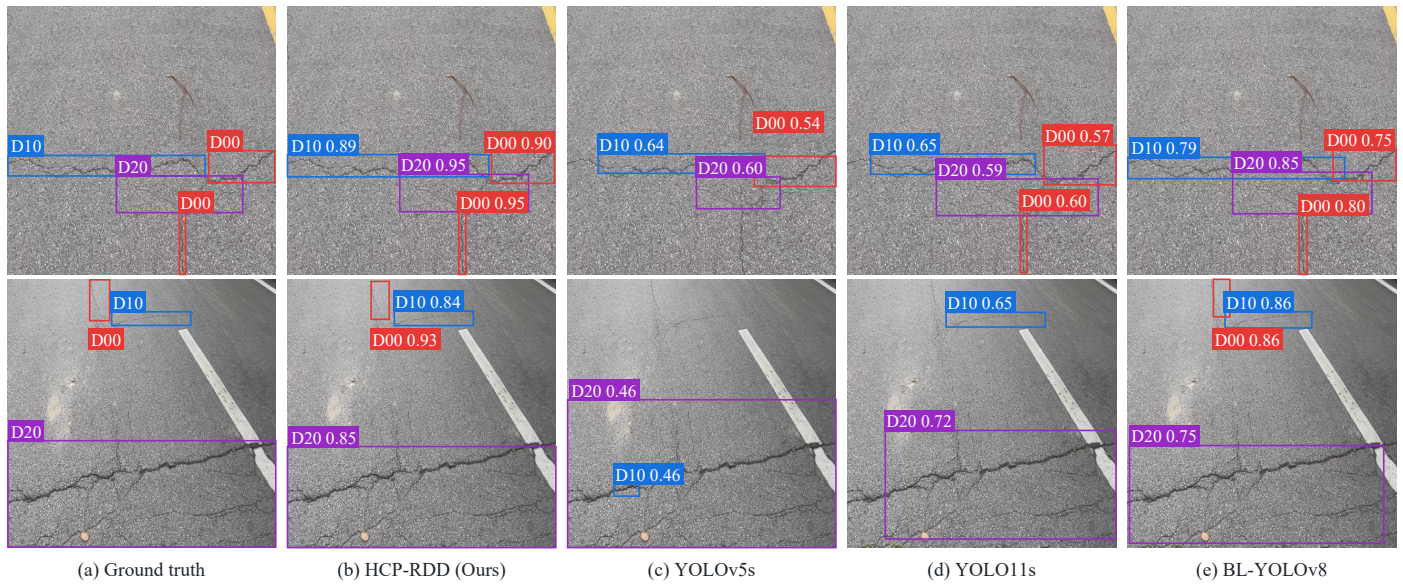


Fig. 5. Qualitative comparison on two representative images from the RDD2022-China MotorBike training split. Columns (a)–(e) show ground-truth annotations, HCP-RDD, YOLOv5s, YOLO11s, and BL-YOLOv8, respectively. Box labels give the damage category and, for model outputs, the confidence score. D00, D10, and D20 denote longitudinal, transverse, and alligator cracks, respectively. The first row shows intersecting thin cracks and co-existing damage regions; the second row shows large-area cracking together with small cracks.

TABLE V. CROSS-REGION EVALUATION ON RDD2022-JAPAN

Model	Recall (%)	Precision (%)	mAP@0.50 (%)	mAP@0.50:0.95 (%)	F1-score (%)
YOLOv5s	47.06	55.64	50.21	22.93	50.99
YOLO11s	51.09	58.73	53.71	26.08	54.64
HCP-RDD	54.89	61.10	58.06	30.26	57.83

Bold values indicate the best result per metric column.

stronger cross-modal semantic learning capabilities, which can accurately extract key damage information such as crack morphology, texture, and damage distribution on pavements. On the other hand, weaker or less damage-oriented captions may contain insufficient crack semantics or irrelevant background descriptions, introducing redundant text noise. Such noise interferes with image–text cross-modal feature interaction and ultimately restricts the downstream detection performance.

In summary, this experiment demonstrates that the quality of caption generators directly affects the effectiveness of textual priors. Captions with noise or semantic deficiency will degrade the performance of image–text fusion. This finding validates the design philosophy of the proposed HCP-RDD framework: high-quality disease-oriented captions are a prerequisite for textual priors to bring performance improvements. Only precise and effective scene semantic descriptions can assist the visual network in addressing the challenge of fine-grained damage recognition caused by illumination variations and shadow interference. According to the comparative experimental results, Qwen3-VL-8B-Instruct is adopted as the default caption generator for HCP-RDD.

