Research on Vision-Based Recognition Algorithm for Roadside Cross-Field Targets

Liangdong Zuo[1,2,*], Jia Liu[1], Jie Li[3], Hejia Li[1]
[1]*Chongqing College of Architecture And Technology, Chongqing, China*
[2]*Chongqing Research Institute of Shanghai Jiao Tong University , Chongqing, China*
[3]*Chongqing University of Science and Technology, Chongqing, China*
**Corresponding Author*

Abstract: The realization of intelligent transportation systems and vehicle-infrastructure cooperation demands robust road-side perception capabilities. However, vision-based recognition algorithms face significant performance degradation when deployed across different domains—varying geographical locations, weather conditions, lighting environments, and traffic scenarios. This paper investigates recognition algorithms for roadside cross-field targets, addressing the fundamental challenge of domain generalization in visual perception systems. We systematically analyze the limitations of conventional detection methods in cross-domain scenarios, propose a novel framework integrating domain adaptation techniques with multi-scale feature extraction, and present experimental validations using diverse roadside datasets. Our approach achieves substantial improvements in cross-domain recognition accuracy while maintaining real-time performance requirements for roadside deployment. The research con-tributes to the foundational technology for large-scale implementation of vehicle-infrastructure cooperative systems.

Keywords: Cross-Field Target Recognition; Roadside Perception; Domain Adaptation; Computer Vision; Intelligent Transportation Systems

1. Introduction

1.1 Background and Motivation

The paradigm shift towards "vehicle-infrastructure-cloud integration" has positioned roadside perception systems as critical components in the intelligent transportation ecosystem. Unlike onboard vehicular sensors that suffer from inherent blind spots and line-of-sight limitations, roadside vision systems provide comprehensive, bird's-eye-view monitoring of traffic scenes, enabling capabilities ranging from collision avoidance to traffic flow optimization[1-3].

However, the practical deployment of roadside vision systems faces a fundamental challenge: models trained in specific domains often fail when deployed in new environments[4]. A recognition algorithm that performs reliably in sunny Shanghai may exhibit significant degradation when deployed in snowy Harbin or foggy Chongqing. This cross-domain performance gap represents one of the most pressing obstacles to large-scale intelligent transportation implementation[5].

1.2 Definition of Cross-Field Targets

In the context of roadside perception, "cross-field targets" encompasses multiple dimensions of domain variation[6-8]:

1) Geographical domains: Different cities, countries, and road network topologies.
2) Environmental domains: Variations in weather (sunny, rainy, snowy, foggy), illumination (day, night, dawn, dusk), and seasonal conditions.
3) Traffic scenario domains: Highway, urban intersection, rural road, tunnel, and parking lot scenes.
4) Temporal domains: Changes over time in the same location (traffic pattern evolution, infrastructure modifications).

1.3 Research Objectives

This paper aims to address the cross-domain recognition challenge through the following objectives[9]:1) To analyze the mechanisms underlying performance degradation in cross-domain roadside recognition. 2)To develop a novel algorithmic framework that enhances model generalization across domains. 3)To

validate the proposed approach through comprehensive experiments on diverse datasets. 4)To provide practical guidelines for deploying robust roadside perception systems[10].

2. Related Work

2.1 Visual Road Recognition

Visual road recognition has evolved substantially over the past two decades. Conventional approaches followed a sequential pipeline: image preprocessing, feature extraction, and model fitting. Early methods relied on handcrafted features such as Scale-Invariant Feature Transform (SIFT) and Speeded-Up Robust Features (SURF) to distinguish road regions from non-road areas. These approaches typically incorporated geometric cues including horizon detection and vanishing point localization to establish regions of interest for subsequent processing[11].

Horizon detection algorithms, both edge-based and edge-less, have been employed to reduce computational requirements by cropping image frames to relevant portions below the skyline. Vanishing point detection, through methods such as line voting and the Locally Adaptive Self-Voting (LASV) algorithm, provides vehicle localization relative to the road geometry[12].

The advent of deep learning revolutionized this domain. Convolutional Neural Networks (CNNs) demonstrated superior performance in road segmentation and object detection tasks, trained on benchmark datasets including KITTI and Daimler[13]. More recently, transformer-based architectures have emerged, offering enhanced capability to capture global contextual information.

