applied sciences

MDPI

Article

Open-Vocabulary Crack Object Detection Through Attribute-Guided Similarity Probing

Hyemin Yoon and Sangjin Kim *

Department of Management Information Systems, Dong-A University, Busan 49236, Republic of Korea; yoonlea1205@donga.ac.kr

* Correspondence: skim10@dau.ac.kr; Tel.: +82-051-200-7484

Abstract

Timely detection of road surface defects such as cracks and potholes is critical for ensuring traffic safety and reducing infrastructure maintenance costs. While recent advances in image-based deep learning techniques have shown promise for automated road defect detection, existing models remain limited to closed-set detection settings, making it difficult to recognize newly emerging or fine-grained defect types. To address this limitation, we propose an attribute-aware open-vocabulary crack detection (AOVCD) framework, which leverages the alignment capability of pretrained vision–language models to generalize beyond fixed class labels. In this framework, crack types are represented as combinations of visual attributes, enabling semantic grounding between image regions and natural language descriptions. To support this, we extend the existing PPDD dataset with attribute-level annotations and incorporate a multi-label attribute recognition task as an auxiliary objective. Experimental results demonstrate that the proposed AOVCD model outperforms existing baselines. In particular, compared to CLIP-based zero-shot inference, the proposed model achieves approximately a 10-fold improvement in average precision (AP) for novel crack categories. Attribute classification performance—covering geometric, spatial, and textural features—also increases by 40% in balanced accuracy (BACC) and 23% in AP. These results indicate that integrating structured attribute information enhances generalization to previously unseen defect types, especially those involving subtle visual cues. Our study suggests that incorporating attribute-level alignment within a vision–language framework can lead to more adaptive and semantically grounded defect recognition systems.

Keywords: open-vocabulary detection; deep learning; Port Road crack; vehicle-egocentric view data; data analysis

Academic Editor: Michele D'Amato

Received: 13 August 2025
Revised: 15 September 2025
Accepted: 19 September 2025
Published: 24 September 2025

Citation: Yoon, H.; Kim, S. Open-Vocabulary Crack Object Detection Through Attribute-Guided Similarity Probing. *Appl. Sci.* **2025**, *15*, 10350. https://doi.org/10.3390/app151910350

1. Introduction

Road infrastructure serves as a critical component of transportation systems from both social and economic perspectives. Over time, surface defects such as cracks and potholes emerge due to repeated traffic loads and various environmental factors. If these defects are not detected and repaired in a timely manner, they can lead to increased risks of traffic accidents and a significant reduction in road lifespan [1]. Consequently, the timely identification of road surface damage has become a major concern for minimizing maintenance costs. In response, automated detection of road defects using image-based computer vision techniques has seen rapid advancement in the research community. Such automated systems are highly valued for their ability to consistently assess road conditions without human intervention. For instance, Shaoxiang Li et al. [2] proposed RepGD-YOLOv8W, an

enhanced version of YOLOv8, trained on drone footage and motorcycle-mounted camera videos. Their model demonstrated improved accuracy and time efficiency compared to the original YOLOv8. Similarly, Wang et al. [3] introduced a modified architecture named C2f-Faster-EMA with a Partial Convolution approach and a redesigned detection head, which improved the model's expressiveness and practical performance, thereby confirming its feasibility for real-world applications. Beyond these CNN- and YOLO-based architectures, transformer-based methods have also been explored for road crack detection. Notably, CrackFormer [4] leverages self-attention mechanisms to capture long-range dependencies and fine-grained crack structures, while multi-scale feature fusion networks [5] integrate contextual information across scales to restore detailed crack patterns and enhance robustness. These developments illustrate the broad spectrum of approaches in crack detection and highlight the growing role of advanced deep learning architectures in infrastructure inspection.

Conventional object detectors operate in a closed-set regime, predicting only predefined classes and struggling to generalize to unseen categories or capture linguistic descriptions. Vision–language models (VLMs), notably CLIP [6], overcome this by learning alignments between images and text from large-scale datasets, enabling open-vocabulary detection (OVD). OVD replaces fixed classifiers with text embeddings, permitting recognition of novel objects; recent works demonstrate improved zero-shot detection on benchmarks such as COCO [7] and LVIS [8] using this approach. Examples include OVR-CNN [9], which integrates captions and GloVe embeddings, RegionCLIP [10], which aligns regions and phrases to refine localization, VLDet [11], which matches image–text sets with minimal architecture changes, and open-vocabulary attribute detection (OVAD) [12], which matches image–text sets with minimal architecture changes.

Applying OVD to road defect detection raises domain-specific challenges. Existing road datasets contain only a few classes (e.g., RDD includes four types of defects), restricting recognition of rare or emerging patterns. General VLMs trained on web images are not optimized for the subtle textures and degradation cues of road surfaces. Furthermore, many datasets provide coarse labels such as SHREC2022's "crack" or "pothole" [13] and the broad "D00" code in the RDD2020 dataset [14]—hampering discrimination of fine-grained variations. These factors highlight the need for models that adapt flexibly across regions, conditions, and time. This lack of detailed labeling makes it difficult for models to learn discriminative features between subtle variations of road defects and hinders effective fine-grained classification. This study aims to introduce an attribute-aware open-vocabulary crack detection (AOVCD) framework for road defect detection. While prior OVD methods demonstrate strong generalization by aligning visual features with category-level semantics, such category-only alignment is insufficient in crack detection, where novel crack types often emerge as new combinations of geometric/visual attributes rather than entirely new categorical concepts. As illustrated in Figure 1, to address this gap, our framework defines road crack types as combinations of visual attributes and incorporates attribute-level alignment in the embedding space of a pretrained language model, thereby capturing fine-grained variations essential for detecting previously unseen defect types. To the best of our knowledge, this is the first attempt to apply the AOVCD approach to the domain of road crack detection. This paper examines whether attribute alignment improves the scalability of the text-embedding space for open-vocabulary crack detection. Our goal is not to benchmark detection heads; instead, we hold the vision backbone and text encoder fixed across variants and isolate the effect of attribute alignment. Accordingly, we adopt a ViLD-style OVAD recipe as a representative baseline.

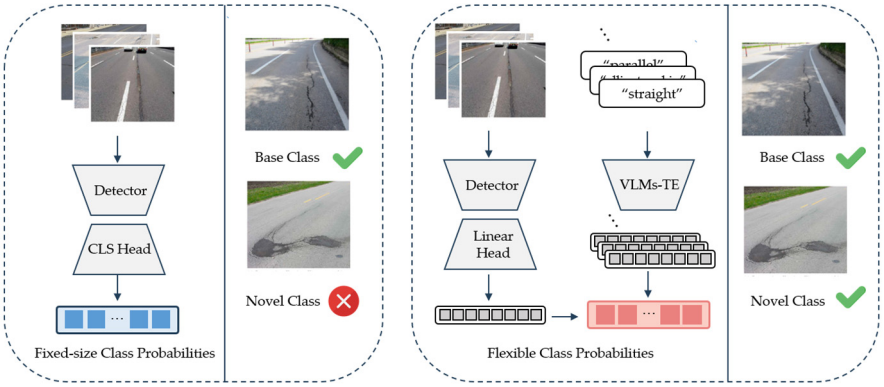


Figure 1. Conceptual overview comparing (**a**) a conventional closed-set detector, which only recognizes predefined classes, and (**b**) our attribute-aware open-vocabulary crack detector. The latter uses text embeddings (VLMs-TE) to align visual features with attribute descriptions, enabling detection of unseen road defects beyond the fixed label space.

The contributions of the proposed method in this paper can be summarized as follows:

- This study addresses the need for a universal system capable of adapting to various defect types and real-world environmental changes and proposes a flexible and scalable detection framework for road maintenance. By applying open-vocabulary detection to the road defect detection domain, it overcomes the limitations of models trained on fixed class sets.
- We introduce an extended annotation for the PPDD dataset in which road crack types are defined as combinations of visual attributes, allowing for representation learning that can generalize to novel classes.
- We empirically demonstrate that a model trained with attribute-based alignment achieves better performance in detecting unseen categories (novel classes) compared to conventional class-only models.

This study aims to present a viable direction for building more flexible and scalable detection systems in the field of road maintenance and proposes a generalized framework capable of handling the diversity of road defects and real-world environmental variability.