4) *Cross-regional generalization*: To further verify the generalization capability of the proposed HCP-RDD method, cross-regional evaluation is conducted on the RDD2022-Japan

subset. As shown in Table V, HCP-RDD achieves the best detection performance among all comparison models, with mAP@0.50 of 58.06%, mAP@0.50:0.95 of 30.26%, precision of 61.10% and recall of 54.89%. Compared with YOLOv5s and the YOLO11s baseline, HCP-RDD obtains consistent improvements in mAP, precision and recall. The results demonstrate that after introducing the caption-conditioned feature enhancement mechanism, the model retains superior damage localization and fine-grained recognition capacity for out-of-distribution road scenes.

In summary, even when road conditions, pavement textures and visual characteristics differ greatly from the China dataset, the HCPM can still output reliable and effective semantic priors to help the model identify pavement damage such as cracks from diverse regional scenarios. This experiment further validates the effectiveness of the HCPM, which can effectively alleviate the challenge of fine-grained damage recognition caused by illumination variations and shadow interference. The proposed scheme is not only suitable for domestic road detection scenarios, but also has substantial potential for application to road damage detection tasks in other countries.

G. Qualitative Results (RQ4)

To investigate the effectiveness of HCP-RDD in complex road scenes and intuitively analyze its localization and missed-detection behavior, qualitative visualization experiments are conducted, as shown in Fig. 5.

On the one hand, compared with the conventional YOLOv5s, HCP-RDD successfully detects all annotated damage regions in both examples, whereas YOLOv5s misses one D00 instance in the first scene and retains only two targets in the second scene. This indicates that the proposed caption-prefix can complement pure visual features and improve fine-grained damage recognition under complex conditions involv-

ing intersecting cracks, small targets, and coexisting pavement distresses.

On the other hand, compared with the stronger BL-YOLOv8, HCP-RDD still exhibits more stable detection performance in complex road scenes. Since BL-YOLOv8 mainly relies on visual features and lacks effective caption-based semantic guidance, its feature representation can still be limited when dealing with fine-grained cracks, small targets, and complex background interference, leading to a higher risk of missed detections. In contrast, HCP-RDD introduces disease-oriented textual priors through the hierarchical caption-prefix mechanism to provide additional semantic guidance for multi-scale visual features, further demonstrating the effectiveness of the proposed caption-prefix mechanism in complex road damage detection.

Overall, the visualization results further confirm the effectiveness of our framework, demonstrating that HCP-RDD can alleviate fine-grained recognition difficulties caused by complex visual interference and provide a practical approach for vision-text fusion in road damage detection.

V. CONCLUSION

This study targets the limitations of existing road damage detection methods, including reliance on visual features alone, degraded fine-grained recognition under illumination and shadow interference, and negative effects brought by noisy scene captions. A hierarchical caption-prefix road damage detection framework (HCP-RDD) based on YOLO11s is proposed. By embedding the cross-attention module and dynamic gate module, the framework adopts BERT to encode disease-oriented textual priors generated by LLMs. It realizes image-text interaction incremental learning, scale-adaptive weighting and residual incremental reprojection to achieve hierarchical multi-scale feature enhancement, while retaining raw visual information. A bypass mechanism ensures stable baseline performance when valid captions are absent. Quantitative experiments, ablation studies and cross-region tests on the RDD2022 dataset demonstrate that HCP-RDD outperforms SOTA by 1.62 percentage points in mAP@0.50 and achieves higher precision and recall, effectively improving fine-grained road damage detection under complex interference.

In conclusion, this work verifies the effectiveness of incorporating textual semantic priors into road damage detection and provides a new perspective for controllable vision-text fusion in pavement distress recognition. Future research will optimize the cross-modal fusion mechanism, explore the influence of noisy and missing captions, balance detection accuracy and inference speed, and conduct more experiments on complex and cross-regional road scenes to further improve the practical applicability and generalization ability of the proposed framework in real-world road inspection tasks.

ACKNOWLEDGMENT

This work was supported by the Liaoning Provincial Science and Technology Plan Joint Program (General Project of Natural Science Foundation) (JH4/4800, No. 2024-MSLH-483).