2.2 Cross-Domain Object Detection

The cross-domain object detection problem has garnered increasing research attention. As noted in a comprehensive review, traditional detection models assume that training and test data originate from identical or similar distributions—an assumption rarely satisfied in practice. When deployed across different environments, these models experience substantial accuracy degradation.

Existing approaches to cross-domain detection can be categorized into three primary paradigms:

1) Transfer Learning-Based Methods: These integrate domain adaptation techniques with detection frameworks to enhance model adaptability across environments. Approaches range from aligning feature distributions to minimizing domain discrepancy through adversarial training.

2) Self-Learning Methods: These leverage pseudo-labeling strategies to improve model transferability on target domains. The model iteratively generates predictions on target data and refines itself using high-confidence samples.

3) Image Generation-Based Methods: These employ Generative Adversarial Networks (GANs) to synthesize target-domain images, augmenting training data and improving model performance in the target domain.

Recent work in cross-domain 3D object detection has revealed that most models overfit to training domains, and existing adaptation solutions shift knowledge domains rather than improving fundamental generalization ability. Novel approaches focusing on detection quality for bounding box surfaces and corners closer to sensors have shown promise in enhancing cross-domain robustness.

2.3 Multimodal and Fusion-Based Approaches

The limitations of single-modality perception have motivated research into multimodal fusion. Cross-modal data fusion emerges as a potential solution to enhance perceptual capabilities. Integrating visible and infrared imagery through dual-modality feature fusion and coupled attention mechanisms has demonstrated average accuracy improvements of 30.4% over single-modality visible detection.

Fusion strategies typically operate at multiple levels: early fusion (combining raw sensor data), intermediate fusion (integrating feature representations), and late fusion (merging decision outputs). Transformer-based fusion mechanisms have shown particular effectiveness in capturing cross-modal correlations.

3. Methodology

3.1 Overall Architecture

We propose a Cross-Domain Adaptive Recognition Network (CDARNet) designed specifically for roadside perception requirements. The architecture comprises four main components:

1) Multi-Scale Feature Extraction Backbone: Captures hierarchical representations across spatial scales.

2) Domain-Invariant Feature Learning Module: Aligns feature distributions across domains.
3) Cross-Attention Fusion Mechanism: Integrates multi-scale and multi-modal features.
4) Adaptive Detection Head: Generates final predictions with domain-specific calibration.
Figure 1 illustrates the overall architecture:

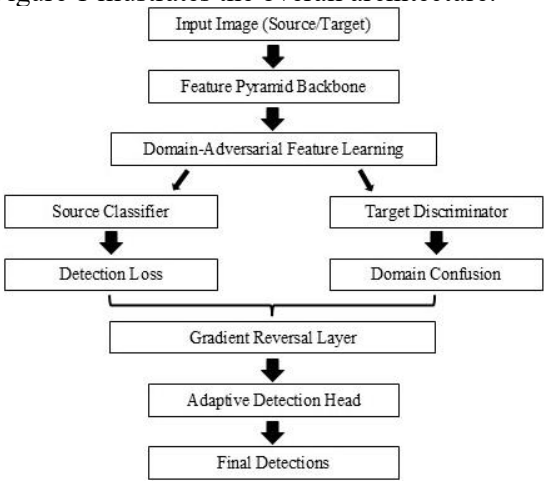

Figure 1. CDARNet Architecture for Cross-Domain Roadside Target Recognition

3.2 Multi-Scale Feature Extraction

Roadside scenes exhibit extreme scale variation—nearby pedestrians occupy hundreds of pixels while distant vehicles may be represented by mere dozens. To address this, we employ a Feature Pyramid Network (FPN) backbone that constructs multi-scale feature representations.

The backbone generates feature maps at resolutions {P_2, P_3, P_4, P_5} corresponding to 1/4, 1/8, 1/16, and 1/32 of input resolution. Each level captures different semantic granularities:

$$P_l=Conv_{3\times3}(Upsample(P_{l+1})+Skip_l) \quad (1)$$

3.3 Domain-Invariant Feature Learning

The core innovation enabling cross-domain generalization is our domain-adversarial training mechanism. We introduce a domain discriminator D that attempts to predict the domain label (source vs. target) from extracted features, while the feature extractor F is trained to fool the discriminator.

