Engineering Applications of Artificial Intelligence 182 (2026) 115851

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

A two-stage detection and segmentation framework for pedestrian crosswalk inventory and condition assessment from aerial imagery

Xintong Yan [a], Zubin Bhuyan [a], Jimi Oke [b], Guanhe Wu [a], Yuanchang Xie [a,*]

[a] Department of Civil and Environmental Engineering, University of Massachusetts Lowell, 1 University Avenue, Lowell, MA, 01854, USA
[b] Department of Civil and Environmental Engineering, University of Massachusetts Amherst, 130 Natural Resources Road, Amherst, MA, 01003, USA

ARTICLE INFO

Keywords:
Aerial imagery
Crosswalk detection
Safety
Semantic segmentation
Transportation asset management

ABSTRACT

Accurate crosswalk inventories are critical for supporting pedestrian and wheelchair-user safety and accessibility. Inclusion of detailed pavement marking data further enhances crosswalk condition assessment, lifecycle planning, and data-driven asset management. This study introduces a unified deep-learning-based artificial intelligence framework for large-scale crosswalk localization and condition assessment using high-resolution aerial imagery. The framework employs an innovative two-stage approach. The first stage is an oriented bounding box (OBB) module, which is an anchor-free detector designed to localize crosswalks at arbitrary orientations, providing highly accurate spatial positioning. This is followed by a segmentation stage, which incorporates a novel multi-scale attention (MSA) module. The MSA module is crucial for extracting pixel-level details of pavement markings, which are essential for subsequent crosswalk condition analysis. The outputs from both stages are then fed into a condition-assessment pipeline to calculate a normalized condition ratio. This ratio effectively quantifies crosswalk pavement marking degradation, offering a standardized metric across diverse crosswalk types. The efficacy of the proposed two-stage approach is rigorously validated on extensive and diverse datasets sourced from Massachusetts, Florida, and Washington. Performance metrics demonstrate superior capabilities of the proposed framework: the OBB detector achieved a mean average precision of 0.893 across intersection over union (IoU) thresholds from 0.50 to 0.95, and the segmentation model attained a mean intersection over union (mIoU) of 0.925. These results consistently outperform existing state-of-the-art baselines. Ultimately, the proposed framework successfully generates accurate, geographic information system (GIS)-ready vector crosswalk inventories, establishing a highly scalable and automated solution for advanced transportation infrastructure monitoring and proactive maintenance planning.

1. Introduction

Crosswalks are critical transportation assets that support pedestrian and wheelchair-user safety and accessibility (Al-haideri et al., 2025; Bian et al., 2020; Pantangi et al., 2021). A scalable and up-to-date crosswalk inventory is essential for safety analysis, asset management, and planning applications. However, traditional crosswalk inventory construction largely relies on manual efforts, including field surveys and manual mapping, which are time-consuming, labor-intensive, and difficult to scale across large geographic areas (Hosseini et al., 2023).

Recent advances in convolutional neural networks (CNNs), especially in object detection and semantic segmentation, have made automated extraction of transportation assets increasingly feasible (Guo et al., 2024; Han, 2025; Kong et al., 2022; Kortli et al., 2022; Li et al., 2024; Lu et al., 2024; Ning et al., 2022; Tümen and Ergen, 2020; Wang and Chen, 2024; Xiao et al., 2020; Yan et al., 2024; Zhang et al., 2022), offering a promising alternative for large-scale and cost-effective crosswalk inventory generation.

Existing studies have explored multiple imaging sources. Ground-level imagery, such as Google Street View (GSV) or dash-cam videos, provides rich visual details but suffers from limited spatial coverage, inconsistent perspectives, and unreliable geolocation (Kaufmann et al., 2018; Xing et al., 2019), making it ineffective for constructing consistent, city- or state-wide crosswalk inventories suitable for geographic information system (GIS) integration. Satellite imagery offers wide coverage and georeferencing capabilities. However, its typical spatial

* Corresponding author.
E-mail addresses: xintong_yan@uml.edu (X. Yan), Zubin_Bhuyan@uml.edu (Z. Bhuyan), jimi@umass.edu (J. Oke), Guanhe_Wu@student.uml.edu (G. Wu), yuanchang_xie@uml.edu (Y. Xie).

https://doi.org/10.1016/j.engappai.2026.115851
Received 29 January 2026; Received in revised form 13 July 2026; Accepted 30 July 2026
Available online 3 August 2026
0952-1976/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

resolution (30-50 cm per pixel) is insufficient for reliably capturing fine-grained pavement markings such as crosswalk bars, especially under occlusion, shadowing, or degradation (Luttrell et al., 2024).

High-resolution aerial orthoimagery provides a promising alternative by offering both sufficient spatial detail (e.g., 3-6 inch resolution) and accurate georeferencing. This makes it particularly suitable for large-scale, GIS-ready crosswalk inventory generation. Despite this potential, research leveraging aerial imagery for crosswalk detection and observable condition assessment remains limited.

From a methodological perspective, several challenges persist. First, most existing crosswalk detectors rely on horizontal bounding boxes (HBBs), which are inadequate for representing arbitrarily oriented crosswalks and often include irrelevant pavement markings (e.g., stop bars, lane lines, or shadows), thereby reducing localization accuracy and affecting downstream analysis. Second, existing approaches either focus on coarse localization or classification-based condition labeling (Bui et al., 2026; Xie et al., 2023; Yi et al., 2021), without capturing pixel-level degradation patterns necessary for fine-grained condition assessment. Third, semantic segmentation applied directly to full aerial images remains challenging due to small target size, visual similarity to other pavement markings, and complex backgrounds, often leading to missed detections or false positives (Cao et al., 2021; Leventi et al., 2025; Rosa et al., 2024). While Tile2Net (Hosseini et al., 2023) does demonstrate the feasibility of segmenting crosswalks from aerial images, it neither aims to recover bar-level detail nor evaluate crosswalk conditions.

To sum up, the growing demand for scalable crosswalk inventories, advances in deep learning, and the increasing availability of high-resolution aerial imagery make automated crosswalk inventory construction both necessary and technically feasible. However, existing methods remain insufficient to support accurate, cost-effective, and scalable crosswalk inventory construction.

To this end, this study proposes CrosswalkNet (XNet), a unified deep-learning framework for large-scale crosswalk localization and visible condition assessment using high-resolution aerial imagery. The framework integrates crosswalk localization and pixel-level representation within a unified pipeline, where crosswalk regions are first identified and then refined to fine-grained visible marking patterns. The resulting outputs are further incorporated into a geospatially aligned workflow to derive an image-based indicator for crosswalk condition assessment. The major contributions of the current study are summarized as follows.

(1) This study presents a framework for automated crosswalk inventory construction from high-resolution aerial imagery, which can be used to extract crosswalk geometries and fine-grained visible marking patterns across diverse roadway environments.
(2) This study establishes a condition-assessment pipeline that derives an image-based indicator of marking condition, which can be used for large-scale screening and prioritization of crosswalk maintenance and inspection.

The remainder of this paper is organized as follows. Section 2 reviews related work on crosswalk inventory construction, oriented bounding box (OBB) detection, and semantic segmentation. Section 3 describes the dataset. Section 4 presents the proposed methodology, including crosswalk localization, segmentation, and condition assessment. Section 5 reports performance results, along with GIS-based inventory construction and condition assessment outcomes. Section 6 concludes the paper and discusses future research directions.

2. Related work

2.1. Crosswalk inventory construction

Crosswalk inventory construction has been addressed through a variety of data sources and methodological approaches, reflecting differences in data availability, spatial scale, and application requirements. Existing studies can be broadly grouped into several categories.

Traditional approaches rely on manual data collection, including field surveys and visual interpretation of aerial imagery. These methods typically involve human annotation of crosswalk locations and attributes, sometimes combined with on-site verification (Hosseini et al., 2023). While they can produce accurate and reliable datasets, their high labor cost and limited scalability make them difficult to apply across large regions. For instance, Moran (2022) manually reviewed satellite imagery to classify the presence and condition of crosswalks at approximately 6400 intersections in San Francisco, a process that required about 90 h of effort by a single researcher. Crowd-sourced platforms such as OpenStreetMap have also been used to create and update crosswalk inventories (Berriel et al., 2017; Moran and Laefer, 2026). However, the available crosswalk information is often incomplete, outdated, and inconsistent across regions, limiting its reliability for large-scale applications.

With the increasing availability of high-resolution remote sensing data, Light Detection and Ranging (LiDAR)-based approaches have emerged as a promising solution for automatically extracting crosswalk features by leveraging both the geometric and radiometric properties of point clouds (Esmorís et al., 2023; Fern and Díaz-vilari, 2025; Riveiro et al., 2015). Specifically, Esmorís et al. (2023) propose two LiDAR-based algorithms to characterize multiple crosswalk attributes, such as paint quality and painted area, based on 48 detected zebra crossings. While their work demonstrates the capability of LiDAR data to generate detailed and accurate attributes, such approaches remain difficult to scale to large-area crosswalk inventory construction due to limitations in data availability, acquisition cost, and processing complexity.

As discussed earlier, computer vision techniques have been widely adopted for crosswalk detection and mapping, largely due to the increasing availability of street-level, satellite, and aerial imagery. Ground-level imagery has been the most commonly used data source in prior studies (e.g., Haselhoff and Kummert, 2010; Kaya et al., 2023; Malbog, 2019; Murali and Coughlan, 2013; Tian et al., 2021; Tümen and Ergen, 2020; Zhai et al., 2015; Zhang et al., 2022), as it provides rich visual details for detecting crosswalk markings. However, its limited spatial coverage and inconsistent geolocation hinder its application to large-scale inventory construction.

Satellite imagery has also been explored in previous research (e.g., Brkić et al., 2023; Shin and Choi, 2025; Verma and Ukkusuri, 2024; Wijnands et al., 2021; Xu et al., 2023), as it offers broader geographic coverage. Nevertheless, its relatively coarse spatial resolution often limits its ability to reliably capture fine-grained pavement markings. In contrast, high-resolution aerial orthoimagery has emerged as a promising data source by balancing spatial detail and geographic coverage, enabling more consistent large-scale crosswalk mapping. However, several challenges remain, including efficient crosswalk localization, robust assessment of marking condition from aerial imagery, and effective integration of such information into scalable crosswalk inventory systems.

Currently, few studies provide unified frameworks that simultaneously support both large-scale inventory construction and condition assessment. Moreover, achieving consistent performance across large geographic areas remains challenging due to variations in image quality, occlusion, and complex background conditions. These limitations highlight the need for advanced approaches that can fully leverage high-resolution aerial imagery to enable scalable and reliable crosswalk inventory construction.

2.2. OBB detection

Object detection is a fundamental task in computer vision that involves locating objects in an image and identifying their categories.

Since the rise of deep learning, the field has progressed rapidly, forming two major families of approaches: two-stage and one-stage detectors. Representative two-stage frameworks include Region-based Convolutional Neural Network (R-CNN) (Girshick et al., 2014), Spatial Pyramid Pooling Networks (SPPNet) (He et al., 2015), Fast R-CNN (Girshick, 2015), Mask R-CNN (He et al., 2017), and Feature Pyramid Networks (FPN) (Lin et al., 2017), while one-stage detectors emphasize real-time performance and include models such as You Only Look Once (YOLO) (Redmon et al., 2016), Single Shot MultiBox Detector (SSD) (Liu et al., 2016), and Single-Shot Refinement Detector (RefineDet) (Zhang et al., 2018).

Traditional detectors typically predict HBBs. However, when objects exhibit arbitrary sizes, elongated shapes, or large variations in orientation, HBBs often enclose large amounts of irrelevant background, which degrades detection accuracy. This limitation is especially pronounced in high-resolution remote sensing imagery, where objects such as ships, vehicles, and buildings frequently appear at diverse orientations and may be densely distributed (Wang et al., 2025; Wen et al., 2023). Motivated by these challenges, oriented object detection has emerged as an important research direction.

Early oriented detectors extend horizontal box frameworks by introducing rotational data augmentation, directional anchors, or rotation-aligned feature extraction. Representative anchor-based methods include the Rotation-Invariant CNN (RICNN) (Cheng et al., 2016), Rotated Region Proposal Networks (R2PN) (Zhang et al., 2018), Region of Interest (RoI) Transformer (Ding et al., 2018), and Single-Shot Alignment Network (S2A-Net) (Han et al., 2022). Although these approaches achieve strong accuracy, they rely heavily on manually designed anchors, making them sensitive to hyperparameter choices and computationally expensive when handling multiple scales and angles (Wen et al., 2023). These drawbacks have created a clear need for more flexible and lightweight solutions, especially for high-resolution scenes where object orientations vary continuously.

To overcome these limitations, recent research has shifted toward anchor-free oriented detection to simplify model design, remove the burden of handcrafted anchor engineering, and achieve better generalization across complex object geometries. A notable trend is the adoption of polar coordinate representations, where object geometry is described using radial distances and angular deviations, offering a more natural way to model orientation. For example, Liu et al. (2021) introduce the center-boundary dual attention network (CBDA-Net), which leverages dual attention to highlight discriminative features of rotating objects while suppressing background noise. Zhou et al. (2020) propose an anchor-free polar remote sensing object detector (P-RSDet), where arbitrary-oriented object detection can be achieved by predicting the center point and regressing one polar radius and two polar angles. Yi et al. (2021) design box boundary-aware vectors (BBAVectors), representing each oriented box using four boundary-aware vectors extending from the center to its quadrantal edges. Zhang et al. (2022) further develop a dense anchor-free rotated object detector (DARDet), which directly predicts five parameters of an OBB at each foreground pixel in a dense anchor-free manner. Li et al. (2026) propose the geometry-aware detector (GADet), which combines adaptive convolution, rotation-sensitive attention, and pruning to achieve a strong accuracy-efficiency balance.

These anchor-free designs converge on a common goal to reduce architectural complexity while preserving or improving detection accuracy.

Building upon these insights, this paper follows this anchor-free paradigm and further adapts it to the specific requirements of the crosswalk detection task. By leveraging orientation-aware design choices, the aim is to retain the efficiency of one-stage detectors while achieving robust performance under diverse crosswalk orientations.

2.3. Semantic segmentation

Semantic segmentation assigns a semantic label to every pixel of an image and has advanced significantly with the development of deep convolutional neural networks. Early approaches such as Fully Convolutional Networks (FCN) (Long et al., 2015) enable end-to-end dense prediction but rely on low-resolution feature maps, often resulting in coarse boundaries and limited spatial detail. This limitation is particularly critical for tasks involving thin and elongated structures.

To improve spatial localization, encoder-decoder architectures such as U-Net (Ronneberger et al., 2015) introduce skip connections to recover high-resolution features. While effective in preserving spatial details, these models primarily rely on local feature aggregation and may struggle to capture long-range contextual dependencies, especially in complex scenes with repetitive patterns. Subsequent methods such as Pyramid Scene Parsing Network (PSPNet) (Zhao et al., 2017) and the DeepLab series (Chen et al., 2017a, 2017b, 2018) address this issue by incorporating multi-scale context through pyramid pooling and atrous convolution. However, although these approaches improve global context modeling, they are still designed for generic segmentation tasks and do not explicitly exploit structural priors of target objects.

Recent advances further explore transformer-based and high-resolution architectures. Models such as SegFormer (Xie et al., 2021) and Swin-UNet (Cao et al., 2022) leverage transformer-based encoders to capture long-range global context, while Mask2Former (Cheng et al., 2022) introduces a unified framework using masked attention mechanisms. Despite their strengths, these approaches primarily prioritize dense feature representation and global modeling. This focus may be suboptimal for detecting crosswalk markings in aerial imagery, where targets appear as small-scale, repetitive patterns that are frequently obscured by shadows or environmental noise.