2. Related Work

2.1. Object Detection of Pavement Detection

Traditional manual inspection methods for road defect maintenance require substantial human labor, are inherently subjective, and become inefficient in large-scale environments. As a result, automated road defect detection using deep learning approaches has emerged as a critical task for intelligent transportation systems and infrastructure maintenance and is being actively studied. One of the representative benchmark datasets in this domain is the Road Damage Detection (RDD) dataset proposed by Maeda et al. [14]. It provides road surface images and bounding box annotations collected via vehicle-mounted smartphone cameras in multiple countries. The dataset includes instances of common road defects such as longitudinal cracks, transverse cracks, potholes, and alligator cracks. Various deep learning-based detectors have been trained and evaluated using this benchmark, with models such as YOLO, Faster R-CNN, and SSD demonstrating promising performance under controlled conditions [15,16]. The Crack500 dataset [17], another benchmark designed for defect-background segmentation in road imagery, offers 500 road images encompassing a wide range of backgrounds and crack types, along with precise pixel-level annotations of crack boundaries. Additionally, the GAPs384 dataset [18], collected in Germany, comprises asphalt road images captured nearly vertically and includes annotations for six types of

defects. These benchmarks reflect the visual characteristics of defects in real-world road environments and serve as a critical foundation for evaluating the performance of road defect detection models.

Early applications of deep learning for pavement defect detection primarily focused on pixel-level segmentation to distinguish cracks from non-crack regions. SDDNet [19] leveraged depthwise convolution to achieve real-time crack segmentation while maintaining accuracy, thereby enhancing computational efficiency. Tang et al. [20] addressed the class imbalance issue by designing EDNet, an encoder–decoder architecture that improves robustness and precision by balancing the learning between crack and background regions. Guo et al. [21] proposed an architecture that adaptively refines crack boundaries and enhances edge detection. Qu et al. [22] introduced a CNN-based architecture that enables effective segmentation of cracks at various scales by directly transferring high-level features to low-level feature maps through multi-scale feature fusion. To incorporate global contextual information in crack images, transformer-based architectures have also been adopted to improve segmentation performance. Liu et al. [4] proposed CrackFormer, which integrates self-attention blocks into a SegNet structure. In this model, the CNN backbone captures local textures while the transformer architecture learns long-range dependencies. This approach is particularly effective in enhancing performance by detecting textural differences between cracks and the road surface, rather than relying solely on shape variations.

Meanwhile, object detection tasks have also been actively explored in the context of road defect detection, particularly for localizing and classifying anomalies such as potholes and complex structural damage. Faster R-CNN, for example, achieved an F1-score of 49% on benchmark datasets like RDD2022 [23], demonstrating the utility of two-stage detectors for road defect detection [24]. However, due to their relatively slow inference speed, these models face limitations in real-time applications. As a result, one-stage detectors such as the YOLO family have been increasingly studied to balance both efficiency and accuracy. Among the YOLO variants, models ranging from YOLOv5 to YOLOv8 have shown outstanding performance in terms of both accuracy and speed, making them popular choices as backbones in many studies. The C2f-Faster-EMA module proposes an architecture that replaces the standard C2f module in YOLOv8s with a partial convolution block and redesigns the detection head, achieving a 5.8% improvement in mAP@0.5 while reducing model size and complexity by over 20% [3]. YOLOv8-PD [25] enhances feature extraction for road defects and reduces computational cost by incorporating a global information extraction module and a Large Separable Kernel Attention mechanism. These advancements in road defect detection have proven effective in achieving both high accuracy and real-time performance. Nonetheless, most existing studies have been conducted under the closed-set assumption, constrained by the labels provided in the training datasets. For instance, the widely used RDD2020 dataset includes only four types of road defects—longitudinal cracks, transverse cracks, alligator cracks, and potholes—limiting its applicability to the recognition of novel or more fine-grained defect types. This highlights a gap in the literature and underscores the need for open-vocabulary or flexible detection approaches that can generalize to evolving road conditions.

2.2. Open-Vocabulary Detection

Early vision–language models (VLMs) such as CLIP [6] and ALIGN [26] have been developed to align visual and textual information in a joint embedding space through contrastive learning on millions of image–text pairs. These pretrained models achieve superior performance in zero-shot image classification by aligning the global semantic meaning between entire images and their textual descriptions. However, their focus

on image-level alignment imposes limitations in distinguishing fine-grained details or attributes of specific localized regions within an image. In domains where precise attribute identification—such as size, color, or texture—is essential, global alignment alone proves insufficient. To address this issue, subsequent research studies such as RegionCLIP [10], MDETR [27], and GLIP [28] have attempted region-level or phrase-level alignment to better capture localized visual information. RegionCLIP learns alignment by automatically generating region–phrase pairs without relying on predefined bounding boxes, achieving strong performance on benchmarks such as COCO [7] and LVIS [8]. MDETR, a transformer-based detection model conditioned on text, enables phrase-to-region alignment for handling referring expressions. GLIP proposes a unified framework that integrates object detection with phrase grounding, allowing models to detect objects based on text prompts. While these models demonstrate the feasibility of more localized vision–language alignment, challenges remain in preserving high-resolution features, adapting to specific domains, and generalizing fine-grained alignments. These challenges are particularly salient in domains like road defect detection, which require distinguishing subtle differences in small cracks or textures, thereby necessitating more sophisticated alignment techniques.

Open-vocabulary object detection (OVD) is an approach that enables the detection of novel objects based on arbitrary text prompts, without relying on predefined class labels. Zareian et al. [9] first introduced OVD by leveraging image–caption pairs as weak supervision signals. Since then, various models such as ViLD [29], RegionCLIP [10], GLIP [28], Detic [30], and VLDet [11] have been developed. ViLD injects CLIP's language embeddings into the detector's classifier, while Detic is trained on over 20,000 categories using web-crawled tag-based data and achieves strong performance on the LVIS benchmark. These models are typically evaluated on benchmarks like COCO and LVIS for their zero-shot detection capabilities. The application potential of OVD is particularly high in road defect detection, where defect types are diverse and new forms frequently emerge. However, road imagery differs significantly from web images in terms of visual characteristics, making it challenging for pretrained models to generalize—necessitating domain adaptation strategies. As an extension of OVD, open-vocabulary attribute detection (OVAD) [12] has gained attention for its ability to recognize object attributes in a zero-shot manner using text-based prompts. While traditional attribute classification has been treated as a multi-label classification problem over a fixed attribute set, OVAD evaluates fine-grained attribute prediction by incorporating attribute annotations into datasets such as COCO. Models like CLIP and OpenCLIP [31] predict attributes by computing the similarity between visual embeddings of object regions and textual embeddings of attribute terms. OvarNet [32] enhances semantic alignment by using sentence-form prompts that include attributes and further improves performance by organizing attributes into higher-level categories such as color and material. Existing open-vocabulary detection frameworks largely rely on category-level supervision (e.g., Detic, RegionCLIP, OVAD), whereas attribute-based recognition literature has long emphasized that attributes function as transferable and compositional cues in zero-shot tasks. However, in the context of road defect detection, category-only alignment is insufficient because practical maintenance requires fine-grained details such as the location, size, and orientation of defects. Unlike OvarNet, which focuses on attribute recognition in natural image domains through a two-stage training process, our study integrates attribute-level signals into the OVD pipeline and specifically addresses the domain shift challenge inherent in road defect imagery. We propose attribute-aware open-vocabulary crack detection (AOVCD), which establishes a baseline for the task and lays the foundation for future OVD applications in this critical domain.

3. Methods

3.1. Preliminaries

The AOVCD aims to accomplish two tasks: (1) category alignment and (2) attribute alignment to enable generalized identification of a given crack patch. The category alignment is learned through the open-vocabulary detection (OVD) task, while the attribute alignment is formulated as a multi-label classification (MLC) task, trained on vision–language matching derived from natural language descriptions of crack objects.

The OVD task considers two disjoint sets of object categories: the base categories O^B and the novel categories O^N. Each input image $I \in \mathbb{R}^{H\times W\times 3}$, where the last dimension corresponds to the RGB color channels, is annotated with a set of object category labels $y^c = \{y_1^c, \cdots, y_K^c\}$, where K denotes the number of object instances present in the image and $c \in C$. Here, C represents the predefined set of crack categories. Each label y_k^c belongs to either the set of base categories O^B or novel categories O^N, such that the total number of category classes is $C = |O^B| + |O^N|$. During training, the model learns from the base categories O^B, and during evaluation, it must identify correct categories from a dataset that includes both unseen categories O^N and unseen samples from the seen categories O^B. We refer to this as the generalized OVD task. To verify the effectiveness of the proposed idea, we adopt a conventional object detection framework.