REFERENCES

- [1] H. M. S. Badar, I. Hussain, A. K. Bashir, N. Alturki, G. Fan, and C. Zhang, "Edge-optimized lightweight and transformer backbones for real-time road damage detection in IIoT systems," *IEEE Internet of Things Journal*, 2025.
- [2] C. Wu, M. Ye, J. Zhang, and Y. Ma, "YOLO-lwnet: A lightweight road damage object detection network for mobile terminal devices," *Sensors*, vol. 23, no. 6, p. 3268, 2023.
- [3] M. W. Khan, M. S. Obaidat, K. Mahmood, B. Sadoun, H. M. S. Badar, and W. Gao, "Real-time road damage detection using an optimized YOLOv9s-fusion in IoT infrastructure," *IEEE Internet of Things Journal*, vol. 12, no. 11, pp. 17 649–17 660, 2025.
- [4] W. Pan, X. Wang, and W. Huan, "Real-time dynamic scale-aware fusion detection network: take road damage detection as an example," *Journal of Real-Time Image Processing*, vol. 22, no. 2, p. 55, 2025.
- [5] H.-C. Dan, P. Yan, J. Tan, Y. Zhou, and B. Lu, "Multiple distresses detection for asphalt pavement using improved you only look once algorithm based on convolutional neural network," *International Journal of Pavement Engineering*, vol. 25, no. 1, p. 2308169, 2024.
- [6] Y. Du, N. Pan, Z. Xu, F. Deng, Y. Shen, and H. Kang, "Pavement distress detection and classification based on YOLO network," *International Journal of Pavement Engineering*, vol. 22, no. 13, pp. 1659–1672, 2021.
- [7] G. Guo and Z. Zhang, "Road damage detection algorithm for improved YOLOv5," *Scientific Reports*, vol. 12, no. 1, p. 15523, 2022.
- [8] N. Kheradmandi and V. Mehranfar, "A critical review and comparative study on image segmentation-based techniques for pavement crack detection," *Construction and Building Materials*, vol. 321, p. 126162, 2022.
- [9] Q. Song and J. Tian, "Hybrid graph convolutional and deep convolutional networks for enhanced pavement crack detection," *Engineering Applications of Artificial Intelligence*, vol. 145, p. 110227, 2025.
- [10] Y. Zang, W. Li, J. Han, K. Zhou, and C. C. Loy, "Contextual object detection with multimodal large language models," *International Journal of Computer Vision*, vol. 133, no. 2, pp. 825–843, 2025.
- [11] Y. Zhou, S. Zhao, Y. Chen, Z. Wang, C. Jin, M. Zhao, and D. N. Metaxas, "LED: LLM enhanced open-vocabulary object detection without human curated data generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 9643–9653.
- [12] J. Cao, B. Peng, M. Gao, H. Hao, X. Li, and H. Mou, "Object detection based on CNN and vision-transformer: a survey," *IET Computer Vision*, vol. 19, no. 1, p. e70028, 2025.
- [13] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "DETRs beat yolos on real-time object detection," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024, pp. 16 965–16 974.
- [14] W. Wang, X. Li, X. Lyu, T. Zeng, J. Chen, and S. Chen, "Multi-attribute NMS: an enhanced non-maximum suppression algorithm for pedestrian detection in crowded scenes," *Applied Sciences*, vol. 13, no. 14, p. 8073, 2023.
- [15] P. Yan, X. Qi, and L. Jiang, "Li-YOLOv8: Lightweight small target detection algorithm for remote sensing images that combines GSConv and PConv," *PLOS ONE*, vol. 20, no. 5, p. e0321026, 2025.
- [16] J. Hua, Z. Wang, X. Tian, Q. Zou, J. Xiao, and J. Ma, "Full perception head: Bridging the gap between local and global features," *IEEE/CAA Journal of Automatica Sinica*, vol. 12, no. 7, pp. 1391–1406, 2025.
- [17] H. Xie, Q. Dong, S. Duan, K. Wang, Z. Li, and S. Zhao, "Echf-YOLO: a remote sensing object detection network integrating efficient convolution and high-low frequency features," *Big Earth Data*, vol. 10, no. 1, pp. 491–524, 2026.
- [18] X. Liu, G. Li, Y. Qi, Z. Han, A. van den Hengel, N. Sebe, M.-H. Yang, and Q. Huang, "Consistency-aware anchor pyramid network for crowd localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [19] H. Liu, Q. Tong, L. Miao, and X. Liu, "Effective anchor adaptation and feature enhancement strategies for tiny object detection in aerial images," *IEEE Access*, vol. 12, pp. 69 677–69 689, 2024.