To mitigate these limitations, attention mechanisms offer a promising solution by selectively emphasizing informative regions while suppressing irrelevant features. General-purpose modules such as Convolutional Block Attention Module (CBAM) (Woo et al., 2018) and Dual Attention Networks (DANet) (Fu et al., 2019) have already demonstrated significant improvements in feature discriminability. Building on these foundations, recent studies in crack detection have also proven that integrating attention can enhance the model's performance under challenging scenarios (Li et al., 2024; Sun et al., 2022; Zhu et al., 2024, 2025). For instance, OUR-Net (Li et al., 2024) pioneers a multi-frequency feature modeling approach to simultaneously capture global structures and fine-grained details, effectively improving the detection of cracks with varying widths. Zhu et al. (2024) propose RHACrackNet, which integrates hybrid attention and residual structures to further refine feature representation, achieving a favorable balance between accuracy and computational efficiency. Pushing this efficiency further, AutoCrackNet (Zhu et al., 2025) utilizes neural architecture search to automatically design lightweight models, demonstrating the potential of combining efficient architectures with attention mechanisms.

Despite the success of these models in capturing fine structures at close range, their performance remains underexplored when transitioning to aerial platforms. Aerial crosswalk segmentation is uniquely characterized by reduced spatial resolution, sub-optimal target scales, and potential environmental interference from shadows or complex backgrounds. To date, few studies have investigated whether the architectural advantages of attention-based models can be effectively leveraged to resolve the repetitive, small-scale patterns of crosswalk bars under these constraints. Consequently, evaluating the efficacy of these mechanisms in the aerial domain serves as a primary objective of the current research.

3. Data

For model training and validation, high-resolution orthorectified

aerial imagery is utilized. Compared to raw aerial imagery, which is affected by sensor orientation, platform instability, and geometric distortions, orthorectified imagery is corrected to provide uniform scale and accurate georeferencing, enabling direct transformation between image and geographic coordinates (Hosseini et al., 2023).

In recent years, high-resolution orthoimagery has become increasingly accessible. In the United States, the U.S. Geological Survey (USGS) provides nationwide coverage, and many states routinely acquire aerial imagery. The aerial imagery used for training in this study is obtained from multiple state and regional datasets, including Massachusetts, Florida, and Washington. Each image contains approximately 100 million pixels (10,000×10,000), with a ground sampling distance of 6 inches (15 cm) per pixel. This spatial resolution is sufficient to capture fine-scale details, enabling reliable identification of small features such as pedestrian crosswalks.

Two categories of crosswalks were defined in this study: (1) zebra/continental/ladder crosswalks (hereafter "stripe"), characterized by broad white stripes and grouped together due to their visual similarity and common usage; and (2) two-parallel-line crosswalks (hereafter "parallel"), consisting of two solid boundary lines with or without painted infill. Since both variants share the same geometric structure, they were consolidated into a single annotation class. Representative examples of the two types are illustrated in Fig. 1.

The annotation process is conducted in two stages. First, a subset of aerial images (10,000×10,000 pixels) is manually labeled by trained research assistants to create a seed dataset. An initial detection model is then trained on this dataset and applied to additional images to generate preliminary annotations. All the annotations are subsequently verified and corrected through manual inspection by two domain experts to ensure consistency. In addition, a bounding box rectification strategy is applied to refine annotation quality and improve geometric accuracy.

Note that due to the large spatial resolution of the original aerial images (10,000×10,000 pixels), a pre-processing strategy is applied to generate smaller tiles suitable for deep learning. Each image is divided into 1024×1024 patches using a sliding-window approach with an overlap of 276 pixels. This procedure produces manageable image segments that can be processed individually by the model, while preserving contextual information across patch boundaries, which is particularly beneficial for detecting features that span multiple tiles. Following annotation and tiling, the dataset is partitioned into a training set of 18,879 tiles (80%) and a test set of 4740 tiles (20%), as shown in Table 1.

4. Methodology

The proposed XNet framework consists of two modules: an oriented bounding box detection module (XNet-OBB) and a semantic segmentation module (XNet-Seg). As shown in Fig. 2, the input aerial images (10,000 × 10,000) are first partitioned into smaller tiles of size 1024 × 1024 with an overlap of 276 pixels to facilitate efficient processing. The XNet-OBB module is then applied to accurately localize crosswalk instances, generating oriented bounding boxes that define regions of interest for subsequent analysis. Based on these detected regions, the corresponding image patches are extracted and fed into the XNet-Seg module, which performs pixel-level inference to segment crosswalk markings and distinguishes between intact and degraded areas.

The segmentation results within each detected OBB are subsequently used to generate GIS-compatible layers and to compute a condition ratio, defined as the proportion of valid crosswalk pixels relative to the total OBB area. This ratio serves as an image-based indicator of marking visibility and completeness. It should be noted that this indicator reflects only the visible condition of crosswalk markings in aerial imagery and does not capture physical factors such as retroreflectivity, material thickness, skid resistance, paint aging, or nighttime visibility. Therefore, the proposed framework is intended as an efficient large-scale screening tool to identify potentially degraded crosswalk markings and support subsequent detailed inspection or multi-sensor evaluation.

4.1. XNet-OBB

The OBB stage follows a standard object detection pipeline consisting of three key modules: (1) a backbone network for multi-scale feature extraction, (2) a neck module for hierarchical feature fusion, and (3) a detection head for OBB prediction. The subsequent sections provide detailed descriptions of each module, along with the employed loss functions and evaluation metrics.

4.1.1. Framework

4.1.1.1. Backbone for multi-scale feature extraction. The backbone transforms input aerial tiles (1024 × 1024 × 3) into hierarchical feature representations through a sequence of lightweight convolutional and residual-attention blocks. It begins with a stem convolution for rapid spatial downsampling, followed by multiple C3K2 blocks for early-stage feature extraction, a Spatial Pyramid Pooling-Fast (SPPF) module for multi-scale receptive field aggregation, and a final C2PSA block for global context enhancement. Fig. 3 presents the network structure of the C3K2, C2PSA, and SPPF.

In the shallow and mid-level stages, the C3K2 blocks employ cross-stage residual connections with depthwise separable convolutions to efficiently capture edge and texture details while maintaining low computational cost. The SPPF module aggregates multi-scale context by performing parallel pooling operations with different receptive field sizes, enhancing the network's ability to encode global spatial dependencies. Finally, the C2PSA block integrates Cross-Stage Partial (CSP) connections with Parallel Spatial Attention (PSA) to jointly model spatial and channel dependencies, thereby improving feature discriminability for small or occluded crosswalks.

A generic residual transformation within a backbone block can be formulated as:

$$\boldsymbol{y} = \Phi(\boldsymbol{x}) + BN(DWConv_{3\times3}(\sigma(BN(Conv_{3\times3}(\boldsymbol{x})))))) \quad (1)$$

where $\sigma(\cdot)$ denotes the SiLU activation function. The operator $\varnothing(\cdot)$ is implemented as an identity mapping when input and output channel dimensions match, or as a 1×1 convolution followed by batch normalization when dimensionality alignment is required during channel expansion or downsampling.

4.1.1.2. Neck for multi-scale feature fusion. Given that crosswalks may



Fig. 1. Examples of crosswalks from the dataset.

Table 1
Image (1024×1024) count and instance distribution of the dataset.

	Massachusetts	Florida	Washington	Total	Stripe	Parallel
Train	11,511	5872	1496	18,879	20,718	5215
Test	2891	1472	376	4740	5232	1361

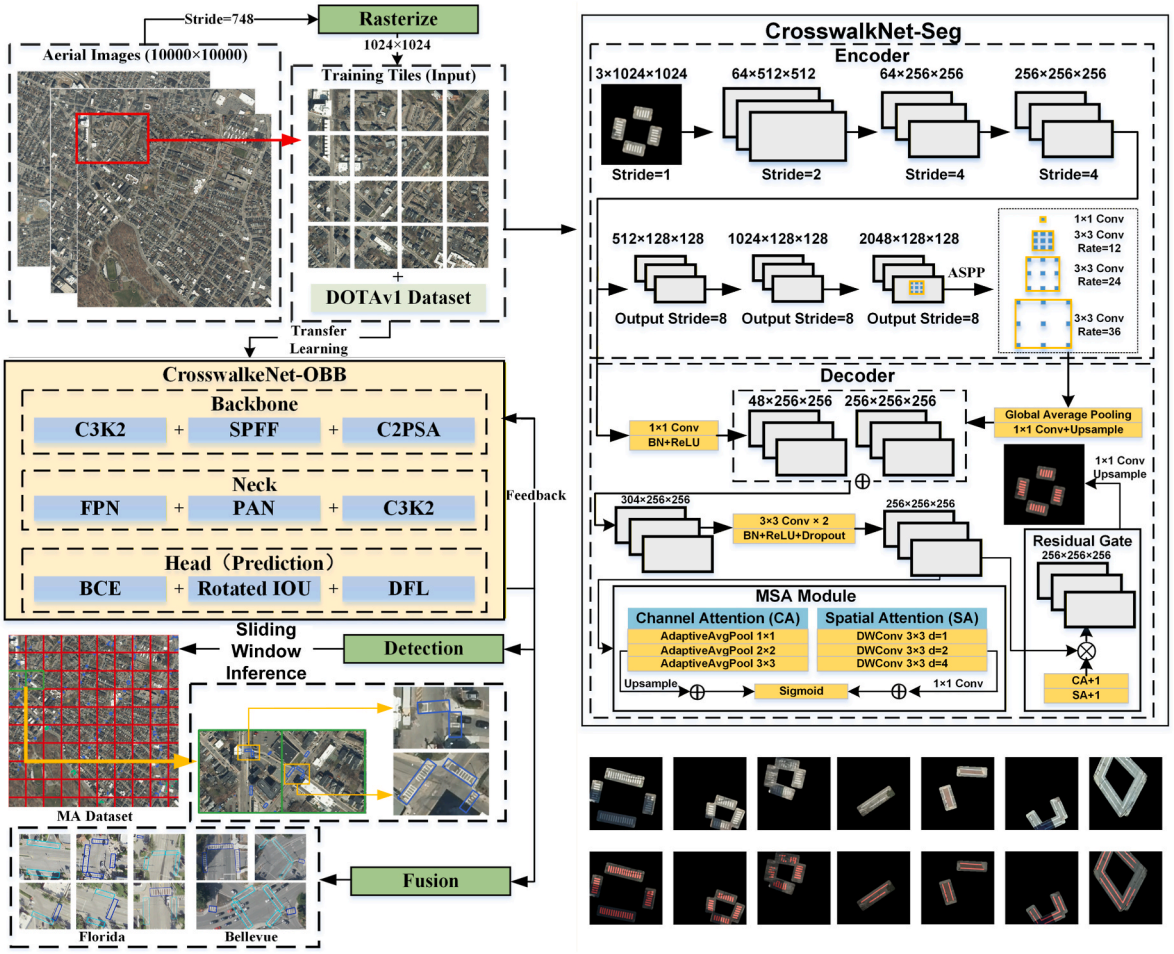


Fig. 2. XNet framework.

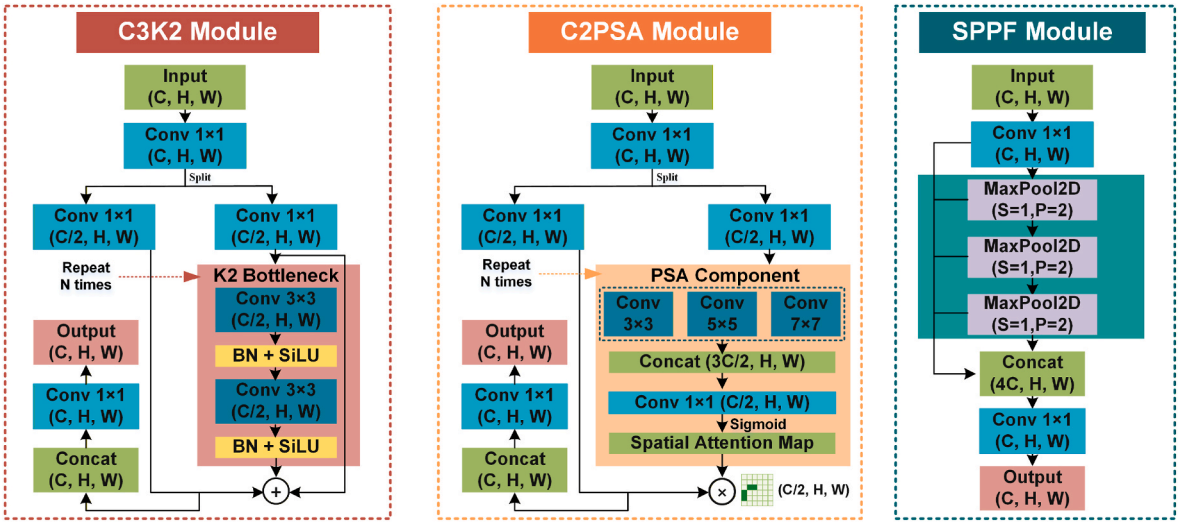


Fig. 3. Network structure of C3K2, C2PSA, and SPPF.

appear at varying scales in aerial imagery, the neck is designed to achieve robust cross-scale feature integration by combining a Feature Pyramid Network (FPN) with a Path Aggregation Network (PAN).

The FPN introduces a top-down pathway, where deep semantic features are progressively upsampled and concatenated with higher-resolution features from shallower layers, enriching low-level representations with contextual information. Conversely, the PAN adds a bottom-up pathway that passes fine-grained spatial details upward, enhancing localization precision and geometric consistency.

At each fusion stage, feature maps from adjacent levels (with strides 8, 16, and 32) are merged through channel concatenation and refined by lightweight C3K2 fusion blocks, which employ depthwise convolutions and residual shortcuts to reduce redundancy and maintain computational efficiency.

This bidirectional feature aggregation process can be expressed as:

$$\{P_3, P_4, P_5\} = \{\psi_3(C_3 \oplus Up(C_4)), \psi_4(C_4 \oplus Up(C_5)), \psi_5(C_5)\} \tag{2}$$

where C_3, C_4 and C_5 are the feature maps extracted from the backbone at strides 8, 16, and 32, respectively; $\oplus$ represents channel concatenation, and $\psi_i(\cdot)$ $(i = 3, 4, 5)$ are C3K2 fusion modules that refine the combined features through residual attention and depthwise separable convolution.

As a result, the neck produces semantically rich and spatially precise multi-scale features (P_3, P_4, P_5), which are subsequently passed to the detection head for OBB regression and classification.

4.1.1.3. Detection head with anchor-free oriented bounding boxes. Crosswalks in aerial imagery typically appear as elongated rectangular patterns that may be arbitrarily rotated according to road orientation. Representing such objects with conventional horizontal bounding boxes (HBBs) introduces excessive background, lowers intersection-over-union (IoU) under oblique angles, and increases conflicts during non-maximum suppression (NMS). To overcome these limitations, the proposed framework adopts oriented bounding boxes (OBBs) to explicitly encode object rotation. Each OBB is parameterized as (C_x, C_y, w, h, θ), where (C_x, C_y) denotes the box center, w and h represent its width and height, and θ is the orientation angle.

The detection head transforms the multi-scale feature maps (P_3, P_4, P_5) from the neck into OBB predictions. Each detection layer corresponds to a specific stride (8, 16, and 32), enabling the model to detect crosswalks of varying sizes with a balance between fine localization and high-level semantic precision. At each grid location (m, n) of a feature map with stride s, the regression branch predicts a parameter vector:

$$t = (t_x, t_y, t_w, t_h, t_\theta) \tag{3}$$

These parameters are decoded into an oriented bounding box as:

$$C_x = s \times (m + \sigma(t_x)) \tag{4}$$

$$C_y = s \times (n + \sigma(t_y)) \tag{5}$$

$$w = s \times e^{t_w} \tag{6}$$

$$h = s \times e^{t_h} \tag{7}$$

$$\theta = \pi \times \sigma(t_\theta) - \frac{\pi}{2} \tag{8}$$

where $\sigma(\cdot)$ is the sigmoid function.

To ensure a unique and consistent representation, two post-processing steps are applied. First, canonicalization enforces $w \geq h$:

$$(w, h, \theta) = \begin{cases} \left(h, w, \theta + \frac{\pi}{2}\right), \text{if } w < h \\ (w, h, \theta), \text{otherwise} \end{cases} \tag{9}$$

Subsequently, the angle is normalized into the interval $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right)$:

$$\theta = \left(\left(\theta + \frac{\pi}{2}\right) mod\ \pi\right) - \frac{\pi}{2} \tag{10}$$

In parallel, the classification branch outputs the foreground probability p for each grid cell.