In the multi-label classification (MLC) task, the model evaluates the assignment of multiple labels to each object instance from a fixed-size label space A. To enable generalized identification and scalability of crack categories, we construct a fixed-size set of attribute category labels $y^a = \{y_1^a, \cdots, y_K^a\}$ based on diverse text descriptions for each crack category, where $y_k^a \in \{0, 1\}^{|A|}$ and $a \in A$. Here, A denotes the predefined attribute set. Each instance is then assigned one or more attribute categories. Notably, the attributes associated with the novel categories O^N include both seen and unseen labels. This design reflects the fact that a specific object instance can possess shared attributes at a coarse-grained level while also exhibiting distinct attributes at a fine-grained level. Furthermore, the model allows for the insertion of fine-grained attribute categories without additional training after deployment. For clarity, we denote the object-level crack detection baseline as OVCD (object-level crack detection) and our proposed attribute-augmented variant as AOVCD (attribute-aware open-vocabulary crack detection). These terms will be used consistently throughout the paper.

3.2. AOVCD Baseline Methods

For the AOVCD baseline, the architecture of the open-vocabulary detector is adapted based on the ViLD and OVAD. Figure 2 illustrates the overall structure of our proposed AOVCD baseline, which consists of a Faster R-CNN based object detector D_{det} and a pretrained vision–language model composed of an image encoder Φ_V and a text encoder Φ_T. We modify the original classification head—typically trained to predict fixed category indices—by replacing it with a linear projection layer that maps region features to the same embedding space as the text encoder. This allows region-level visual features to be aligned with category-level text embeddings. The proposal features generated by the Region Proposal Network (RPN) are further trained to align closely with the global image embedding produced by the image encoder, encouraging semantic consistency between region-level and global representations. Here, p_i denotes the i-th region proposal generated by the RPN, and $f(p_i)$ is the corresponding region-level feature vector. For the vision–language backbone, we adopt the CLIP model, a dual-encoder architecture that encodes both images and text into a shared embedding space.

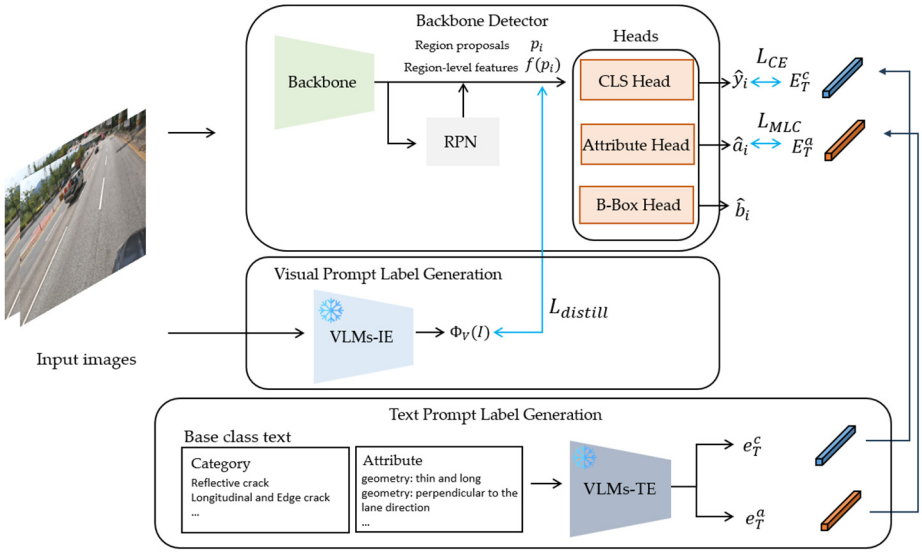


Figure 2. Proposed training framework for AOVCD. Input images are processed by the backbone and RPN to generate region proposals and features $f(p_i)$, which are fed into classification head (CLS Head), attribute head (Attribute Head), and bounding box head (B-Box Head) to produce predictions $\hat{y}_i$, $\hat{a}_i$, $\hat{b}_i$. The text prompt label generation module encodes predefined class and attribute texts via VLMs-TE to obtain class embeddings E_T^c and attribute embeddings E_T^a. The visual prompt label generation module extracts image-level embeddings $\Phi_V(I)$ using VLMs-IE, which are used for the distillation loss $L_{distill}$. The final objective combines cross-entropy loss L_{CE}, multi-label contrastive loss L_{MLC}, and distillation loss $L_{distill}$.

3.3. Crack Category Alignment Loss

To enable open-vocabulary object classification, we leverage category-level language supervision from CLIP. Each predicted proposal feature $f(p_i)$ is trained to match the corresponding class embedding $t^c_{y_i^c} \in E_T^c$, where y_i^c is the ground-truth category of i-th region. Here, the text prompt label generation module encodes predefined class texts using VLMs-TE to construct the set of class embedding E_T^c. We formulate the object classification as a standard cross-entropy loss over the similarity scores between the proposal feature and all class text embeddings:

$$s_i^c = sim(f(p_i), E_T^c) \in \mathbb{R}^{|C|} \tag{1}$$

$$L_{CE} = -\sum_{i=1}^{K} \left(\log(\exp(s_i^c[y_i^c])) - \log\left(\sum_{j=1}^{|C|} \exp(s_i^c[j])\right)\right) \tag{2}$$

Here, $sim(\cdot)$ denotes the cosine similarity (or dot product) between the proposal feature and each class text embedding, normalized if needed. The classifier is non-parametric and defined entirely through CLIP-derived language embeddings, allowing our detector to generalize beyond the seen categories during training. Importantly, since the class vocabulary includes both seen categories (O^B) and novel categories (O^N), this formulation naturally supports generalized open-vocabulary detection (G-OVD). At inference time, predictions are scored over the entire set $O^B \cup O^N$, without retraining or finetuning.

3.4. Crack Attribute Alignment Loss

In addition to classifying object categories, our framework also predicts a set of visual attributes associated with each detected instance. Given a fixed vocabulary of attribute categories A and their corresponding CLIP text embeddings E_T^a, we treat attribute prediction as a multi-label classification (MLC) task. Here, the text prompt label generation module encodes predefined attribute texts via VLMs-TE to construct the set of attribute

embeddings E_T^a. For each proposal feature $f(p_i)$, we compute similarity scores against all attribute embeddings:

$$s_i^a = sim(f(p_i), E_T^a) \in \mathbb{R}^{|A|} \tag{3}$$

The predicted attribute probabilities are obtained by applying a sigmoid activation:

$$\hat{y}_i^a = \sigma(s_i^a) \tag{4}$$

The attribute loss is then defined as the binary cross-entropy loss over the ground-truth attribute labels $y_i^a \in \{0,1\}^{|A|}$:

$$L_{MLC} = -\sum_{i=1}^{K} \sum_{j=1}^{|A|} \left(y_i^a[j] \cdot \log \hat{y}_i^a[j] + (1 - y_i^a[j]) \cdot \log\left(1 - \hat{y}_i^a[j]\right)\right) \tag{5}$$

This design enables the model to assign multiple attribute labels per instance, capturing both shared and instance-specific properties. Moreover, by using CLIP-based embeddings as label anchors, our attribute classifier can generalize to unseen attribute descriptions semantically aligned with the vision–language space without modifying the classification head.

3.5. Total Training Objective

Our final training objective integrates the following three components: (1) object alignment loss L_{CE}, (2) attribute alignment loss L_{MLC}, and (3) distillation loss L_{distil}, which aligns the proposal feature $f(p_i)$ with the corresponding global or region-specific CLIP visual embedding. The distillation loss encourages the detector to inherit CLIP's rich visual semantics by minimizing the discrepancy between region features and CLIP activations over the same regions:

$$L_{distill} = -\sum_{i=1}^{M} \|f(p_i) - \Phi_V(p_i)\|_2^2 \tag{6}$$

where $\Phi_V(p_i)$ is the CLIP visual embedding of the cropped region p_i, or alternatively, a region-aware feature pooled from the CLIP image encoder.

The total objective is defined as

$$L_{total} = \lambda_c \cdot L_{CE} + \lambda_a \cdot L_{MLC} + \lambda_d \cdot L_{distill} \tag{7}$$

where λ_c, λ_a, and λ_d are hyperparameters that control the balance among classification, attribute prediction, and knowledge distillation. This joint objective enables the detector to benefit from vision–language priors for both category and attribute reasoning, while preserving the region-level localization capabilities of Faster R-CNN. The balancing weights (λ_c, λ_a, and λ_d) were fixed at 1:1:1 in order to directly compare the impact of category alignment versus attribute alignment, while avoiding additional hyperparameter complexity. This design follows the practice of prior OVD works [29] where loss weights are empirically chosen. Once trained, the model supports open-vocabulary and attribute-aware detection in a unified framework without the need for explicit retraining on novel categories or attribute compositions.