- [20] P. Nguyen, "Ha-rdet: Hybrid anchor rotation detector for oriented object detection," in *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE, 2025, pp. 2910–2919.
- [21] W. Tao, X. Wang, T. Yan, Z. Liu, and S. Wan, "Esf-YOLO: an accurate and universal object detector based on neural networks," *Frontiers in Neuroscience*, vol. 18, p. 1371418, 2024.
- [22] H. Ouyang, "DEYO: DETR with YOLO for step-by-step object detection," *arXiv preprint arXiv:2211.06588*, 2022.
- [23] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS—improving object detection with one line of code," in *2017 IEEE international conference on computer vision (ICCV)*. IEEE, 2017, pp. 5562–5570.
- [24] S. Liu, D. Huang, and Y. Wang, "Adaptive NMS: Refining pedestrian detection in a crowd," in *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE, 2019, pp. 6452–6461.
- [25] C. Symeonidis, I. Mademlis, I. Pitas, and N. Nikolaidis, "Neural attention-driven non-maximum suppression for person detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 2454–2467, 2023.
- [26] Z. Luo, Z. Fang, S. Zheng, Y. Wang, and Y. Fu, "NMS-loss: Learning with non-maximum suppression for crowded pedestrian detection," in *proceedings of the 2021 international conference on multimedia retrieval*, 2021, pp. 481–485.
- [27] S. Liu, T. Ren, J. Chen, Z. Zeng, H. Zhang, F. Li, H. Li, J. Huang, H. Su, J. Zhu *et al.*, "Detection transformer with stable matching," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 6468–6477.
- [28] D. Jia, Y. Yuan, H. He, X. Wu, H. Yu, W. Lin, L. Sun, C. Zhang, and H. Hu, "DETRs with hybrid matching," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 19 702–19 712.
- [29] Y.-X. Huang, H.-I. Liu, H.-H. Shuai, and W.-H. Cheng, "Dq-DETR: DETR with dynamic query for tiny object detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 290–305.
- [30] S. Xu, Y. Li, M. Lin, P. Gao, G. Guo, J. Lü, and B. Zhang, "Q-DETR: An efficient low-bit quantized detection transformer," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 3842–3851.
- [31] J. Jiang, K. Xu, Y. Song, L. Gao, and J. Zhang, "Investigation of cracking behavior in asphalt pavement using digital image processing technology," *Frontiers in Built Environment*, vol. 11, p. 1580379, 2025.
- [32] Z. Li, C. Yin, and X. Zhang, "Crack segmentation extraction and parameter calculation of asphalt pavement based on image processing," *Sensors*, vol. 23, no. 22, p. 9161, 2023.
- [33] L. Fan, S. Li, Y. Li, B. Li, D. Cao, and F.-Y. Wang, "Pavement cracks coupled with shadows: A new shadow-crack dataset and a shadow-removal-oriented crack detection approach," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 7, pp. 1593–1607, 2023.
- [34] K. Kirschke and S. Velinsky, "Histogram-based approach for automated pavement-crack sensing," *Journal of Transportation Engineering*, vol. 118, no. 5, pp. 700–710, 1992.
- [35] L. Ying and E. Salari, "Beamlet transform-based technique for pavement crack detection and classification," *Computer-Aided Civil and Infrastructure Engineering*, vol. 25, no. 8, pp. 572–580, 2010.
- [36] Y. Shi, L. Cui, Z. Qi, F. Meng, and Z. Chen, "Automatic road crack detection using random structured forests," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 12, pp. 3434–3445, 2016.
- [37] H. Oliveira and P. L. Correia, "Automatic road crack detection and characterization," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 155–168, 2012.
- [38] L. Ali, F. Alnajjar, N. Zaki, and H. Aljassmi, "Pavement crack detection by convolutional adaboost architecture," in *8th Zero Energy Mass Custom Home International Conference, ZEMCH*, vol. 2021, 2021, pp. 383–394.
- [39] N.-D. Hoang and Q.-L. Nguyen, "A novel method for asphalt pavement crack classification based on image processing and machine learning," *Engineering with Computers*, vol. 35, no. 2, pp. 487–498, 2019.
- [40] C. Kyriakou, S. E. Christodoulou, and L. Dimitriou, "Spatial road-way condition-assessment mapping utilizing smartphones and machine learning algorithms," *Transportation Research Record*, vol. 2675, no. 9, pp. 1118–1126, 2021.