4.1.1.4. Model configurations. The baseline model architecture is presented in Table 2. To explore the trade-off between accuracy and efficiency, five model configurations with varying channel width scales are evaluated in this study, as summarized in Table 3. Specifically, two scaling dimensions are considered, including depth scaling and width scaling. Depth scaling determines the number of repeated blocks within each stage, which controls the network depth and thus affects the model capacity and the level of feature abstraction. Width scaling, on the other hand, controls the number of channels in each layer, thereby adjusting the feature dimensionality and computational cost.

4.1.2. Loss function design

Accurate crosswalk detection requires simultaneously addressing three challenges: (1) handling the severe imbalance between background and crosswalk samples, (2) distinguishing crosswalks from visually similar linear road markings under occlusion or illumination changes, and (3) capturing the orientation of crosswalk stripes relative to the roadway axis.

To this end, a multi-task loss is adopted in the current study, which integrates object-class prediction (L_{cls}), rotated-box regression ($L_{rot-box}$), and distribution focal refinement (L_{dfl}). The overall objective is formulated as:

$$L_{obb} = \lambda_{cls} L_{cls} + \lambda_{box} L_{rot-box} + \lambda_{dfl} L_{dfl} \tag{11}$$

where λ_{cls}, λ_{box} and λ_{dfl} represent the weighting factors.

For L_{cls}, a binary cross-entropy (BCE) loss on sigmoid-activated logits is employed for each class (stripe, parallel):

$$L_{cls} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{2} \left[y_{i,k} \log p_{i,k} + \left(1 - y_{i,k}\right) \log \left(1 - p_{i,k}\right) \right] \tag{12}$$

where $p_{i,k} \in [0, 1]$ represents the predicted probability that candidate location i belongs to class k, $y_{i,k} \in [0, 1]$ is the corresponding ground-truth label, and N is the total number of training instances.

Table 2
Original model architecture of XNet-OBB.

Section	Stage	Module	Output size
Backbone	0	Stem Conv (stride = 2)	512 × 512 × 64
	1	Conv (stride = 2) →P2	256 × 256 × 128
	2	C3K2	256 × 256 × 256
	3	Conv (stride = 2) →P3	128 × 128 × 256
	4	C3K2	128 × 128 × 512
	5	Conv (stride = 2) →P4	64 × 64 × 512
	6	C3K2	64 × 64 × 512
	7	Conv (stride = 2) →P5	32 × 32 × 1024
	8	C3K2	32 × 32 × 1024
	9	SPPF	32 × 32 × 1024
	10	C2PSA	32 × 32 × 1024
Neck	11-13	Top-down: Upsample (P5) + Concat (P4) + C3K2 →P4′	64 × 64 × 512
	14-16	Top-down: Upsample (P4′) + Concat (P3) + C3K2 →P3′	128 × 128 × 256
	17-19	Bottom-up: Downsample (P3′) + Concat (P4′) + C3K2 →P4″	64 × 64 × 512
	20-22	Bottom-up: Downsample (P4″) + Concat (P5) + C3K2 →P5′	32 × 32 × 1024
Head	23	Anchor-free OBB head (Detect (P3′, P4″, P5′))	128 × 128 × 71 (P3′) 64 × 64 × 71 (P4″) 32 × 32 × 71 (P5′)

Table 3
Model configuration of XNet-OBB.

Model	Scale (Depth)	Scale (Width)	Backbone channels	Neck channels
XNet-OBB-XSmall	0.5	0.25	[16, 32, 64, 128, 256]	[128, 64, 128, 256]
XNet-OBB-Small	0.5	0.50	[32, 64, 128, 256, 512]	[256, 128, 256, 512]
XNet-OBB-Medium	0.5	1.00	[64, 128, 256, 512, 512]	[512, 256, 512, 512]
XNet-OBB-Large	1.0	1.00	[64, 128, 256, 512, 512]	[512, 256, 512, 512]
XNet-OBB-XLarge	1.0	1.50	[96, 192, 384, 768, 768]	[768, 384, 768, 768]

For the rotated-box regression loss $L_{rot-box}$, a rotated IoU loss is employed to directly reflect the spatial overlap between predicted and ground-truth quadrilaterals:

$$IoU_R(\boldsymbol{b},\boldsymbol{b}^*)=\frac{A(\boldsymbol{b}\cap\boldsymbol{b}^*)}{A(\boldsymbol{b}\cup\boldsymbol{b}^*)} \tag{13}$$

where $A(\cdot)$ denotes polygon area, and $\boldsymbol{b},\boldsymbol{b}^*$ are predicted and ground-truth oriented boxes. The regression loss is then expressed as:

$$L_{rot-box}=1-IoU_R(\boldsymbol{b},\boldsymbol{b}^*) \tag{14}$$

which penalizes positional, scale, and angular misalignments jointly.

L_{dfl} is utilized to further refine the predicted edge distances by modeling them as discrete probability distributions:

$$L_{dfl}=-\frac{1}{M}\sum_{m=1}^{M}q_m log r_m \tag{15}$$

where q_m denotes the ground-truth distribution of edge distances and r_m is the predicted probability across M (=16 in this study) discrete bins.

4.1.3. Evaluation metrics

To comprehensively evaluate the detection performance of XNet-OBB, multiple quantitative metrics are employed to measure the model's capability in distinguishing crosswalks from background clutter. These include Precision (*Pr*), Recall (*Re*), and mean Average Precision (*mAP*) under different IoU thresholds.

The Precision (*Pr*) and Recall (*Re*) are defined as:

$$Pr=\frac{TP_k}{TP_k+FP_k} \tag{16}$$

$$Re=\frac{TP_k}{TP_k+FN_k} \tag{17}$$

where TP_k, FP_k and FN_k represent the number of true positives, false positives, and false negatives, respectively. Precision measures the reliability of predictions (the fraction of detected crosswalks that are correct), while recall reflects completeness (the fraction of ground-truth crosswalks that are successfully detected).

As for the *mAP*, which is obtained by averaging Average Precision (AP), two widely used variants are adopted in the current study:

$$mAP_{50}=AP_k^{IoU=0.5} \tag{18}$$

$$mAP_{50-95}=\frac{1}{10}\sum_{t=0}^{9}AP_k^{IoU=0.5+0.05t} \tag{19}$$

where $AP_k^{IoU=\tau}$ represents the AP of class k at IoU threshold τ.

4.2. XNet-Seg

Following the OBB-based detection stage, the XNet-Seg module refines crosswalk analysis through pixel-level segmentation and condition assessment. Each detected crosswalk region is cropped according to its OBB, while all areas beyond the OBB are masked to remove irrelevant background. Building on this localized input, the segmentation stage employs a semantic segmentation framework that transforms detection-level geometry into interpretable condition indicators, forming the basis for quantitative analysis of crosswalk wear and degradation. The following sections describe the segmentation framework, loss functions, evaluation metrics, and the derivation of the normalized condition ratio used for crosswalk condition assessment.

4.2.1. Framework

An end-to-end segmentation framework is established consisting of an encoder and a decoder. The encoder extracts multi-level contextual features using a ResNet-50 backbone augmented with Atrous Spatial Pyramid Pooling (ASPP) to enlarge the receptive field and capture scale-invariant context. The decoder progressively restores spatial resolution and refines boundary details through low-level feature fusion. To further enhance feature discrimination, a Multi-Scale Attention (MSA) module is embedded in the decoder to adaptively recalibrate multi-scale feature responses across both channel and spatial dimensions.

4.2.1.1. Encoder network and ASPP module. Convolutional encoders are known to lose spatial detail because repeated pooling and striding reduce feature resolution. While deeper layers capture high-level semantics beneficial for classification, the loss of fine-grained localization can hinder precise segmentation. To mitigate this, the proposed framework employs a ResNet-50 backbone with modified strides: the last two convolutional stages (layers 3 and 4) are replaced by dilated convolutions, enlarging the receptive field without increasing parameters.

To capture multi-scale contextual information, an ASPP module is applied to the high-level feature map. It comprises parallel atrous convolutions with dilation rates of 12, 24, and 36, together with a global average-pooling branch. The outputs of these branches are concatenated and projected to a 256-channel feature space, effectively aggregating context from multiple receptive fields.

Formally, the output $y(i)$ of a convolution operation with a kernel size of J is defined as:

$$y(i)=\sum_{j=1}^{J}x(i+r\cdot j)\cdot w(j) \tag{20}$$

where $y(i)$ denotes the output feature value at spatial position i, $x(i+r\cdot j)$ represents the input feature value at an offset determined by the dilation rate r and kernel index j; $w(j)$ is the kernel weight at index j, and J is the total size of the convolution kernel, and r denotes the dilation rate, which controls the spacing between kernel elements in a dilated convolution.

4.2.1.2. Decoder and MSA module. The architecture of the decoder and the MSA module is shown in Fig. 4. Following the ASPP stage, the high-level feature map is upsampled and fused with the low-level encoder feature from layer 1 to recover spatial precision. Before fusion, the low-level feature is reduced to 48 channels using a 1×1 convolution and batch normalization to balance scale differences. The concatenated tensor is refined through two successive 3×3 convolutional blocks with ReLU activation and a 0.1 dropout rate. This step restores structural details lost during encoding, improving crosswalk boundary completeness and overall segmentation accuracy.

The decoder further incorporates an MSA module to dynamically recalibrate feature responses across both channel and spatial dimensions. Unlike fixed-weight fusion strategies that simply sum up or concatenate multi-scale features, the MSA adaptively emphasizes

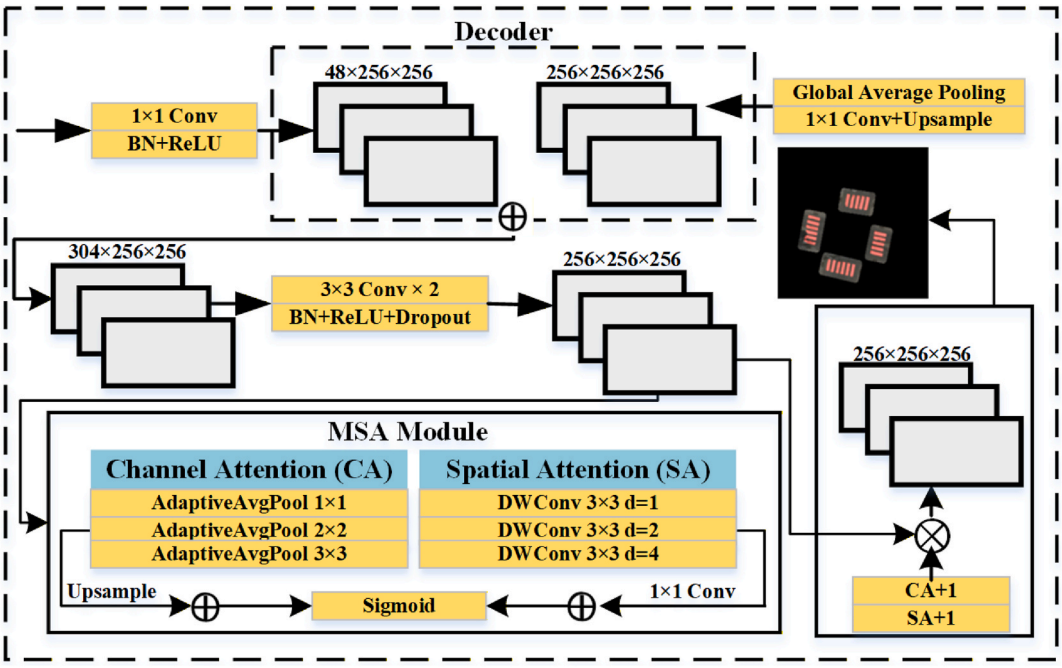


Fig. 4. The architecture of the decoder and MSA.

informative regions and suppresses background noise through self-attention within a single feature map. The module consists of two complementary branches: channel attention and spatial attention. The channel attention branch applies adaptive average pooling at three scales (1, 2, 4) followed by shared multi-layer perceptrons (MLPs) to generate scale-aware channel weights, enabling the network to strengthen discriminative feature channels. The spatial attention branch utilizes depthwise dilated convolutions with dilation rates of 1, 2, and 4 to capture contextual dependencies and texture variations over multiple spatial extents.

The attention responses from all scales are aggregated and applied through a residual gating formulation:

$$\widehat{X} = X \odot \left(1 + \sigma\left(\sum_{d\in D}\left(f_C^{(d)}(X) + f_D^{(d)}(X)\right)\right)\right) \tag{21}$$

where X is the fused decoder output and $\widehat{X}$ is the refined feature map; $D = (1, 2, 4)$ represents the set of spatial scales used in the MSA module; $f_C^{(d)}(X)$ and $f_D^{(d)}(X)$ are the channel- and spatial-attention responses computed at scale d.

4.2.2. Loss function and metrics

In the crosswalk segmentation task, pixels from the background and the crosswalks are significantly imbalanced. Imbalanced datasets can cause instability in training a well-established, generative, and discriminative model (Sun et al., 2022). To address this problem and improve the dice coefficient performance, a composite loss function including both weighted binary cross-entropy (BCE) and dice loss is used in the current study:

$$L_{segmentation} = \lambda_{BCE} L_{BCE} + \lambda_{Dice} L_{Dice} \tag{22}$$

with

$$L_{BCE} = -\frac{1}{N}\sum_{n=1}^{N}[y_n \log p_n + (1-y_n)\log(1-p_n)] \tag{23}$$

$$L_{Dice} = 1 - \frac{2\sum_{n=1}^{N} y_n p_n + \varepsilon}{\sum_{n=1}^{N} p_n^2 + \sum_{n=1}^{N} y_n^2 + \varepsilon} \tag{24}$$

where λ_{BCE} and λ_{Dice} represent the weight factors of the BCE and dice loss (both of which are set as 1 for the current study); p_n and y_n represent the predicted and ground-truth probabilities for pixel n, and ε can ensure the loss function stability by avoiding the numerical issue of dividing by 0. N denotes the total number of pixels.

Similar to XNet-OBB, the Precision (*Pr*) and Recall (*Re*) are used to evaluate the performance of the segmentation model; besides, mean Intersection-over-Union (*mIoU*) and mean Dice coefficient (*mDice*) are reported over the crosswalk and background classes:

$$mIoU = \frac{1}{2}\left(\frac{TP_c}{TP_c + FP_c + FN_c} + \frac{TP_b}{TP_b + FP_b + FN_b}\right) \tag{25}$$

$$mDice = \frac{1}{2}\left(\frac{2TP_c}{2TP_c + FP_c + FN_c} + \frac{2TP_b}{2TP_b + FP_b + FN_b}\right) \tag{26}$$

where c and b denote the crosswalk and background classes, respectively. *TP*, *FP* and *FN* represent the number of true positives, false positives, and false negatives for each class.

4.2.3. Condition assessment

The condition of each detected crosswalk is estimated using an image-based indicator derived from the proportion of visible pavement markings within its OBB. After the OBB-based detection and segmentation stages, the segmented stripe regions are spatially intersected with the corresponding crosswalk polygons to isolate valid marking pixels. The total segmented area inside each OBB (A_{seg}) and the OBB's geometric area (A_{obb}) are computed in a projected coordinate system. The raw condition ratio (RCR) of the s-th crosswalk, RCR_s, is then defined as the fraction of the segmented area within the OBB:

$$RCR_s = \frac{A_{seg,s}}{A_{obb,s}} \tag{27}$$

To ensure fair comparison across structurally distinct crosswalk

types, a per-class normalization strategy is adopted. Since parallel crosswalks inherently contain less painted areas within their OBBs than stripe types, their RCRs are naturally lower even under intact conditions. As a result, applying global normalization would introduce systematic bias in condition-level interpretation across different crosswalk types. To address this issue, a normalized condition ratio (NCR) is proposed using per-class min-max scaling:

$$NCR_s^{(k)} = \frac{RCR_s^{(k)} - \min\limits_{q\in\Omega_k}\left(RCR_q^{(k)}\right)}{\max\limits_{q\in\Omega_k}\left(RCR_q^{(k)}\right) - \min\limits_{q\in\Omega_k}\left(RCR_q^{(k)}\right)} \tag{28}$$

where $NCR_s^{(k)}$ represents the NCR of the s-th crosswalk in class k, and Ω_k denotes all samples of that class.