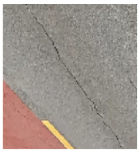
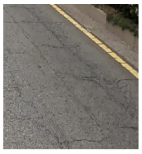
4. Experiments

4.1. Dataset and Preprocessing

To evaluate the effectiveness of the proposed method, we utilized the PPDD dataset [33], a road crack dataset collected from port pavements characterized by high traffic volume and heavy load. The dataset contains a total of 204,839 images, each accompanied by polygon-based annotations, and includes six distinct crack categories. The six types of cracks are as follows: reflective crack (RC), longitudinal and edge crack (LEC), corrugation,

shoving, and slippage crack (CSSC), rutting and depression crack (RDC), construction joint crack (CJC), alligator crack (AC). The number of annotated instances and visual examples for each crack category are summarized in Table 1. The representative images for each crack type were cropped from original images in the dataset. Among the six classes, RDC and CSSC are used only in the zero-shot evaluation setting as novel classes to assess the generalization ability of the proposed method.

Table 1. Data image and data distribution of PPDD dataset.

	Reflective Crack	**Longitudinal Edge Crack**	**Corrugation Shoving Slippage Crack**	**Rutting Depression Crack**	**Construction Joint Crack**	**Alligator Crack**
Visual pattern						
Number of crack instances	124,780	183,094	20,210	68,434	109,625	88,571

The original images with a resolution of 3840 × 2160 were resized to 640 × 640 for input into the model. This resolution matches the standard input size commonly used in existing object detection models and was chosen to strike a balance between computational efficiency and detection accuracy. This resizing and bounding-box conversion process followed the PPDD dataset protocol [33]. Furthermore, the polygon-based annotations originally intended for segmentation tasks were converted into bounding box format to suit our proposed detection-based framework. To align class names and attribute sets in the vision–language model (VLM), we employed a generative model to define visual attributes for each crack type. These attributes were categorized into three dimensions: geometry, spatial distribution, and texture. The attributes were constructed and validated by domain experts, and the complete attribute definitions are summarized in Table 2. The structure and annotation protocol follow the same format used in prior OVAD studies. To construct attribute annotations, we adopted a semi-automated procedure combining generative modeling and expert validation. First, domain experts described typical causal factors and visual characteristics for each crack category. These descriptions were provided as prompts to a generative language model, which produced a set of candidate attribute expressions. The candidates were subsequently reviewed and refined by experts, ensuring domain relevance and linguistic clarity. Through iterative filtering, a final set of 18 attributes was consolidated. This process balances the creativity and breadth of generative models with the domain knowledge necessary for reproducibility.

While Tables 1 and 2 summarize the categories and visual attributes of cracks, it should be noted that road cracks exhibit a wide spectrum of patterns depending on their mechanical causes, material properties, and environmental conditions. Capturing this diversity with only a few representative images would be insufficient, and a full taxonomy is beyond the scope of this work. Instead, we provide references to standard crack morphology guides [14,33,34] for readers seeking detailed visual illustrations. In this study, the tables are intended to clarify the dataset design for readers from both road engineering and computer vision backgrounds.

Table 2. Attribute annotation table for crack categories. Each row lists a crack type, its characteristic geometric, textural, and spatial descriptors, and the corresponding class code.

Crack	Attribute of Crack	Category (Class)
Reflective crack	"geometry: thin and long", "geometry: perpendicular to the lane direction"	RC
Longitudinal crack	"geometry: thin and long", "geometry: parallel to the lane direction", "spatial: pavement edge"	LEC
Edge crack		
Corrugation	"geometry: perpendicular to the lane direction", "texture: wave-like", "geometry: crescent-shape", "geometry: curved"	CSSC
Shoving		
Slippage crack		
Rutting	"geometry: depressed", "geometry: being lower", "texture: widely fragmented and finely broken", "spatial: aligned with wheel paths"	RDC
Depression		
Construction joint crack	"geometry: straight", "geometry: thin and long", "geometry: parallel to the lane direction", "spatial: aligned with lane markings"	CJC
Alligator crack	"geometry: polygonal", "texture: finely cracked", "texture: alligator skin", "texture: finely fragmented"	AC

4.2. Experimental

4.2.1. Experimental Details

In this study, we implemented the attribute-aware open-vocabulary crack detection (AOVCD) framework by extending the methodology of OVAD in a ViLD-style OVD recipe [12]. Our detection model is based on the Faster R-CNN architecture with a ResNet-50 backbone pretrained on ImageNet, comprising approximately 41 million parameters. For the language encoder, we adopted the CLIP-ViT-B/32 model, which was kept frozen during training to preserve the general-purpose semantic alignment capabilities learned through large-scale contrastive pretraining. Our study focuses on analyzing the effect of attribute alignment in the open-vocabulary detection setting for the road crack domain. To isolate and highlight the contribution of our alignment strategy, we intentionally employed a conventional architecture.

We report object detection performance using the evaluation metrics provided by the COCO evaluation protocol implemented in Detectron2. Specifically, average precision (AP) is computed across multiple Intersection-over-Union (IoU) thresholds ranging from 0.50 to 0.95 with a step size of 0.05, and the mean of these values is denoted as mAP. For reference, we also report AP_{50}, which corresponds to AP at a single IoU threshold of 0.50 (PASCAL VOC style). Unless otherwise stated, the reported values are averaged over all categories.

We adopted a two-branch training strategy to align visual and textual embeddings in a shared semantic space. Specifically, the visual region proposals generated by the detector were aligned with corresponding attribute-level text embeddings using cosine similarity. During training, we supervised alignment using base class labels annotated with composite visual attributes. This allowed the model to learn attribute-level representations while preserving generalization capacity for unseen (novel) defect types. For training, we used the PyTorch (version 2.1.0) framework along with the scikit-learn (sklearn) library for dataset splitting and standard normalization. The dataset was split into a base class-only training set and a test set containing both base and novel classes (i.e., base ∪ novel). The proportion of base-class samples for training was fixed at 80%, with the remaining 20%

reserved for evaluation. All experiments were conducted with a batch size of 16, and the initial learning rate was set to 0.01 for the main training loop. To ensure stable convergence, we employed the RAdam optimizer [35] and cosine annealing with warm restarts [36] for learning rate scheduling. Warm-up training was performed with an initial learning rate of 0, maximum learning rate (η_max) of 0.01, first cycle length T_0 = 20, and multiplier T_mult = 1. The model was trained for 100 epochs on a single NVIDIA RTX 3080 GPU, and training time per epoch averaged approximately 6 min, with inference taking less than 0.5 s per image on average. The RPN used anchor sizes 32, 64, 128, 256, 512 with aspect ratios 0.5, 1.0, 2.0, and NMS IoU 0.7. During training, we kept the top 2000 proposals before NMS and 1000 after. ROI label assignment used an IoU threshold of 0.5. Training augmentation applied horizontal flip with probability 0.5 and no rotation or color jitter was used. At inference, we used a score threshold of 0.05, class-wise NMS with IoU 0.5, and retained at most 100 detections per image. All other settings followed Detectron2 defaults.

Our method enables zero-shot generalization to novel crack types through visual-attribute alignment in the shared embedding space, rather than relying solely on fixed class labels. Empirical evaluations confirm the superiority of our model in detecting unseen classes when compared to traditional class-based baselines.

4.2.2. Code Availability

Our code is available at https://github.com/LeaYoon/AOVCD (accessed on 7 September 2025).

5. Results

5.1. Comparison of Baseline Models by Attribute Aligned

In Table 3, the Baseline refers to the zero-shot performance of a pretrained CLIP model evaluated on the PPDD test set. CLIP-ViT-B32 (category) represents the performance of the model fine-tuned using only the base class category names. CLIP-ViT-B32 (category + attribute) denotes the performance of the model fine-tuned with both base class category names and the attribute information defined in Table 2. All results are evaluated using the mAP and AP_{50} metric. The Baseline corresponds to the zero-shot performance of the pretrained CLIP model without any additional supervision, yielding negligible accuracy on both base (0.14%) and novel (1.21%) classes and an overall performance of just 0.51%. This result indicates that generic CLIP embeddings are insufficient for road crack detection without domain-specific adaptation. Fine-tuning CLIP using only the base class category names (CLIP-ViT-B32 (category)) leads to a substantial improvement, achieving 68.02% AP_{50} on base classes and 3.24% on novel classes. However, generalization to unseen categories remains limited due to the lack of semantic diversity in the supervision. When both category and attribute information are used during fine-tuning (CLIP-ViT-B32 (category + attribute), AOVCD), the model achieves further gains: 70.98% AP_{50} on base classes and a significantly higher 13.83% AP_{50} on novel classes, resulting in the best total performance of 42.40%. This demonstrates that aligning visual features with structured attribute descriptions enhances the model's ability to detect novel crack types, suggesting improved semantic generalization through attribute-aware supervision. In addition to the numerical results, Figure 3 presents the macro-averaged Precision–Recall (PR) curves for the OVCD and AOVCD at IoU = 0.50. The PR curves clearly illustrate that while the baseline model suffers from rapidly declining precision as recall increases, the attribute-aligned model maintains higher precision across a wider recall range. This visualization further supports the advantage of attribute-aware supervision in improving detection robustness and generalization. Although the absolute AP values for novel categories remain modest, the relative improvements over the baseline are substantial (from 3.24% to 13.83%). These

results confirm the feasibility of transferring open-vocabulary detection into the road crack domain, where pretrained vision–language models have limited prior exposure. By contrast, the smaller gains on base categories reflect the fact that these categories are already well covered in the training data.