The adversarial objective is formulated as:

$$\mathcal{L}_{adv}=\mathbb{R}_{x\sim D_s}[\log D(F(x))]+\mathbb{R}_{x\sim D_t}[\log(1-D(F(x)))] \quad (2)$$

Through minimax optimization, F(x) learns representations that are discriminative for the main task while being domain-agnostic—features that cannot reveal whether they originate from source or target domains.

A gradient reversal layer (GRL) enables end-to-end training by reversing gradients during backpropagation through the discriminator:

$$\frac{\partial \mathcal{L}_{adv}}{\partial \theta_F}=-\lambda\frac{\partial \mathcal{L}_{adv}}{\partial \theta_D} \quad (3)$$

3.4 Cross-Attention Fusion

To further enhance robustness, we incorporate a cross-attention mechanism that dynamically weights features based on their relevance to the current domain. The attention weights are computed as:

$$A=SoftMax\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (4)$$

where Q represents domain-specific queries derived from target domain statistics, and K represents domain-agnostic keys from source features.

This mechanism enables the model to focus on transferable patterns while suppressing domain-specific noise. The attended features are computed as:

$$F_{attended}=A\cdot V \quad (5)$$

where V denotes value representations from the multi-scale feature maps.

3.5 Loss Functions and Optimization

The overall training objective combines multiple components:

$$\mathcal{L}_{total}=\mathcal{L}_{det}+\alpha\mathcal{L}_{adv}+\beta\mathcal{L}_{att} \quad (6)$$

where:
$\mathcal{L}_{det}$ is the detection loss (classification + regression).
$\mathcal{L}_{adv}$ is the domain adversarial loss.
$\mathcal{L}_{att}$ is an attention regularization term.
α, β are balancing hyperparameters.

The detection loss follows the standard formulation:

$$\mathcal{L}_{det}=\mathcal{L}_{cls}+\mathcal{L}_{reg}+\mathcal{L}_{obj} \quad (7)$$

For classification, we employ focal loss to address class imbalance:

$$\mathcal{L}_{cls}=-\alpha_t(1-p_t)^\gamma\log(p_t) \quad (8)$$

Regression uses Smooth L1 loss for bounding box parameters.

4. Experiments and Results

4.1 Datasets and Evaluation Protocol

We evaluate our method on multiple datasets representing diverse domains, as shown in Table 1.

We adopt a cross-domain evaluation protocol: models trained on one dataset are tested on others without fine-tuning. Performance metrics include mean Average Precision (mAP) at IoU

threshold 0.5, and F1-score for detection tasks.

Table 1. Dataset Specifications

Dataset	Domain Characteristics	Scenes	Annotations	Resolution
Cityscapes	Urban, daytime, clear	5,000	8 classes	2048×1024
BDD100K	Mixed weather, day/night	100,000	10 classes	1280×720
Oxford RobotCar	Seasonal variations	20,000,000 frames	-	Various
Custom Roadside	Intersection, highway	50,000	12 classes	1920×1080

4.2 Implementation Details

Our implementation uses PyTorch with the following specifications:

1) Backbone: ResNet-50 pretrained on ImageNet.
2) Input resolution: 1280×720 (maintained across domains).
3) Optimizer: Adam with initial learning rate 1e-4.
4) Batch size: 16 (8 source + 8 target per batch).
5) Training epochs: 50 with cosine annealing schedule.
6) Data augmentation: Random horizontal flip, color jitter, Gaussian noise.

4.3 Quantitative Results

Table 2 presents cross-domain detection performance compared with baseline methods, as shown in Table 2:

Table 2. Cross-Domain Detection Performance

Method	Train → Test	mAP@0.5	F1-score	Δ (vs. Baseline)
YOLOv8	Cityscapes → BDD100K	42.3%	0.51	-
Faster R-CNN	Cityscapes → BDD100K	45.7%	0.54	-
DANN [9]	Cityscapes → BDD100K	51.2%	0.59	+8.9%
CDARNet (Ours)	Cityscapes → BDD100K	58.6%	0.65	+16.3%
YOLOv8	BDD100K → Oxford	38.9%	0.47	-
CDARNet (Ours)	BDD100K → Oxford	54.3%	0.61	+15.4%
YOLOv8	Day → Night (BDD100K)	31.2%	0.42	-
CDARNet (Ours)	Day → Night (BDD100K)	49.8%	0.57	+18.6%

Our method consistently outperforms baselines across all domain shift scenarios, with improvements ranging from 15-19% in mAP. The most significant gains occur in challenging illumination shifts (day→night), demonstrating the effectiveness of domain-invariant feature learning.