5. Results and discussion

5.1. OBB detection

5.1.1. Model performance

All training and evaluation experiments are conducted on a high-performance workstation equipped with an AMD Ryzen Threadripper PRO 7985WX CPU (64 cores, 128 threads), 64 GB of RAM, and two NVIDIA RTX PRO 6000 Blackwell GPUs with 95 GB of memory each. All models are trained for 100 epochs with a batch size of 4 on a single GPU. The model is optimized using stochastic gradient descent (SGD) with an initial learning rate of 0.01, momentum of 0.937, and a weight decay of 0.0005. A warm-up strategy is applied during the first three epochs to stabilize the training process. The loss weights are set as $\lambda_{cls} = 0.5$, $\lambda_{box} = 7.5$, $\lambda_{dfl} = 1.5$. Evaluation is performed on a single GPU using an IoU threshold of 0.7 for oriented bounding box assessment. All models are evaluated under identical hardware and experimental settings to ensure fair and consistent comparison. Tables 4 and 5 summarize the accuracy and efficiency performance of the proposed XNet-OBB models of different scales on the test set.

All model variants achieve strong detection performance, with mAP_{50} consistently exceeding 0.970. Even the lightweight XSmall model reaches mAP_{50} of 0.970 with precision/recall of 0.945/0.935, indicating that the proposed architecture maintains high discriminative capability at low model capacity. As the model scales from XSmall to XLarge, accuracy improves gradually, particularly under the stricter metric $mAP_{50\text{-}95}$ (from 0.836 to 0.893). However, the performance gain becomes marginal beyond the Medium configuration (from 0.871 to 0.893), suggesting diminishing returns as model complexity increases.

From an efficiency perspective, the scaling behavior is clear and consistent. The inference time increases from 5.096 ms (196.235 FPS) for XSmall to 16.492 ms (60.632 FPS) for XLarge, while the computational cost (FLOPs) and peak GPU memory consumption increase substantially from 16.792 G to 518.934 G and from 103.489 MB to 605.881 MB, respectively. This indicates that larger models impose substantially higher computational and memory demands.

Across all configurations, stripe crosswalks consistently achieve higher accuracy than parallel ones. Specifically, at the XLarge scale, stripe crosswalks reach precision/recall of 0.980/0.982 compared to 0.960/0.928 for parallel types. This difference is likely due to the more distinctive geometric structure and denser visual patterns of stripe crosswalks, whereas parallel markings are sparser and more sensitive to occlusion, glare, and degradation. Despite this, the performance on parallel crosswalks remains strong, particularly in larger models, confirming that the OBB-based representation effectively captures orientation and spatial layout.

Overall, the XNet-OBB family demonstrates a clear accuracy-efficiency trade-off. While larger models achieve slightly higher detection accuracy, lightweight and medium configurations retain comparable performance with significantly lower computational cost and faster inference speed. This scalability highlights the flexibility of the proposed framework, allowing practitioners to select an appropriate model variant based on application requirements, ranging from high-throughput large-scale mapping to resource-constrained real-time deployment.

Table 4
Accuracy performance comparison of different XNet-OBB models.

Model	Category	Precision	Recall	mAP_{50}	$mAP_{50\text{-}95}$
XNet-OBB-XSmall	All	0.945	0.935	0.970	0.836
	Stripe	0.966	0.969	0.990	0.851
	Parallel	0.924	0.901	0.950	0.822
XNet-OBB-Small	All	0.966	0.939	0.976	0.854
	Stripe	0.980	0.973	0.992	0.869
	Parallel	0.952	0.905	0.960	0.840
XNet-OBB-Medium	All	0.972	0.944	**0.981**	0.871
	Stripe	**0.981**	0.975	**0.993**	0.881
	Parallel	0.962	0.913	**0.969**	0.860
XNet-OBB-Large	All	**0.973**	0.949	0.979	0.871
	Stripe	0.980	0.975	**0.993**	0.883
	Parallel	**0.966**	0.923	0.965	0.860
XNet-OBB-XLarge	All	0.970	**0.955**	0.980	**0.893**
	Stripe	0.980	**0.982**	**0.993**	**0.904**
	Parallel	0.960	**0.928**	0.967	**0.882**

Table 5
Efficiency performance comparison of different XNet-OBB models.

Model	Inference time (ms)↓	FPS↑	FLOPs (G)↓	Params (M)↓	PeakMem (MB)↓
XSmall	**5.096**	**196.235**	**16.792**	**2.654**	**103.489**
Small	5.594	178.761	57.050	9.700	178.266
Medium	6.723	148.736	182.634	20.880	317.041
Large	8.763	114.111	231.092	26.129	369.162
XLarge	16.492	60.632	518.934	58.738	605.881

*This indicates the inference time per 1024×1024 image.

5.1.2. Model comparison and ablation study

Table 6 compares XNet-OBB-XLarge with several state-of-the-art oriented object detection models. Among all methods, XNet-OBB-XLarge achieves the best overall detection performance. Compared with representative two-stage and rotation-aware detectors such as Oriented R-CNN and ReDet, the proposed model shows clear improvements in both localization accuracy and robustness under stricter evaluation metrics. Although ReDet achieves competitive performance in mAP_{50} (0.978), it remains inferior to XNet-OBB-XLarge with lower $mAP_{50\text{-}95}$ (0.861), indicating that the proposed architecture is more effective in capturing complex geometric and orientation patterns of crosswalks.

From an efficiency perspective, different models exhibit distinct trade-offs. ReDet achieves relatively high inference speed (95.093 FPS) and low FLOPs (54.360 G), but at the cost of reduced accuracy. BBA-Vectors is computationally heavy, with the highest FLOPs (578.493 G) and parameter size (72.421 M), while also delivering the lowest detection accuracy ($mAP_{50\text{-}95} = 0.818$), indicating poor efficiency in terms of both accuracy and resource usage. XNet-OBB-XLarge achieves the highest accuracy but requires higher computational resources.

Table 7 presents an ablation study on XNet-OBB-XLarge, examining the effects of loss weighting factors and different training data configurations. For the loss weighting factors, the baseline configuration ($\lambda_{cls} = 0.5$, $\lambda_{box} = 7.5$, $\lambda_{dfl} = 1.5$) achieves the best overall balance between precision, recall, and detection accuracy. Increasing λ_{cls} or decreasing λ_{box} tends to improve precision but leads to a noticeable drop in recall, suggesting that excessive emphasis on classification may suppress positive detections. In terms of detection metrics, mAP_{50} remains relatively insensitive to variations in loss weighting factors, indicating stable object detection capability across different configurations. In contrast, $mAP_{50\text{-}95}$ is more sensitive to the balance between localization and classification. Specifically, when the ratio of λ_{box} to λ_{cls} becomes

Table 6
Comparison between XNet-OBB-XLarge and other state-of-the-art OBB models.

Model	Precision	Recall	mAP_{50}	mAP_{50-95}	FPS↑	FLOPs(G)↓	Params(M)↓
Oriented R-CNN	0.925	0.919	0.945	0.821	86.886	198.530	41.352
S^2A-Net	0.962	0.933	0.972	0.850	92.161	196.310	38.765
ReDet	0.967	0.942	0.978	0.861	**95.093**	**54.360**	**31.573**
BBAVectors	0.921	0.916	0.940	0.818	46.153	578.493	72.421
XNet-OBB	**0.970**	**0.955**	**0.980**	**0.893**	60.632	518.934	58.738

Table 7
Ablation study of XNet-OBB-XLarge.

	Precision	Recall	mAP_{50}	mAP_{50-95}
loss weighting factors (λ_{cls}, λ_{box}, λ_{dfl})				
(0.25, 7.5, 1.5)	0.967	0.974	0.980	0.893
(0.5, 7.5, 1.5) [a]	0.970	0.955	0.980	0.893
(1.0, 7.5, 1.5)	0.973	0.948	0.980	0.890
(0.5, 5.0, 1.5)	0.976	0.947	0.980	0.892
(0.5, 10.0, 1.5)	0.968	0.952	0.979	0.890
(0.5, 7.5, 0.5)	0.967	0.950	0.979	0.888
(0.5, 7.5, 2.5)	0.968	0.955	0.982	0.894
Training Set				
MA	0.966	0.943	0.976	0.876
FL	0.954	0.937	0.970	0.854
MA + FL	0.970	0.948	0.979	0.887
MA + FL + WA[a]	0.970	0.955	0.980	0.893

[a] Baseline configuration.

larger, mAP_{50-95} tends to decrease, implying that overemphasizing bounding box regression may negatively affect precise localization under stricter evaluation criteria.

To investigate the cross-state generalization capability of the model, this study conducts an ablation analysis using different training datasets while maintaining the same test set, as shown in Table 1. In addition to the original training dataset that combines data from Massachusetts (MA), Florida (FL), and Washington (WA), three comparison settings are considered, i.e., MA, FL, and MA + FL.

As shown in Table 7, models trained on a single state exhibit limited generalization capability. For example, training on MA achieves relatively strong performance ($mAP_{50} = 0.976$, $mAP_{50-95} = 0.876$), whereas training on FL results in lower accuracy ($mAP_{50} = 0.970$, $mAP_{50-95} = 0.854$), likely due to differences in data distribution and the dominance of MA samples in the overall dataset.

The MA + FL configuration increases mAP_{50} to 0.979 and $mAP_{50–95}$ to 0.887, demonstrating enhanced robustness through exposure to more diverse training samples. Further incorporating WA data (MA + FL + WA) leads to the best overall performance ($mAP_{50} = 0.980$, $mAP_{50–95} = 0.893$), indicating that cross-state training improves model generalization across heterogeneous environments, while the marginal gains are diminishing as the dataset increases.

5.1.3. Demonstration of OBB detection

Fig. 5 illustrates examples of OBB detections based on the XLarge model, featuring both stripe and parallel types of crosswalks. To enhance clarity and detail, the images are selectively cropped to closely focus on the crosswalk areas. Notably, the model effectively detects crosswalks located in shadowed areas, such as those cast by buildings, and accurately identifies crosswalks adjacent to no parking hash markings. The model demonstrates versatility by accurately detecting crosswalks of varying orientations and lengths across diverse environmental settings, further illustrating its robustness and adaptability to complex urban landscapes.

As illustrated in Fig. 6, crosswalks are detected under challenging conditions, including building shadows, partial occlusions from trees, and faded markings. This performance is supported by a large and diverse training dataset that covers a wide range of crosswalk appearances under varying occlusion, wear, and environmental conditions.

Parking lots are identified as an additional challenge due to the presence of various markings, such as no-parking hash lines, lane dividers, and directional arrows, some of which visually resemble pedestrian crosswalks. These similarities can lead to misclassifications; however, as shown in Fig. 7, XNet is able to successfully differentiate crosswalks from other pavement markings, ensuring high detection accuracy even in complex environments.

To evaluate the model's generalization capability, it is tested on unseen datasets from other states, including Maine, New Hampshire, and Virginia, which are not included in the training or validation sets. 2025 Massachusetts aerial imagery is also used while the model is trained on 2023 Massachusetts aerial imagery. Preliminary results (Figs. 8 and 9) indicate that the model performs well across most scenarios. However, failures occur in cases where crosswalks are completely obscured by shadows. Fig. 9 presents an example of such a case. This limitation suggests that the model robustness could be further improved by incorporating additional data sources or multi-year aerial imagery when generating crosswalk inventory.

5.2. Semantic segmentation

5.2.1. Overall model comparison

All segmentation models are trained for 100 epochs with a batch size of 2 on the same hardware platform as the OBB experiments, using a single GPU. During the training process, all the models adopt the AdamW optimizer with an initial learning rate of 0.0001, and a weight decay of 0.01. All evaluations are conducted on the same GPU, and identical hardware and experimental settings are maintained across all models to ensure fair and consistent comparison.

Table 8 presents a comprehensive comparison between the proposed XNet-Seg and several state-of-the-art segmentation models on the validation set. Among all compared methods, XNet-Seg achieves the best overall performance, reaching an mIoU of 0.925 and an mDice score of 0.960, while maintaining a well-balanced precision (0.962) and recall (0.957). This suggests that the proposed framework is able to simultaneously achieve accurate boundary localization and complete region coverage.

In contrast, the baseline CNN achieves the highest inference speed (200.905 FPS) and the smallest model size (0.186M parameters), demonstrating strong computational efficiency. However, it yields the lowest segmentation accuracy (mIoU = 0.864, mDice = 0.922), reflecting the limitations of shallow architectures in capturing multi-scale contextual information. PSPNet and U-Net increase mIoU to approximately 0.915-0.918, probably benefiting from pyramid pooling and skip connections that enhance global context and spatial detail recovery, albeit at the cost of reduced inference speed. DeepLabv3 and DeepLabv3+ achieve slightly higher accuracy (mIoU = 0.918 and 0.920), demonstrating the effectiveness of atrous convolutions, though they require substantially higher computational cost (>650 GFLOPs).

More advanced architectures further illustrate the trade-off between accuracy and efficiency. HRNet maintains relatively strong performance with moderate computational cost, benefiting from high-resolution feature representations. Swin-Unet achieves competitive recall (0.958) but does not significantly outperform other models in overall accuracy.



Fig. 5. Sample outputs of crosswalks in different orientations.

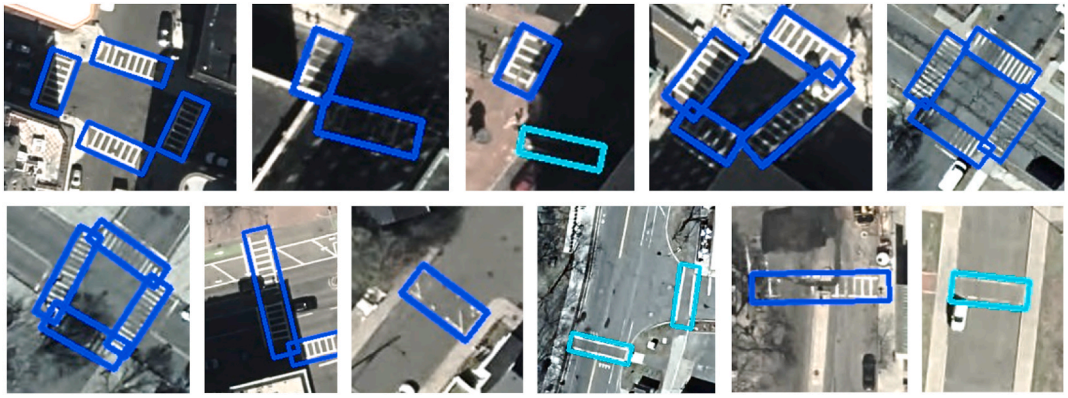


Fig. 6. Examples of crosswalks detected in shaded, occluded conditions, or with faded markings.

Transformer-based models exhibit distinct characteristics, i.e., SegFormer achieves the lowest computational cost (27.074 GFLOPs) and relatively high FPS, but sacrifices accuracy (mIoU = 0.903), while Mask2Former shows limited performance gains (mIoU = 0.908) despite its more complex design and lower inference speed (13.811 FPS), suggesting that general-purpose architectures may not fully capture the

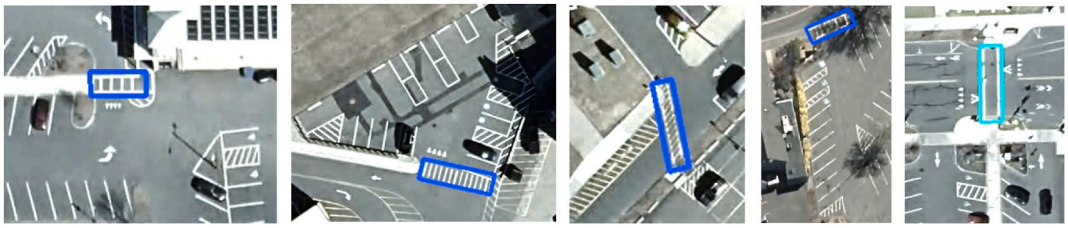


Fig. 7. Sample detections in parking lots.