Table 3. Comparison of baseline models by attribute alignment.

	Base Class	Novel Class	Total
Baseline (CLIP zero-shot)	0.05/0.14	0.74/1.21	0.12/0.51
CLIP-ViT-B32 (category)	38.28/68.02	2.20/3.24	17.69/35.63
CLIP-ViT-B32 (category + attribute)	40.52/70.98	6.27/13.83	22.32/42.40

Comparison of attribute alignment effects across crack categories. Results are reported as mAP and average precision at 50% intersection-over-union threshold (AP_{50}, %). Columns represent performance on base classes (RC, LEC, CJC, AC), novel classes (CSSC, RDC), and total classes.

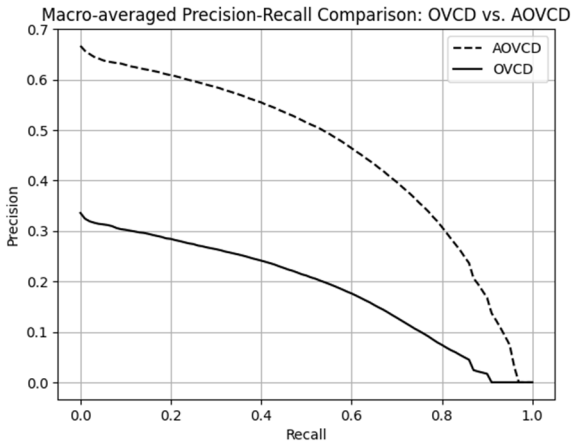


Figure 3. Comparative PR curves at IoU = 0.50. Macro-averaged curves for OVCD vs. AOVCD.

5.2. Performance Comparison by Crack Type

Table 4 reports the performance comparison of crack detection across both base and novel classes under different attribute alignment settings. The baseline model, which directly applies CLIP in a zero-shot setting without attribute guidance, shows marginal detection performance across all categories, with AP_{50} values below 0.2 for base classes and negligible results for novel classes (e.g., 0.11% for RC, 0.0% for AC, and 2.17% for RDC). This highlights the limitation of CLIP's general-purpose vision–language alignment when applied to fine-grained crack types. The OVCD model, which leverages object-level vision–language alignment without explicit attribute supervision, demonstrates significant improvement, particularly in base classes such as AC (76.30%) and CJC (67.78%). However, its performance on novel categories remains limited, showing low AP_{50} values for CSSC (1.98%) and RDC (4.03%), indicating insufficient generalization to unseen crack types. In contrast, the proposed AOVCD framework achieves the highest performance across all categories. Notably, it improves novel class detection by a substantial margin, achieving 15.56% AP_{50} for CSSC and 10.83% for RDC, corresponding to a 4 to 8-fold improvement over OVCD. For base classes as well, AOVCD consistently outperforms OVCD, with AP_{50} gains observed in all four categories (e.g., +5.46% for RC and +4.86% for LEC). These results demonstrate that incorporating attribute-aware alignment facilitates better semantic grounding and generalization, especially for visually ambiguous or previously unseen crack types.

Table 4. Comparison of attribute alignment effect across crack categories.

	Base Class				Novel Class	
	RC	LEC	CJC	AC	CSSC	RDC
Baseline (CLIP zero-shot)	0.11	0.12	0.16	0.0	0.16	2.17
OVCD	40.87	47.20	67.78	76.30	1.98	4.03
AOVCD	46.33	52.06	70.80	78.58	15.56	10.83

Comparison of attribute alignment effects across crack categories. Results are reported as average precision at 50% intersection-over-union threshold (AP_{50}, %). Columns represent performance on base classes (RC, LEC, CJC, AC) and novel classes (CSSC, RDC).

5.3. Ablation Study of Attribute Detection

Table 5 summarizes the attribute recognition performance under a box oracle setup, where the model is evaluated using ground-truth bounding boxes for each crack region. The task is framed as multi-label classification across three attribute types: geometry, spatial, and texture. As shown in Table 2, the complete attribute set comprises 11 geometry attributes, 2 spatial attributes, and 5 texture attributes. The baseline CLIP zero-shot model shows negligible performance, with total balanced accuracy (BACC) of only 0.11% and average precision (AP) values consistently below 0.2%, indicating poor alignment between general CLIP representations and the fine-grained crack attributes. Compared to the baseline, the OVCD model yields minor improvements across all metrics, achieving 1.12% total BACC and modest increases in both geometric and texture attribute precision (e.g., 1.08% AP for geometry, 2.20% AP for texture). However, the performance remains insufficient for practical use. In contrast, the proposed AOVCD framework achieves substantial improvements in attribute recognition. Specifically, it records a total BACC of 41.04%, with the highest gains observed in texture-related attributes (45.01% AP, 35.54% BACC). Geometry and spatial attributes also see strong performance boosts, with AOVCD reaching 25.73% AP in geometry and 25.59% BACC in spatial attributes. Notably, the AP for geometry improves from 0.09% (baseline) to 25.73%, and BACC for spatial attributes rises from 0.03% to 25.59%. These results validate that attribute-aware supervision enables more discriminative feature learning across multiple visual aspects of cracks, supporting more reliable alignment between visual input and semantic descriptions.

Table 5. Attribute recognition performance.

	Total		Geometry		Spatial		Texture	
	BACC	AP	BACC	AP	BACC	AP	BACC	AP
Baseline(CLIP zero-shot)	0.11	0.09	0.19	0.03	0.04	0.15	0.19	0.03
OVCD	1.12	1.08	0.21	0.05	0.13	0.24	2.23	2.20
AOVCD	41.04	25.73	38.74	25.59	45.01	35.54	43.26	20.12

The performance is reported in percentage (%). BACC: balanced accuracy, AP: average precision.

Table 6 presents an ablation study designed to evaluate how different attribute set sizes and their combinations affect model performance. The attributes are grouped into three categories—geometric, spatial, and texture—and progressively accumulated to form larger attribute sets, allowing us to assess their individual and joint contribution. When only geometric attributes are considered, the model achieves 68.94 on the base set, 10.07 on the novel set, and 40.56 in total, which is the lowest overall performance across all settings. Adding spatial attributes (geometry + spatial) produces a very limited change, with scores of 69.11, 9.89, and 40.59, respectively. However, when all three categories (geometry + spatial + texture) are included, the performance increases significantly to 70.98, 13.83, and 42.40, respectively, the best result observed in this study. These results

indicate that geometric attributes alone are insufficient; the addition of spatial attributes does not yield a meaningful benefit. In contrast, incorporating texture information leads to a clear improvement, particularly for novel cases. This trend highlights the complementary role of diverse attribute categories, suggesting that a richer attribute set provides stronger generalization and overall performance.

Table 6. Ablation study on the impact of attribute set size and combination in terms of mAP@50.

Geometry	Spatial	Texture	Base	Novel	Total
O	X	X	68.94	10.07	40.56
O	O	X	69.11	9.89	40.59
O	O	O	70.98	13.83	42.40

The performance is reported in percentage (%). O and X indicate whether the attribute set in the corresponding column was used (O) or not used (X) in the ablation study.

5.4. Qualitative Evaluation by Attribute Alignment

We conducted a qualitative evaluation of the proposed baseline method to assess its ability to identify novel categories through attribute alignment. Figure 4 illustrates that our method is capable of predicting previously unseen classes, such as CSSC and RDC, despite not being exposed to them during training. It also demonstrates that the method successfully performs predictions for categories it encountered during the training phase. Among the notable failure cases observed in our experiments, most misclassifications occurred when novel categories shared highly similar textures with known categories, leading to incorrect label assignments. However, even in such challenging scenarios, the method was generally able to preserve bounding boxes through objectness cues. We address this limitation and propose future research directions to overcome it in the Discussion section.