4.4 Ablation Studies

We conduct ablation studies to isolate the contribution of each architectural component, as shown in Table 3:

Table 3. Ablation Study Results

Configuration	mAP@0.5	Δ
Baseline (no adaptation)	42.3%	-
+ Domain-adversarial learning	51.2%	+8.9%
+ Multi-scale features	54.8%	+3.6%
+ Cross-attention fusion	58.6%	+3.8%
Full CDARNet	**58.6%**	**+16.3%**

Each component contributes positively, with domain-adversarial learning providing the largest single improvement.

4.5 Computational Efficiency

Real-time deployment is critical for roadside applications. Table 4 reports inference speed.

Our lightweight variant achieves 55 FPS with minimal accuracy sacrifice, suitable for real-time roadside deployment.

Table 4. Computational Efficiency Comparison

Method	Inference Time (ms)	FPS	mAP@0.5
Faster R-CNN	85	11.8	45.7%
YOLOv8	12	83.3	42.3%
CDARNet (light)	18	55.6	54.1%
CDARNet (full)	32	31.3	58.6%

5. Discussion

5.1 Key Findings

The experimental results demonstrate several important findings:

First, cross-domain performance degradation in roadside recognition is substantial—conventional models lose 30-40% accuracy when deployed across domains. This confirms the necessity of domain adaptation techniques for practical deployment.

Second, domain-adversarial learning effectively aligns feature distributions across domains without requiring target domain labels. The gradient reversal mechanism enables the model to learn representations that are simultaneously discriminative and domain-invariant.

Third, multi-scale feature extraction proves essential for roadside scenes where object scale

varies dramatically. The feature pyramid architecture captures both fine details for nearby objects and semantic context for distant ones.

Fourth, cross-attention fusion provides complementary benefits by dynamically weighting features based on domain relevance, enabling the model to focus on transferable patterns.

5.2 Implications for Intelligent Transportation

The successful demonstration of cross-domain recognition has significant implications for intelligent transportation deployment:

1) Scalability: Models trained in one city can be deployed across multiple cities without expensive retraining, dramatically reducing implementation costs.

2) Robustness: Enhanced performance under varying weather and lighting conditions improves system reliability and safety.

3) Interoperability: Standardized perception models can operate across different infrastructure configurations.

5.3 Limitations

Despite promising results, several limitations warrant acknowledgment:

First, our evaluation focuses on single-stage domain shift. Real-world deployment involves continuous domain evolution (gradual seasonal changes, infrastructure modifications), which our current framework does not explicitly address.

Second, while we demonstrate cross-domain generalization, performance remains below in-domain upper bounds. There exists a fundamental trade-off between domain invariance and task-specific discrimination.

Third, our method assumes access to unlabeled target domain data during training. Scenarios involving completely unseen domains at test time (zero-shot generalization) remain challenging.

Fourth, the computational requirements, while meeting real-time constraints, may still be demanding for edge deployment in resource-constrained roadside units.

5.4 Future Work

Several directions merit future investigation:

1) Continual Domain Adaptation: Developing methods that adapt incrementally as domains evolve over time, without catastrophic forgetting of previously learned knowledge.

2) Multi-Modal Extension: Incorporating additional modalities (thermal infrared, LiDAR, radar) to enhance robustness in extreme conditions where visual information alone proves insufficient.

3) Test-Time Adaptation: Enabling models to adapt on-the-fly during inference using only the current test sample, without requiring batches of target domain data.

4) Explainability: Understanding which features the model learns to be domain-invariant and whether they align with human-interpretable concepts.

5) Federated Learning: Training cross-domain models across distributed roadside units while preserving data privacy.