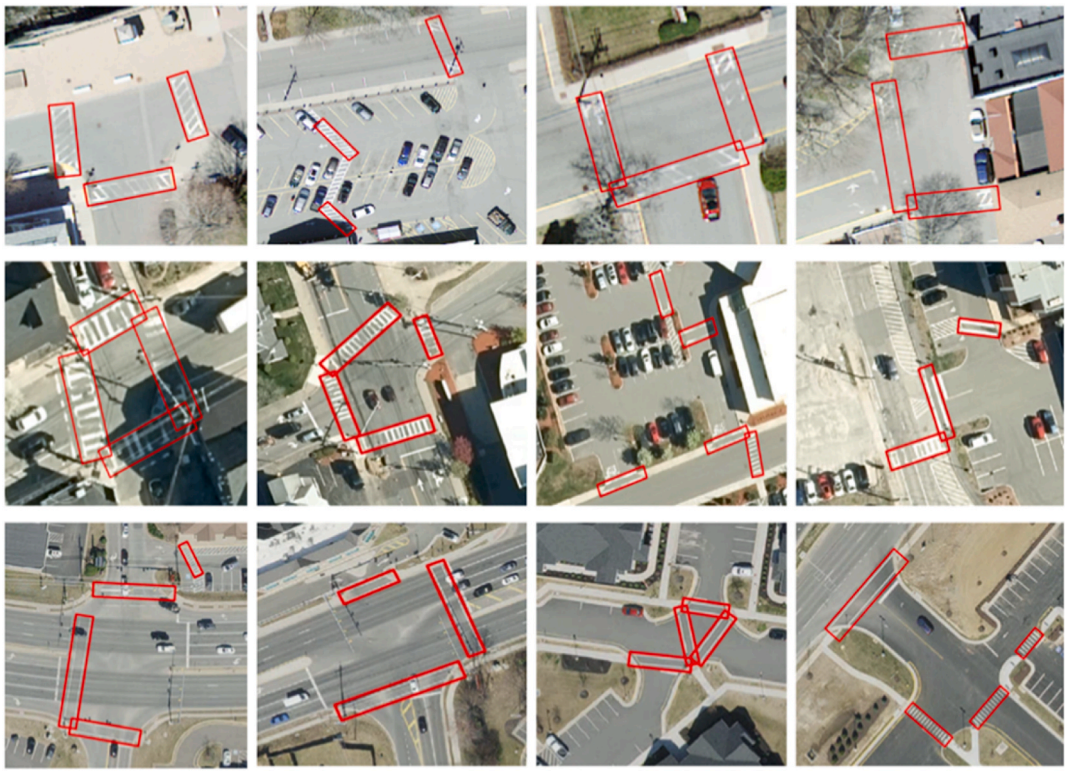


Fig. 8. Examples of detections on image data from other states (top: Maine; middle: New Hampshire; bottom: Virginia).

task-specific structural patterns of crosswalks.

Compared with all baseline and state-of-the-art models, XNet-Seg achieves a favorable balance between accuracy and efficiency. While its computational cost (732.675 GFLOPs) is comparable to other high-performing CNN-based models, it maintains a relatively high inference speed (78.510 FPS), significantly outperforming transformer-based methods in efficiency. Although the absolute accuracy gains are moderate, the improvements are consistent across all evaluation metrics, indicating the robustness of the proposed design. In particular, the MSA mechanism enhances the model's ability to capture fine-grained stripe patterns while preserving global structural consistency. This leads to improved performance under challenging conditions such as partial occlusion, marking degradation, and varying illumination, making XNet-Seg more suitable for large-scale, real-world deployment scenarios.

Figs. 10 and 11 present segmentation results for representative stripe and parallel crosswalk examples, where each column shows the input image, ground truth, and predictions from different models. For stripe crosswalks, the CNN baseline exhibits the weakest performance, producing blurred or merged stripe patterns, missing segments, and introducing false positives by confusing background elements such as sidewalks or shaded regions with crosswalk markings. Its predictions further degrade under shadow or illumination variations. PSPNet and U-Net improve the overall shape reconstruction under clean backgrounds but tend to under-segment when visually similar markings are present, failing to fully recover stripe continuity. DeepLabv3 and DeepLabv3+ show more stable behavior, benefiting from stronger contextual modeling, yet they still miss stripe elements near boundaries or in low-contrast regions. HRNet preserves spatial resolution more effectively and produces sharper structures, but can still be distracted by repetitive background patterns. Transformer-based models such as SegFormer and Mask2Former demonstrate improved global context understanding and generally produce smoother and more coherent predictions; however, they occasionally oversmooth fine-grained stripe details or fail to preserve precise spacing between adjacent bars, especially when the stripes are narrow or partially occluded.

In contrast, XNet-Seg produces predictions that are consistently closer to the ground truth across varying conditions. It successfully recovers nearly all stripe elements even in cluttered backgrounds, while maintaining accurate spacing, orientation, and structural continuity. The model also preserves sharper boundaries under challenging conditions such as shadows, illumination changes, and rotation.

For parallel crosswalks, a similar trend is observed. The CNN baseline again performs the worst, frequently introducing false positives and failing to accurately delineate bar structures. PSPNet and U-Net provide more stable predictions but may still confuse adjacent linear features.

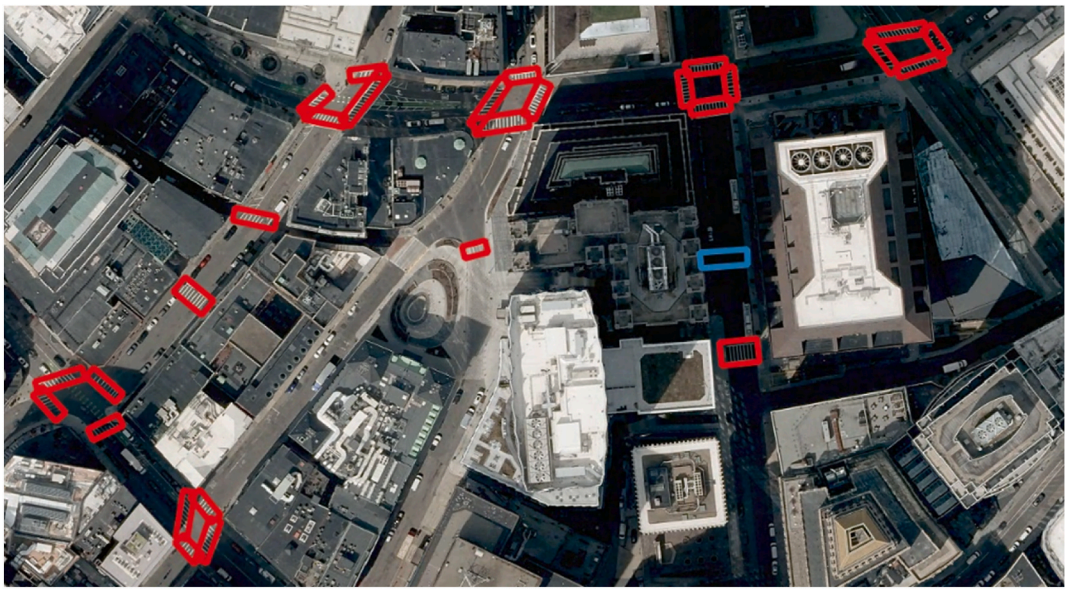


Fig. 9. Examples of failure cases on image data (blue box: FN; red box: TP). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 8
Performance comparison of different segmentation models.

Model	Pr	Re	mIoU	mDice	FPS↑	FLOPs(G)↓	Params(M)↓
CNN	0.927	0.917	0.864	0.922	**200.905**	50.835	**0.186**
PSP-Net	0.961	0.947	0.915	0.954	95.902	435.137	11.134
U-Net	0.961	0.949	0.918	0.955	63.702	772.792	31.038
DeepLabv3	0.959	0.952	0.918	0.956	88.441	656.476	39.634
DeepLabv3+	0.955	0.952	0.920	0.957	81.504	730.421	40.413
Mask2Former	0.959	0.940	0.908	0.949	13.811	237.231	44.954
HRNet	0.950	0.951	0.910	0.951	68.685	99.564	6.562
SegFormer	0.936	0.957	0.903	0.947	76.276	**27.074**	3.715
Swin-Unet	0.951	**0.958**	0.917	0.955	50.366	133.008	25.904
XNet-Seg	**0.962**	0.957	**0.925**	**0.960**	78.510	732.675	40.420

DeepLabv3 series and HRNet achieve reliable segmentation of the main bar structures, while SegFormer and Mask2Former further improve robustness to background complexity through global attention mechanisms. Nevertheless, these transformer-based models can occasionally produce slightly diffused boundaries or minor misalignment in bar edges. In comparison, XNet-Seg maintains clearer boundaries and more precise geometric consistency, particularly in complex scenes or under partial occlusion, leading to the most visually accurate and stable segmentation results among all evaluated methods.

5.2.2. Multi-factor ablation study

To further investigate the contributions of different components, a multi-factor ablation study is conducted by independently varying the attention mechanism, backbone architecture, and dataset, while keeping other components fixed.

As shown in Table 9, the performance differences among the attention modules are relatively small, with all methods achieving similar accuracy (mIoU around 0.923-0.925 and mDice around 0.958-0.960). However, the proposed MSA demonstrates superior efficiency, achieving the highest inference speed (78.510 FPS) while maintaining comparable computational cost (732.675 GFLOPs) and parameter size (40.420M). This indicates that MSA provides a more favorable accuracy-efficiency trade-off compared to other attention modules, making it more suitable for practical deployment.

For the backbone architecture, replacing ResNet50 with ResNet101 only leads to a slight improvement in segmentation accuracy (e.g., Recall from 0.957 to 0.963, mIoU from 0.925 to 0.926). However, this marginal gain comes at a substantial computational cost, with FLOPs increasing significantly from 732.675G to 1045.998G and FPS dropping noticeably from 78.510 to 48.576. This reflects a typical trade-off in deep learning models, indicating that deeper architectures may offer better representational capacity but reduce inference efficiency.

Besides, this study also examines the impact of training data by comparing full images with OBB-based blocks. The results show that using full images leads to a drop in accuracy (mIoU = 0.910, mDice = 0.949), whereas training on OBB-based blocks achieves higher performance (mIoU = 0.925, mDice = 0.960). Moreover, using OBB-based blocks can accelerate the segmentation process, with the FPS increasing from 77.468 to 78.510. These results highlight the effectiveness of the proposed region-focused training strategy. By removing large amounts of irrelevant background and focusing on crosswalk regions, the model is able to learn more discriminative features and alleviate class imbalance.

5.3. GIS inventory construction and condition assessment

5.3.1. Model inference and GIS inventory construction

As discussed in the previous section, XNet-OBB-XLarge and XNet-Seg achieve the best performance for crosswalk detection and segmentation, respectively. Therefore, large-scale inference is conducted on a Massachusetts statewide dataset consisting of 10,218 high-resolution aerial images, each with a spatial resolution of 10,000×10,000 pixels.

All experiments are performed on the same hardware platform used for model training, equipped with an AMD Ryzen Threadripper PRO

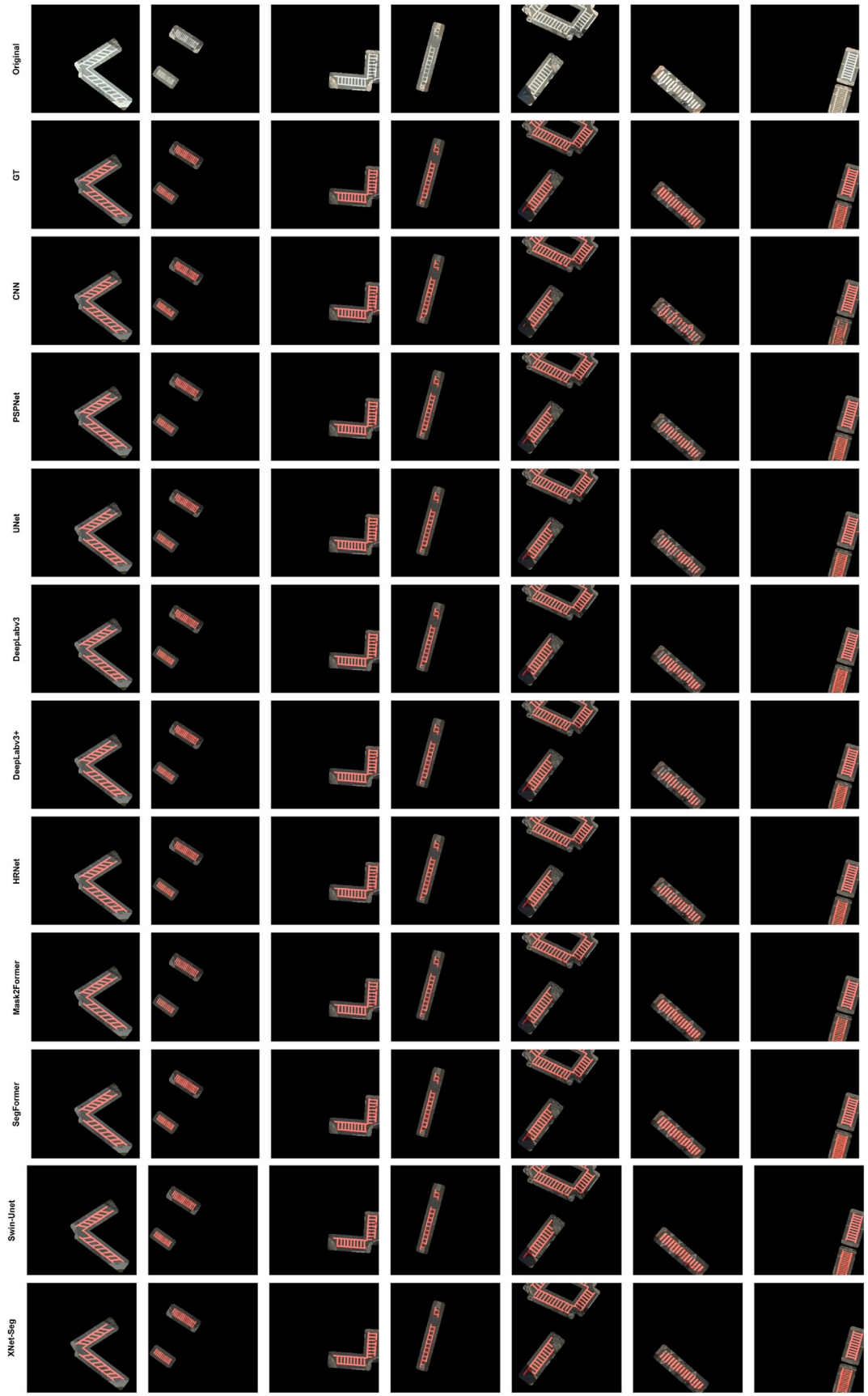


Fig. 10. Segmentation results of the proposed XNet-Seg and other methods on stripe crosswalks.