- [41] A. Cord and S. Chambon, "Automatic road defect detection by textural pattern recognition based on adaboost," *Computer-Aided Civil and Infrastructure Engineering*, vol. 27, no. 4, pp. 244–259, 2012.
- [42] B.-K. Kwon, J.-S. Won, and D.-J. Kang, "Fast defect detection for various types of surfaces using random forest with VOV features," *International Journal of Precision Engineering and Manufacturing*, vol. 16, no. 5, pp. 965–970, 2015.
- [43] B. Liu, J. Kang, H. Guan, L. Ma, Y. Yu, S. Li, and J. Li, "Boundary-guided real-time semantic segmentation and pixel-level quantification of pavement cracks," *IEEE Transactions on Intelligent Transportation Systems*, 2025.
- [44] H. Maeda, Y. Sekimoto, T. Seto, T. Kashiyama, and H. Omata, "Road damage detection using deep neural networks with images captured through a smartphone," *arXiv preprint arXiv:1801.09454*, 2018.
- [45] D. Ma, H. Fang, B. Xue, F. Wang, M. A. Msekh, and C. L. Chan, "Intelligent detection model based on a fully convolutional neural network for pavement cracks," *Computer Modeling in Engineering & Sciences*, vol. 123, no. 3, pp. 1267–1291, 2020.
- [46] F. Yang, L. Zhang, S. Yu, D. Prokhorov, X. Mei, and H. Ling, "Feature pyramid and hierarchical boosting network for pavement crack detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 4, pp. 1525–1535, 2019.
- [47] K. Doshi and Y. Yilmaz, "Road damage detection using deep ensemble learning," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 5540–5544.
- [48] D. Arya, H. Maeda, S. K. Ghosh, D. Toshniwal, and Y. Sekimoto, "RDD2022: A multi-national image dataset for automatic road damage detection," *Geoscience Data Journal*, vol. 11, no. 4, pp. 846–862, 2024.
- [49] Z. Liu, K. Jing, K. Yang, Z. Zhang, and X. Li, "Efficient dense attention fusion network with channel correlation loss for road damage detection," *IET Intelligent Transport Systems*, vol. 18, no. 10, pp. 1747–1759, 2024.
- [50] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.
- [51] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024.
- [52] D. Arya, H. Maeda, S. K. Ghosh, D. Toshniwal, and Y. Sekimoto, "CRDDC2022: Crowd-sensing based road damage detection challenge," in *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 2022, pp. 5228–5235.
- [53] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, vol. 28, 2016.
- [54] Y. Jiang, H. Yan, Y. Zhang, K. Wu, R. Liu, and C. Lin, "RDD-YOLOv5: Road defect detection algorithm with self-attention based on unmanned aerial vehicle inspection," *Sensors*, vol. 23, no. 19, p. 8241, 2023.
- [55] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.
- [56] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018.
- [57] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 7464–7475.
- [58] X. Wang, H. Gao, Z. Jia *et al.*, "BI-YOLOv8: An improved road defect detection model based on YOLOv8," *Sensors*, vol. 23, no. 20, p. 8361, 2023.
- [59] C. Li, L. Li, Y. Geng, H. Jiang, and M. Xu, "YOLOv6: A single-stage object detection framework for industrial applications," 2022.
- [60] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021.
- [61] Gemma Team, "Gemma 3 technical report," 2025. [Online]. Available: <https://arxiv.org/abs/2503.19786>
- [62] Kimi Team, "Kimi-vl technical report," 2025. [Online]. Available: <https://arxiv.org/abs/2504.07491>

- [63] T. Yu, Z. Wang, C. Wang *et al.*, “MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 11 704–11 715.
- [64] GLM-V Team, “GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning,” 2025. [Online]. Available: <https://arxiv.org/abs/2507.01006>
- [65] W. Wang, Z. Gao, L. Gu, H. Pu, L. Cui, X. Wei, Z. Liu, L. Jing, S. Ye, J. Shao *et al.*, “InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency,” 2025. [Online]. Available: <https://arxiv.org/abs/2508.18265>
- [66] Qwen Team, “Qwen3-VL technical report,” 2025. [Online]. Available: <https://arxiv.org/abs/2511.21631>