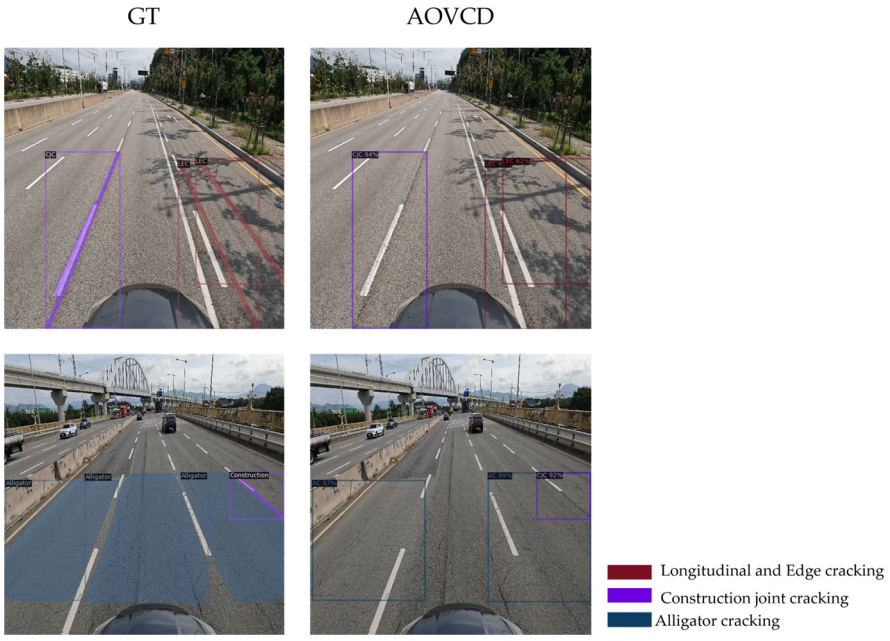


Figure 4. Examples of AOVCD on PPDD test set. AOVCD can recognize seen crack categories while also generalizing to novel categories by virtue of class-agnostic classification head.

Figure 5 illustrates typical failure cases of AOVCD by comparing ground truth with two prediction settings. The left column highlights the confusion between LECs and CJCs. In particular, cracks occurring close to lane markings are often misclassified as CJCs, reflecting the fact that typical construction requirements specify longitudinal joints to be aligned parallel to the roadway centerline and frequently positioned at lane markings

in multi-lane facilities. This domain-specific context causes the model to attend to lane entities when learning CJCs, which in turn leads to false positives for nearby longitudinal cracks. In contrast, when no lane marking is present (Row 2), the model can correctly classify CJCs, demonstrating both the benefit and the risk of attribute alignment with lane-related cues. The right column presents novel class confusion cases. For instance, RDCs are either missed at low confidence or misclassified as CJCs when lane markings dominate the context (Row 1). Detecting RDCs requires texture-level understanding rather than purely geometric cues, but the model appears biased toward contextual priors such as lane features. Similarly, CSSCs are misclassified as LECs (Row 2). Since CSSCs reflect fatigue-induced surface settlement that manifests primarily through repetitive texture patterns, its distinction cannot be reliably inferred from attributes of base categories, which emphasize geometric elongation. These results reveal that pretrained features insufficiently capture domain-specific texture phenomena, underscoring the need for improved structure- and texture-aware detection.

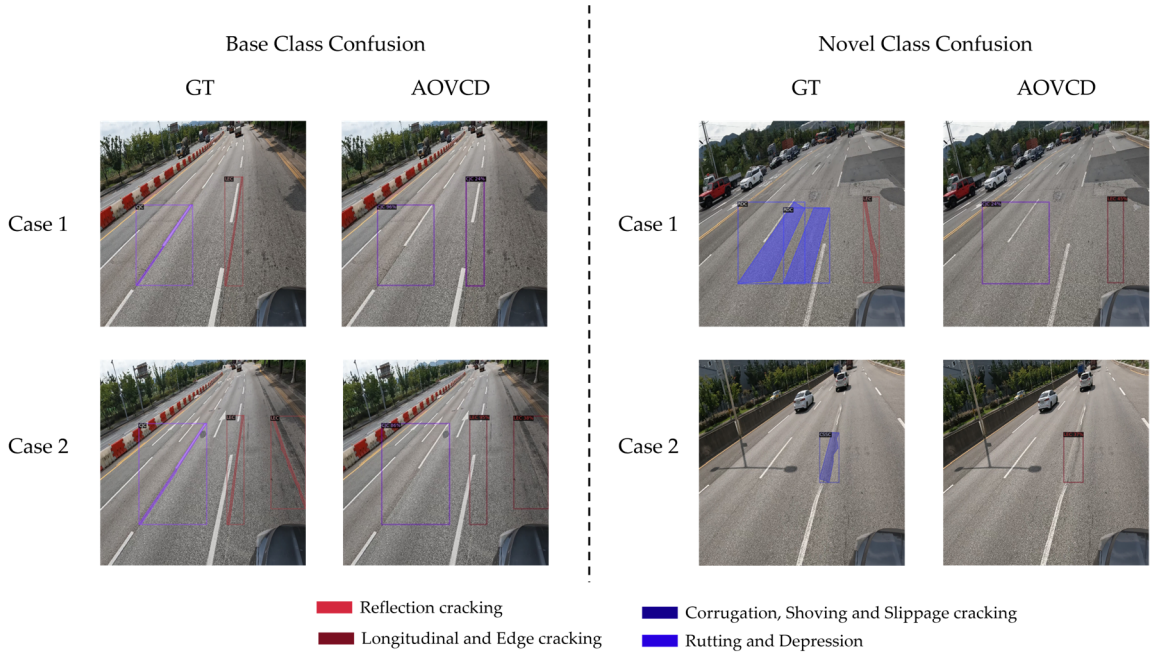


Figure 5. Failure cases of AOVCD. Each row shows representative examples with the input image with ground truth (**left**) and our AOVCD (**right**). Left column: misclassification between longitudinal cracks (LECs) and construction joint crack (CJC). (Row 1) Longitudinal cracks near lane markings are mistaken as CJC due to lane-aligned joint priors. (Row 2) Correct recognition of CJC when lane marking is weakly present. Right column: novel class confusion. (Row 1) Rutting and depression cracks (RDCs) are missed or misclassified as CJC under lane-biased context. (Row 2) Corrugation cracks (CSSCs) are misclassified as LEC due to the lack of texture attribute perception.

6. Discussion

6.1. Road Crack Detection and Segmentation Dataset

While Tables 3–6 provide comparisons with the OVD baseline, it is important to acknowledge that traditional object detectors or segmentation models such as YOLOv9 [37], CrackFormer [4], and EDNet [20] have reported strong performance on crack detection benchmarks. Recent studies on road crack analysis have explored downstream tasks such as crack segmentation. Representative works include DeepCrack [38] and CrackFormer, which were evaluated on patch-level crack datasets [39,40]. Unlike category-based detection, these approaches aim to generate masks that delineate the fractured regions of cracks. Such methods have reported average precision scores ranging from 0.875 to 0.896. The EDNet

model further addressed the severe imbalance between crack and non-crack pixels by proposing a segmentation method tailored to this challenge, achieving a reported F1-score of 97.80. Meanwhile, the crack research community has organized competitions to advance multi-category crack detection models. A representative example is GRDDC 2020 [41], where participants competed to detect crack instances of four categories in vehicle-view images [14]. To account for class imbalance and jointly emphasize recall and precision, the evaluation was based on the F1-score with IoU > 0.5. The reported performance ranged from 0.5368 to 0.6748 for the top ten teams, indicating competitive yet still moderate results. These observations confirm that conventional detectors and segmenters deliver high accuracy on fixed label sets but cannot recognize novel or unseen crack categories without retraining.