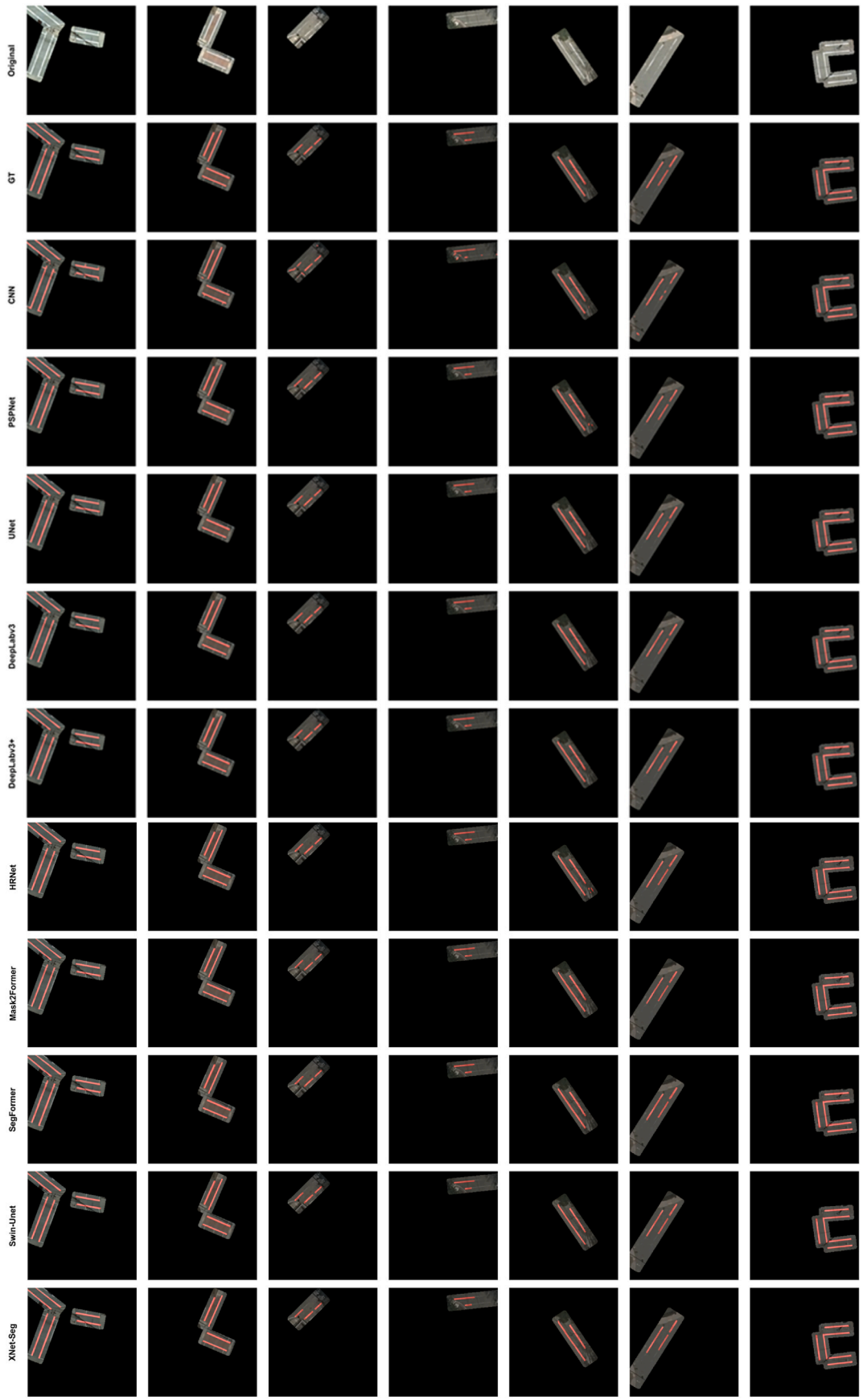


Fig. 11. Segmentation results of the proposed XNet-Seg and other methods on parallel crosswalks.

Table 9
Multi-factor ablation study of XNet-Seg.

	Pr	Re	mIoU	mDice	FPS↑	FLOPs(G)↓	Params(M)↓
Attention							
CBAM	0.959	0.958	**0.925**	**0.960**	77.695	732.873	40.430
DANet	0.957	**0.960**	**0.925**	0.958	76.020	741.540	40.626
ECA	0.958	0.957	**0.925**	0.959	78.150	732.866	40.425
SKNet	0.957	0.956	0.923	0.959	75.063	879.623	42.667
MSA	**0.962**	0.957	**0.925**	**0.960**	**78.510**	**732.675**	**40.420**
Backbone							
ResNet50	**0.962**	0.957	0.925	**0.960**	**78.510**	**732.675**	**40.420**
ResNet101	0.958	**0.963**	**0.926**	**0.960**	48.576	1045.998	59.462
Dataset							
Full image	0.959	0.943	0.910	0.949	77.468	732.675	40.420
OBB blocks	**0.962**	**0.957**	**0.925**	**0.960**	**78.510**	**732.675**	**40.420**

7985WX CPU (64 cores, 128 threads). The PyTorch implementation is executed on a single NVIDIA RTX PRO 6000 Blackwell GPU with 95 GB memory. The total inference time for processing the entire dataset is approximately 15.4 h, corresponding to a throughput of about 664 images per hour (each of size 10,000×10,000). The peak GPU memory allocation is around 2.9 GB, corresponding to approximately 3% of the total available GPU memory. This indicates that the framework maintains low memory overhead despite the extremely high image resolution. These results demonstrate that the proposed framework is capable of handling large-scale, high-resolution aerial imagery with practical computational efficiency.

The efficiency gain of the proposed two-stage framework mainly stems from a detection-guided processing strategy. Specifically, the OBB detection stage is first applied to all images, while the more computationally intensive segmentation stage is selectively performed only on regions of interest associated with detected crosswalks. In practice, only 3,928 out of the 10,218 images proceed to the segmentation stage, as a large portion of the dataset corresponds to rural, forested, or mountainous areas where no crosswalks are present. To further quantify this benefit, a comparative experiment is performed without detection-guided filtering, where segmentation is applied to all images. In this case, the total inference time increases to approximately 27.3 h, compared to 15.4 h with the proposed pipeline, corresponding to about a 77.27% increase in computational cost.

In addition, a GIS inventory is constructed by transforming the raster-based detection and segmentation outputs into vector-based GIS layers containing over 100,000 crosswalks (as shown in Fig. 12). Specifically, the predicted oriented bounding boxes are first converted from image coordinates to geographic coordinates using the georeferencing information of the aerial imagery and then represented as polygon features. Meanwhile, the segmentation results are used to identify valid marking pixels within each detected OBB, from which the image-based condition assessment indicators are computed. The entire pipeline is implemented using Python-based geospatial processing libraries, including GDAL, rasterio, and geopandas, enabling efficient large-scale processing and seamless integration with standard GIS platforms.

5.3.2. Comparison of RCR and NCR

Figs. 13 and 14 illustrate the large-scale distribution of the RCRs computed using the procedure described in Section 4.2.3. To evaluate the robustness of the condition-estimation framework under realistic conditions, forty 10,000×10,000 aerial images are randomly selected from two mapping zones in Massachusetts. Twenty images are drawn from the western region covered by NAD 1983 UTM Zone 18N (Fig. 13), and another twenty from the adjacent region covered by NAD 1983 UTM Zone 19N (Fig. 14). Each aerial image contains more than one hundred crosswalks, offering substantial variation in pavement conditions, illumination, background clutter, and marking degradation that commonly occur across the state.

It can be observed that across both Zone 18N and Zone 19N, stripe crosswalks consistently exhibit greater variability in RCRs, as reflected by wider violin shapes and more dispersed scatter points. This pattern

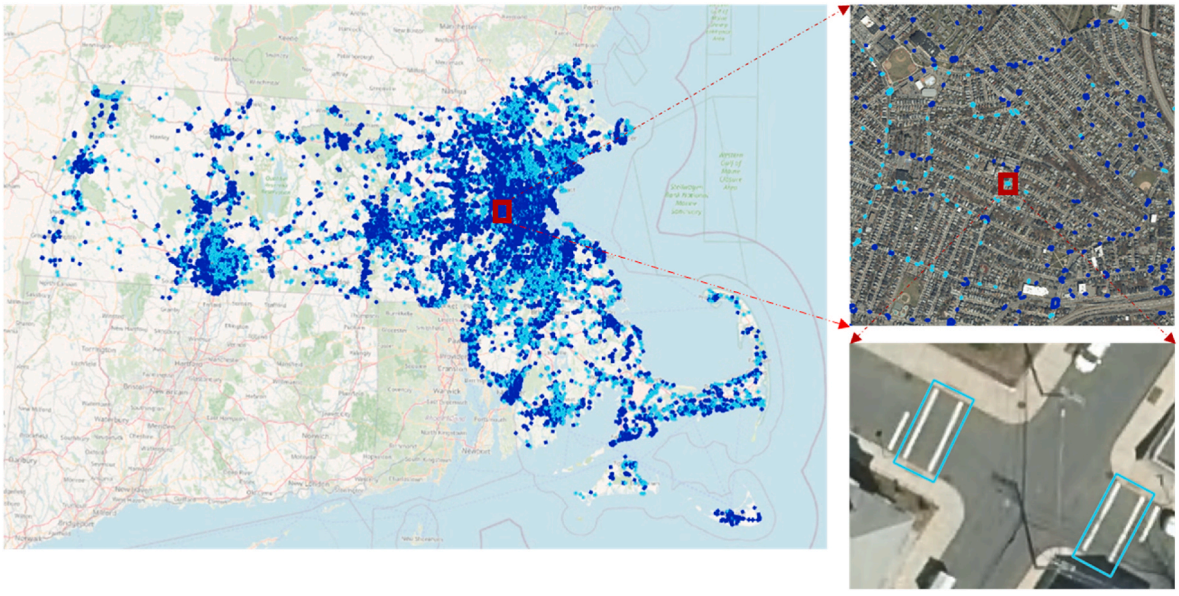


Fig. 12. Statewide crosswalk inventory in Massachusetts with representative local examples.

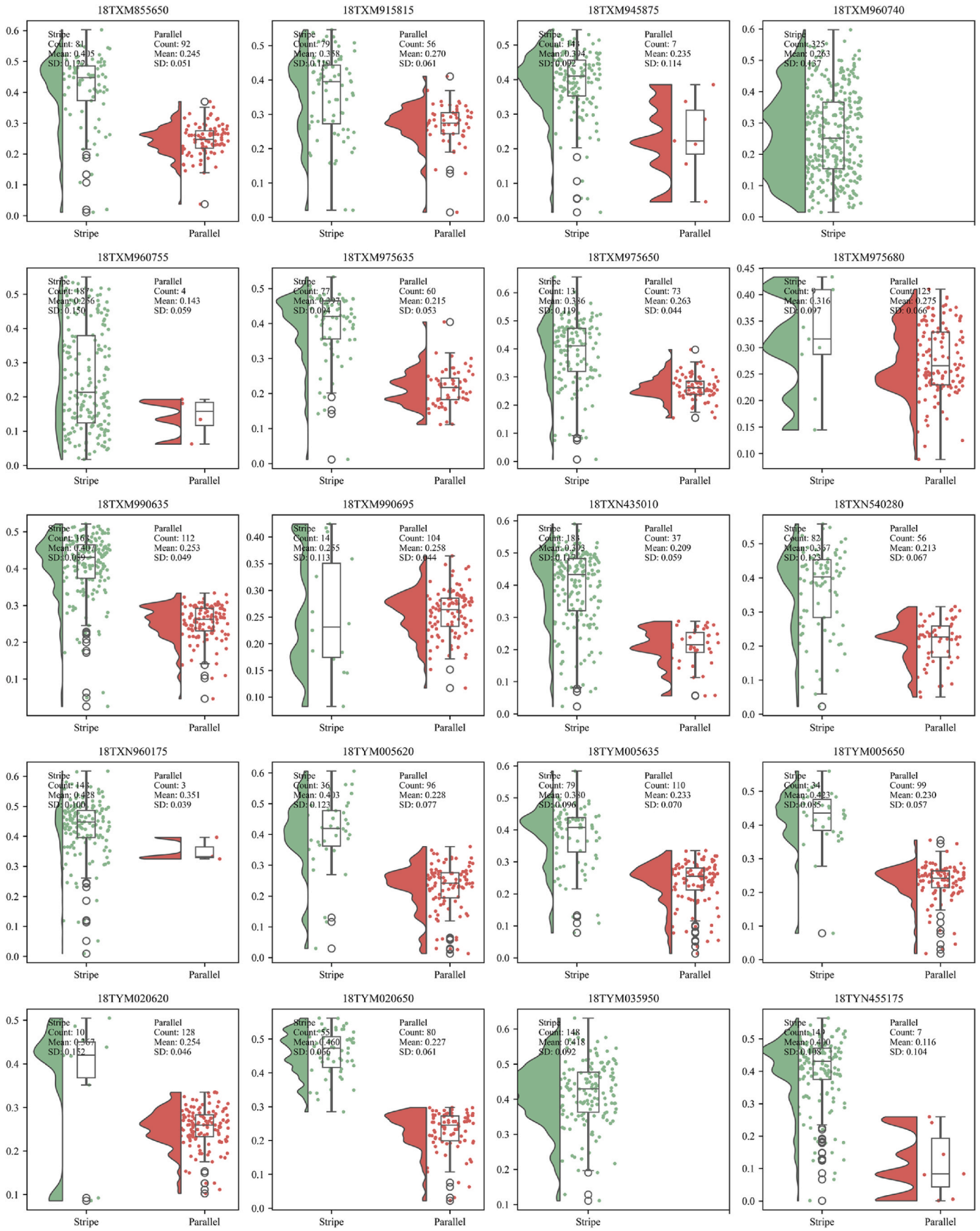


Fig. 13. RCR results for examples from western Massachusetts.

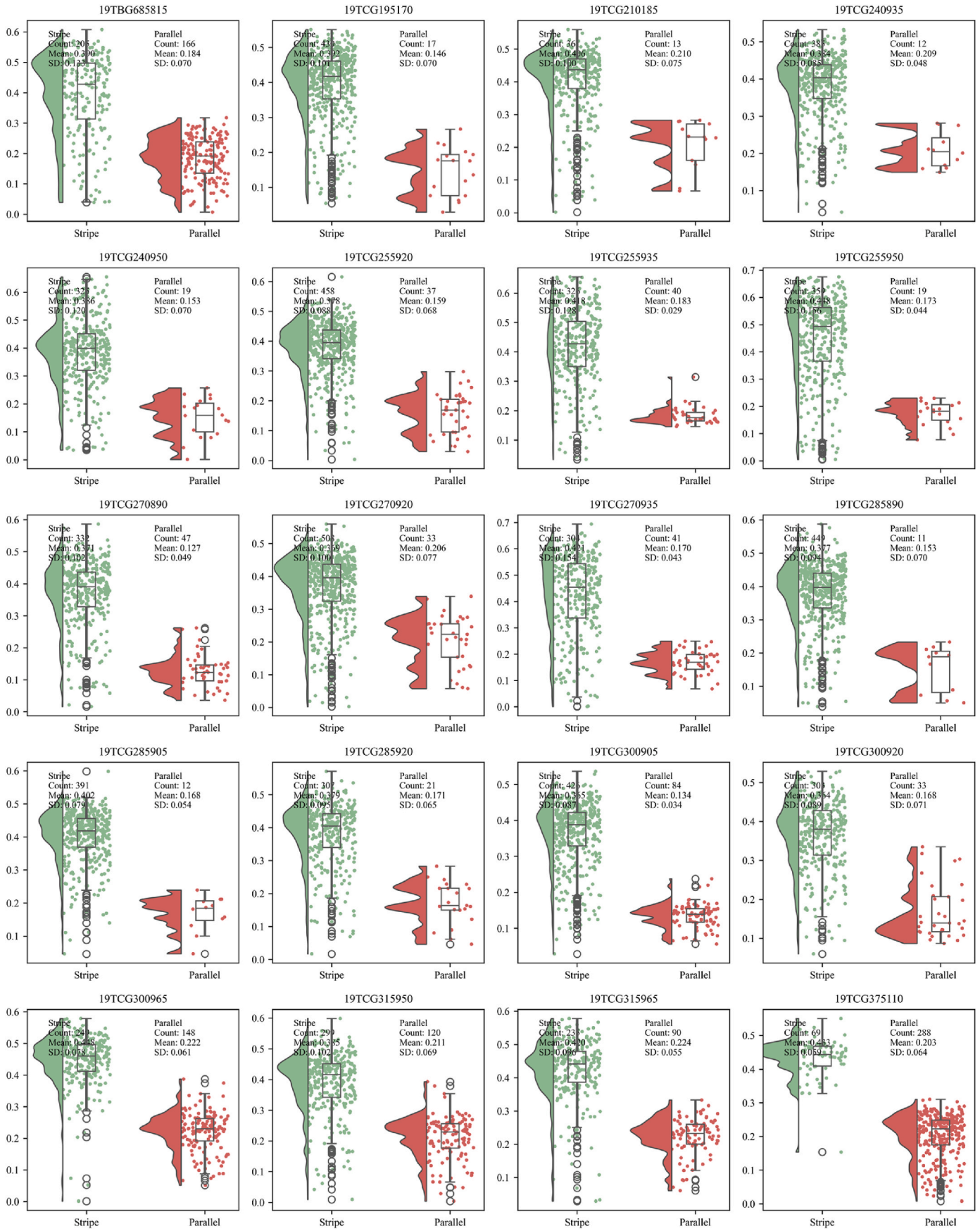


Fig. 14. RCR results for examples from eastern Massachusetts.

arises from the structural complexity of stripe crosswalks. Several images (such as 18TXM990755 in Figs. 13 and 19TCG255920 in Fig. 14) display stripe-crosswalk distributions with two or more distinct peaks, reflecting the simultaneous presence of both highly preserved and severely degraded markings.