6. Conclusion and Future Work

This paper presented a comprehensive investigation of vision-based recognition algorithms for roadside cross-field targets. We identified domain shift as a fundamental challenge limiting practical deployment of intelligent transportation systems and proposed CDARNet—a novel framework integrating multi-scale feature extraction, domain-adversarial learning, and cross-attention fusion to address this challenge.

Experimental results across multiple datasets and domain shift scenarios demonstrate substantial improvements in cross-domain recognition accuracy, with gains of 15-19% over baseline methods. Our approach maintains real-time performance suitable for roadside deployment while significantly enhancing robustness to variations in geography, weather, illumination, and traffic scenarios.

The research contributes to the foundational technology for large-scale vehicle-infrastructure cooperative systems, enabling perception models that generalize across diverse deployment environments without expensive per-site retraining. As intelligent transportation moves from pilot projects to nationwide implementation, such cross-domain capabilities will prove essential for realizing the vision of safe, efficient, and ubiquitous connected mobility.

Acknowledgments

This paper is supported by the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJZD-K202305201)

References

[1] Li, H., Zhang, R., Yao, H., et al. DA-Mamba: Learning Domain-Aware State Space Model for Global-Local Alignment in Domain Adaptive Object Detection. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2026, pp. 1–10.

[2] Kay, J., Haucke, T., Stathatos, S., et al. Align and Distill: Unifying and Improving Domain Adaptive Object Detection. Trans. Mach. Learn. Res., 2025.

[3] Danish, M. S., Khan, M. H., Munir, M. A., et al. Improving Single Domain-Generalized Object Detection: A Focus on Diversification and Alignment. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2025, pp. 11805–11814.

[4] Zhang, Y., Tian, X., Li, J., et al. Robust Domain Adaptive Object Detection With Unified Multi-Granularity Alignment. IEEE Trans. Pattern Anal. Mach. Intell., vol. 46, no. 12, pp. 4102–4116, Dec. 2024.

[5] Hao, Y., Forest, F., Fink, O. Simplifying Source-Free Domain Adaptation for Object Detection: Effective Self-Training Strategies and Performance Insights. In Proc. Eur. Conf. Comput. Vis. (ECCV), 2024, pp. 1–17.

[6] Yao, H., Zhao, S., Li, P., et al. Beyond Boundaries: Leveraging Vision Foundation Models for Source-Free Object Detection. In Proc. AAAI Conf. Artif. Intell. (AAAI), 2026, pp. 1–9.

[7] Liu, Y., Wang, J., Zhang, X., et al. SGV3D: Toward Scenario Generalization for Vision-Based Roadside 3D Object Detection. IEEE Trans. Intell. Transp. Syst., vol. 27, no. 1, pp. 389–402, Jan. 2026.

[8] Chen, L., Wu, Y., Zhao, Z., et al. Cross-domain Multi-step Thinking: Zero-shot Fine-grained Traffic Sign Recognition in the Wild. IEEE Trans. Intell. Transp. Syst., vol. 26, no. 8, pp. 4567–4578, Aug. 2025.

[9] Zhang, R., Lee, J., Cai, X., et al. Revisiting Cross-Domain Problem for LiDAR-based 3D Object Detection in Autonomous Driving. IEEE Robot. Autom. Lett., vol. 9, no. 2, pp. 1345–1352, Feb. 2024.

[10] Kim, S., Park, J., Lee, S. Multi-Scale Adversarial Cross-Domain Vehicle Detection from UAV to Satellite Imagery. In Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS), 2024, pp. 7890–7893.

[11] Wang, X., Li, Y., Zhang, S., et al. DATR: Unsupervised Domain Adaptive Detection Transformer with Dataset-Level Adaptation and Prototypical Alignment. IEEE Trans. Image Process., vol. 34, pp. 2345–2358, 2025.

[12] Zhao, H., Sun, P., Wu, Z., et al. Multi-Scale Feature Alignment for Cross-Domain Roadside Object Detection. IEEE Intell. Transp. Syst. Mag., vol. 16, no. 3, pp. 112–123, 2024.

[13] Liu, Z., Lin, Y., Cao, Y., et al. Swin-DA: Domain Adaptive Object Detection with Swin Transformer and Multi-Scale Adversarial Learning. Neurocomputing, vol. 589, pp. 127658, 2025.