Open-vocabulary object detection (OVD) has rapidly advanced since its formal introduction by Zareian et al. [9]. Early methods relied on image–caption pairs to weakly supervise detection of novel objects. ViLD [29] provided the first substantial gains: ViLD with a ResNet-50 backbone achieved 16.1 mask AP on LVIS rare classes and 22.5 mask AP overall, and using a stronger teacher (ALIGN) increased rare-class performance to 26.3 AP. Detic [30] further expanded the vocabulary by training the classifier on image-level data; it reported a gain of 8.3 mAP on novel classes and reached 41.7 mask mAP and 41.7 mAP on rare LVIS classes using a Swin-B backbone. More recently, VLDet [11] formulated object language alignment as a set-matching problem and reported 32.0 AP on the open-vocabulary COCO benchmark and 21.7 mask AP on LVIS novel categories. These models achieved high longitudinal joint construction solute performance by leveraging large image–text corpora and by using strong backbones and multi-scale training; however, they are evaluated on web-centric datasets with natural objects and backgrounds. Although our absolute novel-class AP_{50} (13.83) remains below the 21.7–32.0 mAP reported for generic OVD models, this gap primarily reflects the difficulty of the road-defect domain (fine cracks, occlusions, poor lighting) and the small number of training classes. Importantly, the attribute-level cues—such as crack width, orientation and shape—enable the detector to discriminate unseen defects far better than category-only alignment. Moreover, the overall AP_{50} of 42.40 is on par with Detic's baseline performance on general datasets. These results demonstrate that adapting OVD techniques to domain-specific imagery with attribute-level supervision can close much of the generalization gap, even when the underlying backbone and training data are modest.

Open-vocabulary attribute detection (OVAD) [12] is a newer extension that our work also relates to. Instead of only predicting object classes, OVAD requires identifying object attributes (properties like color, material, or state) in a zero-shot manner. The first OVAD benchmark augmented COCO with 117 attribute classes to enable rigorous evaluation. Baseline methods (e.g., CLIP and OpenCLIP [31]) detect attributes by measuring similarity between region visuals and text prompts for each attribute or synonym. These baselines show strong results on salient attributes tied to an object's identity (for instance, predicting an object's color or general size), but their performance drops on more subtle attributes like material or object state (e.g., wooden vs. metallic, or open vs. closed state) that are not as explicitly captured by image–text pretraining. Future work could explore stronger backbones and richer attribute vocabularies to further narrow the gap with mainstream OVD benchmarks while maintaining applicability to infrastructure inspection. These findings suggest a broader implication: open-world vision models augmented with attribute recognition can more effectively support real-world decision-making. In our case, detecting a crack and recognizing it as an "edge crack approximately 2 m long" versus a "transverse crack across a lane" directly informs distinct maintenance strategies. Such attribute-aware detection represents the next evolution of vision–language models—one that aligns with the current

research trajectory from image-level alignment towards region-level and attribute-level alignment for comprehensive scene understanding.

6.2. Limitation and Future Work

While the proposed AOVCD framework demonstrates promising results, several limitations present opportunities for improvement. First, the current attribute design can be enhanced to better characterize cracks. Incorporating a richer set of crack attributes (e.g., finer distinctions in width, depth, orientation, or severity) and ensuring high-quality annotations for these attributes would provide more informative guidance to the model [42]. By capturing more nuanced crack descriptors, the detector can more accurately differentiate crack types and conditions, ultimately improving detection accuracy. Nevertheless, the current design still depends on a fixed attribute bank and prompt templates, which may require re-tuning when applied to new road environments. Free-form prompts can also introduce ambiguity, as experts may describe the same crack differently (e.g., "wave-like", "undulating"). While we mitigated this by adopting standardized descriptors and learned templates, ambiguity cannot be fully removed. Although the absolute, we note that the use of discrete attribute prompts may constrain expressivity. Similar to recent findings in NLP regarding the limitations of hard prompt formulations [43], exploring soft or continuous prompt learning strategies in the context of open-vocabulary detection represents an important direction for future research. Second, advanced training strategies should be explored to boost generalization and robustness. For instance, a multi-task learning approach that jointly learns crack detection and attribute classification can enrich the model's feature representations. This has been shown to leverage the semantic characteristics of cracks better and significantly improve precision in crack detection [44]. Similarly, incorporating hard example mining or curriculum learning could help the model handle difficult cases more effectively. Extending this paradigm to a joint learning of numerical indicators (e.g., crack width, depth) with language-based attributes could further improve robustness and expressivity. Such a multi-modal training approach may alleviate the limitations of relying solely on linguistic descriptions. In addition, the AOVCD dataset itself has intrinsic limitations. Annotations are restricted to visually observable cracks within close view (approximately 1000 pixels) and do not cover distant or environment-specific conditions (e.g., night, wet, or back-light). Consequently, scale-stratified or condition-specific evaluation metrics are not directly applicable in this domain. Addressing such gaps would require new data collection protocols rather than model-level modifications. Third, data augmentation and expansion of the training data are crucial for improving generalization. The AOVCD model's performance may be limited by data scarcity or imbalance (e.g., under-represented crack types or rare attributes), which can hinder generalization [45]. Future work should apply diverse augmentation strategies, including synthetic crack generation, photometric augmentations (lighting, noise, weather), and geometric transformations, to increase data diversity and reduce overfitting [46]. They also help the model become more robust to noise and environmental variations commonly encountered in real-world road surfaces. Finally, optimizing the model architecture and inference strategy can further boost performance. One direction is to improve multi-scale feature learning within the detector—effectively capturing both fine, hairline cracks and larger crack patterns. Integrating advanced feature fusion mechanisms (e.g., a bi-directional feature pyramid or attention modules) can enhance the detection of small or low-contrast cracks under challenging conditions [47]. This would address the current difficulty in detecting very thin or morphologically complex cracks, thereby improving overall recall and accuracy. Future work should also explore mapping detected cracks to standardized severity grades, enabling the results to be directly interpretable for maintenance decisions. Additionally, model refinement through

techniques like model compression or lightweight backbone design would be valuable for practical deployments. This can reduce computational complexity and memory usage without sacrificing accuracy, enabling the crack detection model to run on resource-constrained devices in real time. In summary, by pursuing richer attribute modeling, better training regimes, augmented data diversity, and architectural optimizations, future versions of the AOVCD framework can achieve higher detection accuracy, stronger generalization across diverse road conditions, and improved robustness against noise and environmental variability. While the AOVCD framework proposed in this study demonstrated promising performance, several challenges remain to be addressed. The attribute inventory reflects design choices, and alternative inventories may change results. Accordingly, we plan to conduct follow-up research to overcome the identified limitations and further enhance the model's practicality and performance. These efforts will focus on refining attribute design, introducing advanced training strategies, expanding data diversity, and optimizing model architecture, thereby contributing to the development of more accurate and robust open-vocabulary crack detection technology.

7. Conclusions

This study proposed the attribute-aware open-vocabulary crack detection (AOVCD) framework to address the limitations of closed-set object detectors in road defect detection and demonstrated its potential effectiveness. The proposed approach redefines crack categories as combinations of visual attributes and aligns them with pretrained language embeddings, thereby enabling zero-shot generalization to previously unseen defect types. To support this framework, we extended an existing road defect dataset with structured attribute annotations and incorporated a multi-label attribute classification task to enhance semantic understanding. Experimental results show that AOVCD significantly outperforms both the CLIP-based zero-shot baseline and conventional fine-tuned models. In particular, for novel class detection, AOVCD achieves an 11.66% improvement in average precision (AP) compared to the OVCD baseline. Additionally, attribute recognition across geometric, spatial, and textural dimensions improved by more than 30% in terms of balanced accuracy (BACC). These findings empirically demonstrate that attribute-aware alignment effectively enhances both the model's generalization ability and its capacity for fine-grained visual discrimination. However, several challenges remain. The current attribute schema, while functional, requires refinement to capture more nuanced crack characteristics such as width, depth, and severity. Advanced training strategies—such as attribute segmentation or curriculum learning—could further improve the model's generalization. Limited data diversity also poses a constraint; thus, approaches such as GAN-based crack synthesis or domain-specific augmentation techniques are needed to address the sparsity of rare defect patterns. Moreover, enhancing computational efficiency through lightweight backbone designs or improved multi-scale feature fusion could enable practical deployment in resource-constrained environments, such as mobile inspection systems. In summary, this study presents a flexible and scalable framework for open-vocabulary road defect detection and lays the groundwork for building robust and extensible maintenance systems that can withstand real-world environmental variations.

Author Contributions: Conceptualization, H.Y. and S.K.; Methodology, H.Y.; Software, H.Y.; Validation, H.Y. and S.K.; Data Curation, H.Y.; Writing—Original Draft Preparation, H.Y.; Writing—Review and Editing, S.K.; Visualization, H.Y.; Supervision, S.K.; Funding Acquisition, S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Dong-A University, Republic of Korea (10.13039/501100002468).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available at https://github.com/LeaYoon/AOVCD (accessed on 7 September 2025).

Conflicts of Interest: The authors declare no competing interests.