With regard to the parallel crosswalks, they tend to display narrower and more concentrated RCR distributions, typically forming shorter box plots and compact violin shapes. One plausible explanation for this stability lies in their simpler geometric structure, as each parallel crosswalk contains only two long bars. This indicates that even for intact markings, the RCRs of parallel crosswalks are systematically lower than those of stripe crosswalks, because the 2 bars occupy a smaller proportion of the OBB area than the dense stripe patterns. As shown in Figs. 13 and 14, the RCRs of parallel crosswalks (red distributions) usually fall within a relatively tight band (often between 0.15 and 0.30). In contrast, the stripe crosswalks (green distributions) occupy a much wider range of RCRs in almost every subplot. Many images show dense clusters of stripe ratios (approximately 0.30-0.60), along with extended tails that capture degraded or partially faded stripes. As further illustrated in Fig. 15, the difference in RCRs is evident even when the markings are visually intact. In the Zone 18 example (18TYM020650), the stripe crosswalk located in a parking lot yields an RCR of approximately 0.562, reflecting the dense coverage of its multiple bars within the OBB. In contrast, the adjacent parallel crosswalk in the same aerial image remains at a much lower RCR of about 0.291, despite being clearly preserved with minimal fading. A similar pattern also appears in the Zone 19 example (19TCG315965). That is, the intact stripe crosswalk produces an RCR of roughly 0.578, whereas the intact parallel crosswalk remains around 0.292. These paired examples confirm that the lower RCRs of parallel crosswalks do not necessarily indicate poorer condition; instead, they arise from the intrinsic geometric composition of the pattern, where two long bars occupy a substantially smaller portion of the bounding box than the dense stripe arrangement.

These geometric differences further highlight the necessity of applying within-class ratio normalization (as described in Section 4.2.3) in condition assessment. Without such normalization, the RCRs of stripe and parallel crosswalks are not directly comparable, since their baseline values are driven by distinct fill ratios rather than actual deterioration levels. Standardizing each crosswalk type ensures that the resulting condition measures remain comparable and are not biased by these intrinsic differences in marking geometry and coverage.

Fig. 16 presents a side-by-side comparison of the RCR distributions and NCR distributions. After the normalization, the condition ratios for both stripe and parallel crosswalks fall within similar numerical ranges, and their central tendencies become more closely aligned. Specifically, the maximum condition ratio for parallel crosswalks increases from approximately 0.4 (RCR) to above 0.9 (NCR), and the differences in mean values between the 18 Zone (western Massachusetts) and the 19 Zone (eastern Massachusetts) are substantially reduced. This shift indicates that normalization effectively removes geometry-driven disparities and places both crosswalk types onto a common interpretative scale. Consequently, well-preserved parallel crosswalks are no longer penalized by their naturally low fill ratio, and severely faded stripe crosswalks are no longer overshadowed by the higher baseline inherent to their geometry.

5.3.3. Practical implications of NCR

To support practical condition assessment, a data-driven classification scheme is established based on manually labeled crosswalk samples. Specifically, for both stripe and parallel crosswalk types, 400 samples are manually annotated for three condition levels, including good, medium, and poor, resulting in a total of 2400 labeled samples.

The annotation criteria are defined based on visual inspection: “good” corresponds to crosswalk markings with no visible wear, “medium” indicates slight degradation, and “poor” represents severe deterioration or near disappearance of markings. Representative examples of each category are illustrated in Figs. 17 and 18 for stripe and parallel crosswalks, respectively.

To derive objective and reproducible classification thresholds, a data-driven approach based on Gaussian Mixture Modeling (GMM) is adopted (Scikit-learn developers, n.d.). For each crosswalk type, the NCR values are modeled as a mixture of three Gaussian components, corresponding to the latent condition levels. To ensure a fair evaluation and avoid information leakage, the labeled samples are randomly divided into training and validation subsets using an 80/20 split with stratification by condition labels. The GMM is fitted exclusively on the training subset, and the thresholds separating adjacent condition levels are determined as the midpoints between the sorted component means. The NCR distribution of the three condition levels and the determined thresholds are shown in Fig. 19. For stripe crosswalks, the thresholds for different condition levels are determined as: poor (NCR$<$0.196), medium (0.196$\leq$NCR$<$0.454), good (NCR$\geq$0.454). For parallel crosswalks, the corresponding thresholds are: poor (NCR$<$0.156), medium (0.156$\leq$NCR$<$0.332), good (NCR$\geq$0.332).

The learned thresholds are then applied to classify the held-out validation samples. Specifically, each sample is assigned to a condition level (poor, medium, or good) based on its NCR value relative to the two thresholds. The predicted labels are compared against the manually assigned ground truth labels using standard classification metrics, including precision, recall, and F1-score, and the results are shown in Table 10. It can be observed that the proposed method achieves consistently high accuracy across all condition levels, with precision, recall, and F1-scores generally exceeding 0.95. This indicates that the GMM-derived thresholds effectively capture the underlying distribution of NCR values and align well with the manually assigned condition labels.

Among the three condition levels, the classification performance for the “good” and “poor” categories is particularly strong, with F1-scores reaching up to 0.981. This can be attributed to the clear separation of their NCR distributions, as observed in Fig. 19. In contrast, the “medium” category exhibits slightly lower performance, with F1-scores around 0.94-0.96. This is expected, as moderate samples lie in the transitional region between good and poor conditions, leading to partial overlap in NCR values and increased classification ambiguity.

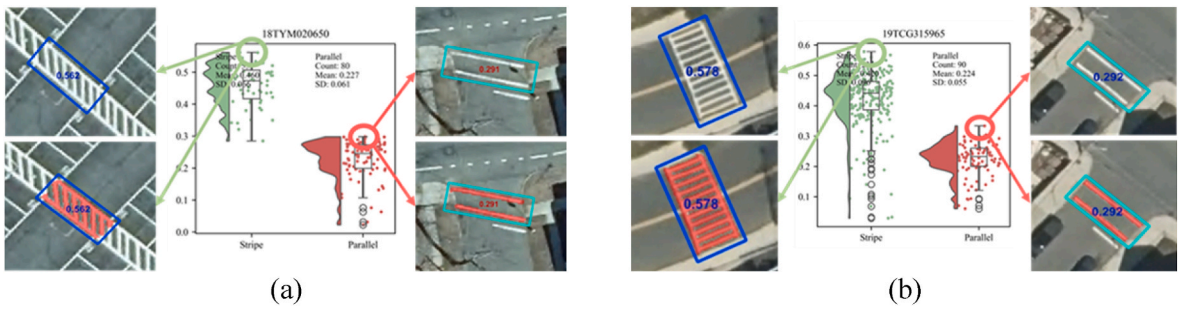


Fig. 15. Examples from aerial images: (a) 18TYM020650 and (b) 19TCG315965.

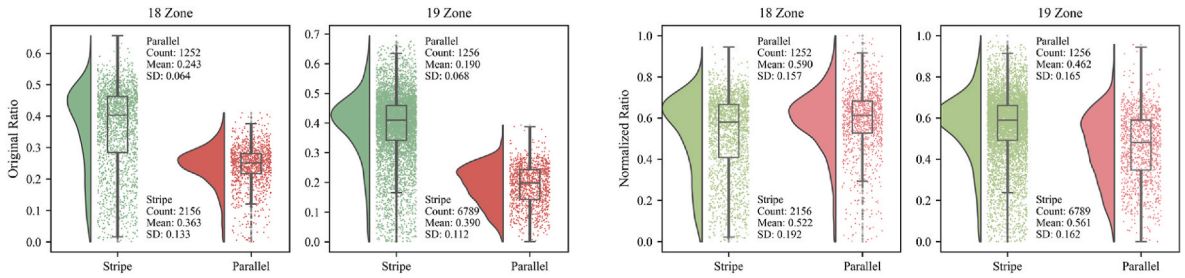


Fig. 16. RCR versus NCR.



Fig. 17. Examples of condition GT for the stripe crosswalks: (a) good, (b) medium and (c) poor.



Fig. 18. Examples of condition GT for the parallel crosswalks: (a) good, (b) medium and (c) poor.

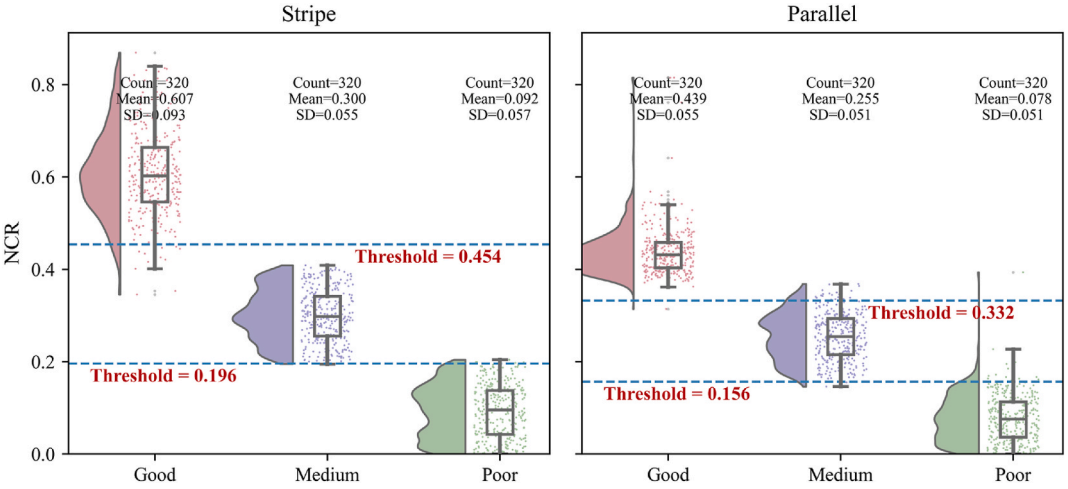


Fig. 19. Distribution of NCR across condition levels with GMM-derived thresholds (training set).

Table 11 further presents an ablation study comparing RCR and NCR. While RCR achieves high precision in certain classes, it shows clear imbalance with low recall for "Good" (0.450) and weak performance in the "Medium" category. In contrast, NCR delivers consistently high and

Table 10
Performance of the proposed threshold on the validation set (NCR).

	Precision			Recall			F1		
	Good	Medium	Poor	Good	Medium	Poor	Good	Medium	Poor
All	0.969	0.956	0.987	0.981	0.956	0.975	0.975	0.956	0.981
Stripe	1.000	0.941	0.987	0.963	0.988	0.975	0.981	0.963	0.981
Parallel	0.941	0.974	0.987	1.000	0.925	0.975	0.970	0.949	0.981

Table 11
Ablation study of RCR and NCR.

	Precision			Recall			F1		
	Good	Medium	Poor	Good	Medium	Poor	Good	Medium	Poor
NCR	0.969	0.956	0.987	0.981	0.956	0.975	0.975	0.956	0.981
RCR	1.000	0.567	0.780	0.450	0.719	1.000	0.621	0.634	0.877

balanced performance across all classes, demonstrating its superiority for robust condition assessment.

Building upon the above findings, the proposed framework provides a practical reference for real-world applications. For example, the derived NCR thresholds and corresponding condition categories can support large-scale screening of crosswalk visible conditions. In particular, crosswalks classified as poor or exhibiting lower NCR values can be efficiently identified for further inspection and prioritized maintenance. This enables scalable, consistent, and cost-effective decision-making for transportation agencies.

Despite the promising results, it should be noted that the condition assessment proposed in this study should be interpreted as an image-based indicator of marking visibility rather than a comprehensive physical measure of crosswalk condition. In particular, it does not capture factors such as retroreflectivity, material properties, or nighttime visibility. This limitation primarily stems from the difficulty of acquiring large-scale, consistently collected field-level condition data, which is often labor-intensive, costly, and challenging to standardize at a statewide level. Nevertheless, the increasing availability of high-resolution sensing technologies, such as LiDAR, provides new opportunities for integrating multi-source data into crosswalk condition assessment. In future work, the proposed framework can be extended by combining aerial imagery with LiDAR data to enable more comprehensive validation of the proposed indicator and to enrich the assessment with additional physical attributes. Such integration is expected to further improve both the reliability and the multidimensionality of crosswalk condition evaluation.

6. Conclusions

To support large-scale crosswalk inventory creation and fine-grained condition assessment from high-resolution aerial imagery, this study develops XNet, a two-stage detection and segmentation framework which integrates an anchor-free OBB detector with a tailored segmentation model.

Trained on aerial imagery from Massachusetts, Washington, and Florida, the proposed OBB detection network effectively captures the geometric complexity of crosswalks by modeling their orientations and elongated structures across diverse roadway environments. Among all model variants, XNet-OBB-XLarge achieves the strongest overall performance, with a precision of 0.970, a recall of 0.955, and an $mAP_{50\text{-}95}$ of 0.893, demonstrating robust localization under varying illumination, shadows, and complex pavement backgrounds. The proposed XNet-Seg outperforms the current state-of-the-art models with the highest mIoU and mDice scores (0.925 and 0.960). Building on these outputs, an image-based condition-assessment pipeline is introduced to quantify crosswalk deterioration, forming an integrated system capable of automated, GIS-ready crosswalk mapping and condition evaluation at scale.

In addition to its strong empirical performance, the study contributes a cohesive framework that links detection, segmentation, and quantitative condition assessment within a single workflow, providing a practical and scalable solution for transportation agencies seeking to modernize crosswalk inventories. The GIS-ready outputs produced by XNet can directly support safety analysis, maintenance prioritization, and accessibility evaluations. Nonetheless, it should be noted that the proposed condition indicator should be interpreted as an image-based proxy rather than a comprehensive physical assessment. Accordingly, the framework is intended as an efficient large-scale screening tool to identify potentially degraded crosswalk markings and support subsequent inspection or multi-sensor evaluation. In practical applications, the proposed system can support crosswalk inventory updates and facilitate maintenance planning by providing network-level screening results. It also offers a scalable foundation for integrating additional sensing modalities or temporal data for more comprehensive condition monitoring in future work.

Some limitations should be noted. The current framework relies on high-resolution aerial imagery and focuses specifically on crosswalks, which may limit its applicability in regions with limited data availability or in broader asset management contexts. In addition, the reliability of the condition indicator is inherently dependent on segmentation performance, and missed or partially detected markings may lead to underestimation of crosswalk condition in certain cases. Future research will extend the framework to additional pedestrian infrastructure elements and incorporate self-supervised or domain-adaptive learning strategies to enhance robustness across diverse geographic and seasonal conditions. Advancing the method toward temporally aware condition tracking may further enable long-term monitoring and predictive maintenance of pedestrian facilities.

CRediT authorship contribution statement

Xintong Yan: Data curation, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Zubin Bhuyan:** Data curation, Formal analysis, Methodology. **Jimi Oke:** Conceptualization, Funding acquisition, Writing – review & editing. **Guanhe Wu:** Data curation, Validation. **Yuanchang Xie:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

No conflict of interest exists in the submission of this manuscript, and all authors for publication approve the manuscript. The authors would like to declare that the work described was original research that has not been published previously, and is not under consideration for

publication elsewhere, in whole or in part. All the authors listed have approved the manuscript that is enclosed.

Acknowledgements

This study was undertaken as part of the Massachusetts Department of Transportation Research Program with funding from the Federal Highway Administration State Planning and Research funds. The authors are solely responsible for the facts, the accuracy of the data and analysis, and the views presented herein.

Special thanks to Kylie Braunius and Bonnie Polin for initiating this idea.

Data availability

The authors do not have permission to share data.

References

Al-haideri, R., Weiss, A., Ismail, K., 2025. A discrete choice latent class method for capturing unobserved heterogeneity in cyclist crossing behaviour at crosswalks. Accid. Anal. Prev. 211, 107850. https://doi.org/10.1016/j.aap.2024.107850.

Berriel, R.F., Rossi, F.S., Souza, A.F. De, Oliveira-santos, T., 2017. Automatic large-scale data acquisition via crowdsourcing for crosswalk classification : a deep learning approach. Comput. Graph. 68, 32–42. https://doi.org/10.1016/j.cag.2017.08.004.

Bian, Y., Liang, K., Zhao, X., Li, H., Yang, L., 2020. Evaluating the effectiveness of new-designed crosswalk markings at intersections in China considering vehicle-pedestrian interaction. Accid. Anal. Prev. 139, 105498. https://doi.org/10.1016/j.aap.2020.105498.

Brkić, I., Ševrović, M., Medak, D., Miler, M., 2023. Utilizing high resolution satellite imagery for automated road infrastructure safety assessments. Sensors 23. https://doi.org/10.3390/s23094405.

Bui, M.D., Lee, J., Choi, K., Kim, H., Kim, C., 2026. Drone-based road marking condition mapping : a drone imaging and geospatial pipeline for asset management. Drones 10, 77. https://doi.org/10.3390/drones10020077.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision. Springer, pp. 205–218.

Cao, Z., Xu, X., Hu, B., Zhou, M., 2021. Rapid detection of blind roads and crosswalks by using a lightweight semantic segmentation network. IEEE Trans. Intell. Transport. Syst. 22, 6188–6197. https://doi.org/10.1109/TITS.2020.2989129.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017a. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40, 834–848. https://doi.org/10.1109/TPAMI.2017.2699184.

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017b. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818.

Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1290–1299.

Cheng, G., Zhou, P., Han, J., 2016. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Trans. Geosci. Rem. Sens. 54, 7405–7415. https://doi.org/10.1109/TGRS.2016.2601622.

Ding, J., Xue, N., Long, Y., Xia, G., Lu, Q., 2018. Learning RoI transformer for detecting oriented objects in aerial images. arXiv preprint arXiv:1812.00155 1–19.

Esmorís, A.M., Vilariño, D.L., Arango, D.F., Cabaleiro, J.C., Rivera, F.F., 2023. Characterizing zebra crossing zones using LiDAR data. Comput. Aided Civ. Infrastruct. Eng. 1767–1788. https://doi.org/10.1111/mice.12968.

Fern, A., Díaz-vilari, L., 2025. A framework for road space extraction from point clouds and integration into 3D city models 143. https://doi.org/10.1016/j.jag.2025.104803.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3146–3154.

Girshick, R., 2015. Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 580–587. https://doi.org/10.1109/CVPR.2014.81.

Guo, W., Xu, C., Jin, S., 2024. Fusion of satellite and street view data for urban traffic accident hotspot identification. Int. J. Appl. Earth Obs. Geoinf. 130, 103853. https://doi.org/10.1016/j.jag.2024.103853.

Han, Y., 2025. Investigating streetscape environmental characteristics associated with road traffic crashes using street view imagery and computer vision. Accid. Anal. Prev. 210, 107851. https://doi.org/10.1016/j.aap.2024.107851.

Han, J., Ding, J., Li, J., Xia, G., 2022. Align deep features for oriented object detection. IEEE Trans. Geosci. Rem. Sens. 60, 1–11. https://doi.org/10.1109/TGRS.2021.3062048.

Haselhoff, A., Kummert, A., 2010. On visual crosswalk detection for driver assistance systems. In: IEEE Intelligent Vehicles Symposium, Proceedings, pp. 883–888. https://doi.org/10.1109/IVS.2010.5548074.

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

He, K., Zhang, X., Ren, S., Sun, J., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 37, 1904–1916.

Hosseini, M., Sevtsuk, A., Miranda, F., Cesar, R.M., Silva, C.T., 2023. Mapping the walk: a scalable computer vision approach for generating sidewalk network datasets from aerial imagery. Comput. Environ. Urban Syst. 101, 101950. https://doi.org/10.1016/j.compenvurbsys.2023.101950.

Kaufmann, S., Kerner, B.S., Rehborn, H., Koller, M., Klenov, S.L., 2018. Aerial observations of moving synchronized flow patterns in over-saturated city traffic. Transport. Res. C Emerg. Technol. 86, 393–406. https://doi.org/10.1016/j.trc.2017.11.024.

Kaya, Ö., Çodur, M.Y., Mustafaraj, E., 2023. Automatic detection of pedestrian crosswalk with faster R-CNN and YOLOv7. Buildings 13, 1–17. https://doi.org/10.3390/buildings13041070.

Kong, W., Zhong, T., Mai, X., Zhang, S., Chen, M., Lv, G., 2022. Automatic detection and assessment of pavement marking defects with street view imagery at the city scale. Remote Sens. 14. https://doi.org/10.3390/rs14164037.

Kortli, Y., Gabsi, S., Voon, L.F.C.L.Y., Jridi, M., Merzougui, M., Atri, M., 2022. Deep embedded hybrid CNN–LSTM network for lane detection on NVIDIA Jetson Xavier NX. Knowl. Base Syst. 240, 107941. https://doi.org/10.1016/j.knosys.2021.107941.

Leventi, H., Habijan, M., Gali, I., 2025. Synthetic and real-world datasets for crosswalk segmentation under diverse weather and lighting conditions 61. https://doi.org/10.1016/j.dib.2025.111755.

Li, C., Feng, G., Li, Y., Liu, R., Miao, Q., Chang, L., 2024. DiffTAD: denoising diffusion probabilistic models for vehicle trajectory anomaly detection. Knowl. Base Syst. 286, 111387. https://doi.org/10.1016/j.knosys.2024.111387.

Li, H., Gong, Y., Zhang, X., Song, Z., Yang, L., Qu, H., 2026. GADet: Geometry-aware oriented object detection for remote sensing. Knowl. Base Syst. 338, 115475. https://doi.org/10.1016/j.knosys.2026.115475.

Li, P., Wang, M., Fan, Z., Huang, H., Zhu, G., Zhuang, J., 2024. OUR-Net: a multi-frequency network with octave max unpooling and octave convolution residual block for pavement crack segmentation. IEEE Trans. Intell. Transport. Syst. 25, 13833–13848. https://doi.org/10.1109/TITS.2024.3405995.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125.

Liu, S., Zhang, L., Lu, H., He, Y., 2021. Center-boundary dual attention for oriented object detection in remote sensing images. IEEE Trans. Geosci. Rem. Sens. 60, 1–14. https://doi.org/10.1109/TGRS.2021.3069056.

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

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965.

Lu, X., Xue, Y., Wang, Z., Xu, H., Wen, X., 2024. X-CDNet : a real-time crosswalk detector based on YOLOX. J. Vis. Commun. Image Represent. 102, 104206. https://doi.org/10.1016/j.jvcir.2024.104206.

Luttrell, J., Zhang, Y., Zhang, C., 2024. Automatically detect crosswalks from satellite view images – a deep learning approach with ground truth verification. Int. J. Transp. Sci. Technol. 1–12. https://doi.org/10.1016/j.ijtst.2024.01.006.

Malbog, M.A., 2019. MASK R-CNN for pedestrian crosswalk detection and instance segmentation. ICETAS 2019 - 2019 6th. In: IEEE International Conference on Engineering, Technologies and Applied Sciences, pp. 2–6. https://doi.org/10.1109/ICETAS48360.2019.9117217.

Moran, M.E., 2022. Where the crosswalk ends: mapping crosswalk coverage via satellite imagery in San Francisco. Environ. Plan. B Urban Anal. City Sci. 49, 2250–2266.

Moran, M.E., Laefer, D.F., 2026. Multiscale analysis of pedestrian crossing distance. J. Am. Plann. Assoc. 92, 290–304. https://doi.org/10.1080/01944363.2024.2394610.

Murali, V.N., Coughlan, J.M., 2013. Smartphone-based crosswalk detection and localization for visually impaired pedestrians. In: Electronic Proceedings of the 2013 IEEE International Conference on Multimedia and Expo Workshops, ICMEW, 2013, pp. 1–7. https://doi.org/10.1109/ICMEW.2013.6618432.

Ning, H., Ye, X., Chen, Z., Liu, T., Cao, T., 2022. Sidewalk extraction using aerial and street view images. Environ. Plan. B Urban Anal. City Sci. 49, 7–22. https://doi.org/10.1177/2399808321995817.

Pantangi, S., Ahmed, S., Fountas, G., Majka, K., Anastasopoulos, P., 2021. Do high visibility crosswalks improve pedestrian safety? A correlated grouped random parameters approach using naturalistic driving study data. Anal. Methods Acc. Res. 30, 100155. https://doi.org/10.1016/j.amar.2020.100155.

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

Riveiro, B., González-jorge, H., Martínez-sánchez, J., Díaz-vilariño, L., Arias, P., 2015. Automatic detection of zebra crossings from mobile LiDAR data. Opt. Laser Technol. 70, 63–70. https://doi.org/10.1016/j.optlastec.2015.01.011.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention — MICCAI 2015. Springer, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

Rosa, G.J.M., Afonso, M.S., Gaspar, P.D., Soares, V.N.G.J., Caldeira, M.L.P., 2024. The development of a prototype solution for detecting wear and tear in pedestrian crossings. Appl. Sci. 14, 6462. https://doi.org/10.3390/app14156462.

Scikit-learn developers, n.d. Gaussian mixture models [WWW Document]. URL https://scikit-learn.org/stable/modules/mixture.html.

Shin, K., Choi, H., 2025. A comprehensive satellite imagery dataset for road marking detection. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 18, 23474–23484. https://doi.org/10.1109/JSTARS.2025.3604411.

Sun, X., Xie, Y., Jiang, L., Cao, Y., 2022. DMA-net : Deeplab with multi-scale attention for pavement crack segmentation. IEEE Trans. Intell. Transport. Syst. 23, 18392–18403. https://doi.org/10.1109/TITS.2022.3158670.

Tian, S., Zheng, M., Zou, W., Li, X., Zhang, L., 2021. Dynamic crosswalk scene understanding for the visually impaired. IEEE Trans. Neural Syst. Rehabil. Eng. 29, 1478–1486. https://doi.org/10.1109/TNSRE.2021.3096379.

Tümen, V., Ergen, B., 2020. Intersections and crosswalk detection using deep learning and image processing techniques. Phys. Stat. Mech. Appl. 543, 123510. https://doi.org/10.1016/j.physa.2019.123510.

Verma, R., Ukkusuri, S.V., 2024. Crosswalk detection from satellite imagery for pedestrian network completion. Transp. Res. Rec. 2678, 845–856. https://doi.org/10.1177/03611981231210545.

Wang, K., Wang, Z., Li, Z., Su, A., Teng, X., Pan, E., 2025. Oriented object detection in optical remote sensing images using deep learning: a survey. arXiv preprint arXiv: 2302.10473. https://doi.org/10.48550/arXiv.2302.10473.

Wang, Q., Chen, B., 2024. A novel transfer learning model for the real-time concrete crack detection. Knowl. Base Syst. 301, 112313. https://doi.org/10.1016/j.knosys.2024.112313.

Wen, L., Cheng, Y., Fang, Y., Li, X., 2023. A comprehensive survey of oriented object detection in remote sensing images. Expert Syst. Appl. 224, 119960. https://doi.org/10.1016/j.eswa.2023.119960.

Wijnands, J.S., Scully, K., Zhao, H., Nice, K.A., Guo, J., Stevenson, M., Thompson, J., 2021. Identifying safe intersection design through unsupervised feature extraction from satellite imagery. Comput. Aided Civ. Infrastruct. Eng. 36, 346–361. https://doi.org/10.1111/mice.12623.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19.

Xiao, D., Yang, X., Li, J., Islam, M., 2020. Attention deep neural network for Lane marking detection. Knowl. Base Syst. 194, 105584. https://doi.org/10.1016/j.knosys.2020.105584.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. 34, 12077–12090.

Xie, K., Sun, H., Dong, X., Yang, H., Yu, H., 2023. Automating intersection marking data collection and condition assessment at scale with an artificial intelligence-powered system. Computational Urban Sci. 3. https://doi.org/10.1007/s43762-023-00098-7.

Xing, L., He, J., Abdel-Aty, M., Cai, Q., Li, Y., Zheng, O., 2019. Examining traffic conflicts of up stream toll plaza area using vehicles' trajectory data. Accid. Anal. Prev. 125, 174–187. https://doi.org/10.1016/j.aap.2019.01.034.

Xu, Q., Long, C., Yu, L., Zhang, C., 2023. Road extraction with satellite images and partial road maps. IEEE Trans. Geosci. Rem. Sens. 61, 1–14. https://doi.org/10.1109/TGRS.2023.3261332.

Yan, X., He, J., Wu, G., Wang, C., Zhang, C., Ye, Y., 2024. Driving risk prediction of urban arterial and collector roads using multi-dimensional real-time data. Eng. Appl. Artif. Intell. 138, 109386. https://doi.org/10.1016/j.engappai.2024.109386.

Yi, J., Wu, P., Liu, B., Huang, Q., Qu, H., Metaxas, D., 2021. Oriented object detection in aerial images with box boundary-aware vectors. In: Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision, pp. 2149–2158. https://doi.org/10.1109/WACV48630.2021.00220. WACV 2021.

Zhai, Y., Cui, G., Gu, Q., Kong, L., 2015. Crosswalk detection based on MSER and ERANSAC. In: IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, pp. 2770–2775. https://doi.org/10.1109/ITSC.2015.448.

Zhang, F., Wang, X., Zhou, S., Wang, Y., 2022. DARDet : a dense anchor-free rotated object detector in aerial images. IEEE Geosci. Remote Sens. Lett. 19, 1–5. https://doi.org/10.1109/LGRS.2021.3122924.

Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z., 2018. Single-shot refinement neural network for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4203–4212.

Zhang, Y., Lu, Z., Zhang, X., Xue, J., Liao, Q., 2022. Deep learning in Lane marking detection : a survey. IEEE Trans. Intell. Transport. Syst. 23, 5976–5992. https://doi.org/10.1109/TITS.2021.3070111.

Zhang, Z., Guo, W., Zhu, S., Yu, W., 2018. Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks. IEEE Geosci. Remote Sens. Lett. 15, 1745–1749. https://doi.org/10.1109/LGRS.2018.2856921.

Zhang, Z. De, Tan, M.L., Lan, Z.C., Liu, H.C., Pei, L., Yu, W.X., 2022. CDNet: a real-time and robust crosswalk detection network on Jetson nano based on YOLOv5. Neural Comput. Appl. 34, 10719–10730. https://doi.org/10.1007/s00521-022-07007-9.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890. https://doi.org/10.1109/CVPR.2017.660.

Zhou, L., Wei, H., Li, H., Zhao, W., Zhang, Yi, Zhang, Yue, 2020. Arbitrary-oriented object detection in remote sensing images based on polar coordinates. IEEE Access 8, 223373–223384. https://doi.org/10.1109/ACCESS.2020.3041025.

Zhu, G., Liu, J., Fan, Z., Yuan, D., Ma, P., Wang, M., Sheng, W., Wang, K.C.P., 2024. A lightweight encoder – decoder network for automatic pavement crack detection. Comput. Aided Civ. Infrastruct. Eng. 39, 1743–1765. https://doi.org/10.1111/mice.13103.

Zhu, G., Shen, S., Yao, J., Wang, M., Zhuang, J., Fan, Z., 2025. Automatic lightweight networks for real-time road crack detection with DPSO. Adv. Eng. Inform. 68, 103610. https://doi.org/10.1016/j.aei.2025.103610.