References

1. Naddaf-Sh, S.; Naddaf-Sh, M.-M.; Kashani, A.R.; Zargarzadeh, H. An efficient and scalable deep learning approach for road damage detection. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 5602–5608.
2. Li, S.; Zhang, D. Deep Learning-Based Algorithm for Road Defect Detection. *Sensors* **2025**, *25*, 1287. [CrossRef] [PubMed]
3. Wang, J.; Meng, R.; Huang, Y.; Zhou, L.; Huo, L.; Qiao, Z.; Niu, C. Road defect detection based on improved YOLOv8s model. *Sci. Rep.* **2024**, *14*, 16758. [CrossRef]
4. Liu, H.; Miao, X.; Mertz, C.; Xu, C.; Kong, H. Crackformer: Transformer network for fine-grained crack detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 3783–3792.
5. Yang, Y.; Niu, Z.; Su, L.; Xu, W.; Wang, Y. Multi-scale feature fusion for pavement crack detection based on Transformer. *Math. Biosci. Eng.* **2023**, *20*, 14920–14937. [CrossRef]
6. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 8748–8763.
7. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the COMPUTER vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part v 13. pp. 740–755.
8. Gupta, A.; Dollar, P.; Girshick, R. Lvis: A dataset for large vocabulary instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5356–5364.
9. Zareian, A.; Rosa, K.D.; Hu, D.H.; Chang, S.-F. Open-vocabulary object detection using captions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14393–14402.
10. Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L.H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y. Regionclip: Region-based language-image pretraining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16793–16803.
11. Lin, C.; Sun, P.; Jiang, Y.; Luo, P.; Qu, L.; Haffari, G.; Yuan, Z.; Cai, J. Learning object-language alignments for open-vocabulary object detection. *arXiv* **2022**, arXiv:2211.14843.
12. Bravo, M.A.; Mittal, S.; Ging, S.; Brox, T. Open-vocabulary attribute detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7041–7050.
13. Thompson, E.M.; Ranieri, A.; Biasotti, S.; Chicchon, M.; Sipiran, I.; Pham, M.-K.; Nguyen-Ho, T.-L.; Nguyen, H.-D.; Tran, M.-T. SHREC 2022: Pothole and crack detection in the road pavement using images and RGB-D data. *Comput. Graph.* **2022**, *107*, 161–171. [CrossRef]
14. Arya, D.; Maeda, H.; Ghosh, S.K.; Toshniwal, D.; Sekimoto, Y. RDD2020: An annotated image dataset for automatic road damage detection using deep learning. *Data Brief* **2021**, *36*, 107133. [CrossRef] [PubMed]
15. Mandal, V.; Mussah, A.R.; Adu-Gyamfi, Y. Deep learning frameworks for pavement distress classification: A comparative analysis. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 5577–5583.
16. Zhao, C.; Shu, X.; Yan, X.; Zuo, X.; Zhu, F. RDD-YOLO: A modified YOLO for detection of steel surface defects. *Measurement* **2023**, *214*, 112776. [CrossRef]
17. Yang, F.; Zhang, L.; Yu, S.; Prokhorov, D.; Mei, X.; Ling, H. Feature pyramid and hierarchical boosting network for pavement crack detection. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 1525–1535. [CrossRef]
18. Eisenbach, M.; Stricker, R.; Seichter, D.; Amende, K.; Debes, K.; Sesselmann, M.; Ebersbach, D.; Stoeckert, U.; Gross, H.-M. How to get pavement distress detection ready for deep learning? A systematic approach. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; pp. 2039–2047.
19. Choi, W.; Cha, Y.-J. SDDNet: Real-time crack segmentation. *IEEE Trans. Ind. Electron.* **2019**, *67*, 8016–8025. [CrossRef]
20. Tang, Y.; Zhang, A.A.; Luo, L.; Wang, G.; Yang, E. Pixel-level pavement crack segmentation with encoder-decoder network. *Measurement* **2021**, *184*, 109914.
21. Guo, J.-M.; Markoni, H.; Lee, J.-D. BARNet: Boundary aware refinement network for crack detection. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 7343–7358. [CrossRef]

22. Qu, Z.; Cao, C.; Liu, L.; Zhou, D.-Y. A deeply supervised convolutional neural network for pavement crack detection with multiscale feature fusion. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 4890–4899.
23. Arya, D.; Maeda, H.; Ghosh, S.K.; Toshniwal, D.; Sekimoto, Y. RDD2022: A multi-national image dataset for automatic road damage detection. *Geosci. Data J.* **2024**, *11*, 846–862.
24. Kortmann, F.; Talits, K.; Fassmeyer, P.; Warnecke, A.; Meier, N.; Heger, J.; Drews, P.; Funk, B. Detecting various road damage types in global countries utilizing faster R-CNN. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 5563–5571.
25. Zeng, J.; Zhong, H. YOLOv8-PD: An improved road damage detection algorithm based on YOLOv8n model. *Sci. Rep.* **2024**, *14*, 12052.
26. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 4904–4916.
27. Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; Carion, N. Mdetr-modulated detection for end-to-end multi-modal understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 1780–1790.
28. Li, L.H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N. Grounded language-image pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10965–10975.
29. Gu, X.; Lin, T.-Y.; Kuo, W.; Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv* **2021**, arXiv:2104.13921.
30. Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; Misra, I. Detecting twenty-thousand classes using image-level supervision. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 350–368.
31. Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; Jitsev, J. Reproducible scaling laws for contrastive language-image learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 2818–2829.
32. Chen, K.; Jiang, X.; Hu, Y.; Tang, X.; Gao, Y.; Chen, J.; Xie, W. Ovarnet: Towards open-vocabulary object attribute recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 23518–23527.
33. Yoon, H.; Kim, H.-K.; Kim, S. PPDD: Egocentric Crack Segmentation in the Port Pavement with Deep Learning-Based Methods. *Appl. Sci.* **2025**, *15*, 5446. [CrossRef]
34. Miller, J.S.; Bellinger, W.Y. *Distress Identification Manual for the Long-Term Pavement Performance Program*; Department of Transportation, Federal Highway Administration: Washington, DC, USA, 2003.
35. Liu, L.; Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; Han, J. On the variance of the adaptive learning rate and beyond. *arXiv* **2019**, arXiv:1908.03265.
36. Loshchilov, I.; Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv* **2016**, arXiv:1608.03983.
37. Yaseen, M. What is YOLOv9: An in-depth exploration of the internal features of the next-generation object detector. *arXiv* **2024**, arXiv:2409.07813.
38. Liu, Y.; Yao, J.; Lu, X.; Xie, R.; Li, L. DeepCrack: A deep hierarchical feature learning architecture for crack segmentation. *Neurocomputing* **2019**, *338*, 139–153. [CrossRef]
39. Zou, Q.; Cao, Y.; Li, Q.; Mao, Q.; Wang, S. CrackTree: Automatic crack detection from pavement images. *Pattern Recognit. Lett.* **2012**, *33*, 227–238. [CrossRef]
40. Zou, Q.; Zhang, Z.; Li, Q.; Qi, X.; Wang, Q.; Wang, S. Deepcrack: Learning hierarchical convolutional features for crack detection. *IEEE Trans. Image Process.* **2018**, *28*, 1498–1512. [CrossRef]
41. Arya, D.; Maeda, H.; Ghosh, S.K.; Toshniwal, D.; Omata, H.; Kashiyama, T.; Sekimoto, Y. Global road damage detection: State-of-the-art solutions. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 5533–5539.
42. Gui, R.; Xu, X.; Zhang, D.; Pu, F. Object-based crack detection and attribute extraction from laser-scanning 3D profile data. *IEEE Access* **2019**, *7*, 172728–172743.
43. Li, X.L.; Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv* **2021**, arXiv:2101.00190. [CrossRef]
44. Wang, H.; Du, W.; Xu, G.; Sun, Y.; Shen, H. Automated crack detection of train rivets using fluorescent magnetic particle inspection and instance segmentation. *Sci. Rep.* **2024**, *14*, 10666. [CrossRef] [PubMed]
45. El-Din Hemdan, E.; Al-Atroush, M. A review study of intelligent road crack detection: Algorithms and systems. *Int. J. Pavement Res. Technol.* **2025**, 1–31. [CrossRef]

46. Kim, J.; Seon, J.; Kim, S.; Sun, Y.; Lee, S.; Kim, J.; Hwang, B.; Kim, J. Generative AI-driven data augmentation for crack detection in physical structures. *Electronics* **2024**, *13*, 3905. [CrossRef]
47. Wang, Y.; Zhu, H.; Wang, Y.; Liu, J.; Xie, J.; Zhao, B.; Zhao, S. GSBYOLO: A lightweight Multi-Scale fusion network for road crack detection in complex environments. *Sci. Rep.* **2025**, *15*, 26615. [CrossRef]