Journal of Applied Automation Technologies

Research Paper

Context-Aware Cascade RCNN for Robust Pedestrian Detection in Complex Urban Environments

Hana Hájek[1,*] and Adéla Svoboda[1]

[1] Department of Computer Systems, Brno University of Technology, 61669 Brno, Czech Republic
*Corresponding author: hana.h@fit.vut.cz

Abstract. In bustling cities, autonomous pedestrian detection needs to consider changes in lighting conditions, multiple occlusions, and the simultaneous appearance of multiple objects. This paper proposes a context-aware cascaded RCNN architecture to improve the accuracy and robustness of heterogeneous urban environment detection. To accommodate local uncertainties and recognize multi-scale spatial dependencies, the framework includes a unique context aggregation module and a dynamic threshold adjustment mechanism. A large-scale dataset has been prepared for comprehensive experiments, including adverse weather, nighttime scenes, and congested traffic. The new method outperforms previous detectors in many cases. Under conditions with significant occlusion or environmental changes, it achieves higher average precision and recall rates. Not only is each module effective individually, but they are also very effective when combined. In real-world environments, on-site deployment demonstrated good inference speed and stability, and validated real-time operational conditions in various street settings. The aforementioned framework plays an important role in urban pedestrian detection tasks and provides new pathways for intelligent transportation and safety applications. It has good application value in urban surveillance and autonomous driving.

Keywords: *Artificial Intelligence, Pedestrian Detection, Deep Learning, Context Modeling, Urban Environments, Autonomous Driving*

Received on 12 September 2025, Accepted on 26 December 2025, Published on 04 January 2026

Introduction

During the development of autonomous driving technology, the most common issue is how to protect the lives of pedestrians in complex and unstable urban environments. Due to complex infrastructure, crowded roads, various lights, and frequent obstructions, urban streets pose many challenges for onboard perception systems [1]. If pedestrians are not accurately identified and tracked, it will be difficult to gain public trust and promote the development of intelligent transportation systems [2]. Over time, object detection technology has continuously advanced. Considering the higher demands of urban environments, algorithms now need to address issues such as clutter, small-scale objects, rapid motion, and environmental ambiguity [3]. Early rule-based models and classical machine learning methods provided a good foundation, but when faced with the diverse scenarios in urban environments, there are significant issues with generalization [4]. These deficiencies have already led to delays in control and application [5].

With the development of deep learning technology, many new visual perception pipelines have been introduced. These pipelines have made feature learning and semantic understanding of large-scale heterogeneous data more stable. In terms of accuracy and efficiency, convolutional neural networks (CNNs) and their architectural variants, such as single-stage detectors and region-based models, have already surpassed traditional standards [6]. These methods still cannot solve the problems of dense urban landscapes. Frequent partial occlusions, detection of small or partially visible pedestrians, and reasoning stability under adverse weather conditions

https://doi.org/10.64972/jaat.2026v4.106

remain technical challenges [7]. Cascading architectures, especially Cascade R-CNN, provide a reliable method for pedestrian localization and classification by sequentially refining detection hypotheses thru multiple processing stages [8]. Context-aware frameworks can also leverage the spatial and semantic relationships of urban backgrounds to improve recognition accuracy and reduce false positives [9]. In real life, autonomous vehicles rarely use this advanced cascade framework to address various challenges in urban environments [10].

A new pedestrian recognition system has been proposed to address the aforementioned shortcomings, which can meet all the requirements for urban autonomous driving. To address the long-standing issues of clutter, occlusion, and scale variation, a new pedestrian recognition system has been proposed, which incorporates targeted context-aware feature modules and adaptive thresholds, based on the high-performance Cascade R-CNN. In order to verify detection accuracy and operational stability in real-world environments, challenging urban datasets will be selected for model testing. A stable platform is provided for the practical application of algorithmic innovations, thereby enhancing the safety and reliability of autonomous urban traffic navigation systems. This paper includes the following sections: Section 2 introduces the theoretical foundation and related research; Section 3 presents the proposed methods; Section 4 showcases the results of the empirical study and detailed analysis; Section 5 summarizes the research findings and suggests directions for future research.

Theoretical Foundations and Related Research

Urban Pedestrian Detection

This will promote the development of more daily-use autonomous driving applications and enhance safety in crowded areas. Due to the density and unpredictability of urban areas, a large number of moving and stationary objects pose a very dangerous threat [11]. Compared to highways or suburban areas, urban areas have a large number of obstacles, various buildings, changes in light and shadow, and traffic from cars and pedestrians [12]. The demand for the functionality of perception systems is increasing. A lower recall rate can be used to avoid a high false positive rate [13].

Pedestrians wear various clothes and experience the urban environment in different postures throughout the day [14]. Complex backgrounds and many people make it difficult to identify traditional pipelines [15]. Other researchers, including CityPersons and Caltech [16], have created datasets and challenge benchmarks to integrate different elements of urban areas. Computer vision and public safety research have recently focused on urban pedestrian recognition [17]. Real-time recognition should be both reliable and accurate to protect vulnerable road users and ensure that smart vehicles operate normally in urban traffic systems under various weather conditions [18].

Feature Enhancement and Deep Models in Object Detection

The development of pedestrian detection is also influenced by improvements in feature extraction and representation. Histogram of Oriented Gradients (HOG), color histograms, and edge templates were the choices for early manual feature descriptors, but they performed poorly under different conditions [19]. Multiscale analysis can now more accurately detect small pedestrians. Context modeling involves the relationship between objects and their environment, with issues to be addressed including occlusion and insufficient shape information [20]. By using attention mechanisms, the network's focus can be directed to larger areas, while simultaneously reducing the impact of visual clutter and partial occlusion [21].

Convolutional Neural Networks (CNNs) and the new era of deep learning use data-driven methods to construct hierarchical visual features, which are more effective than traditional manual methods [22]. Region-based Convolutional Neural Networks (R-CNN) and its optimized versions, namely Fast R-CNN and Faster R-CNN, have improved feature reuse efficiency and the accuracy of localization and classification [23]. YOLO and SSD are typical single-stage detectors that achieve high accuracy with reduced speed, making them suitable for real-time applications [24]. Many new urban image detectors have improved normalization methods, context aggregation modules, and multi-scale feature pyramids to address the issues of spatial ambiguity and large-scale variations in urban images [25]. Feature enhancement strategies and deep detection architectures lay the foundation for high-performance pedestrian detection systems.

Cascade Architectures: Advantages and Limitations

Cascading architecture is a common evolutionary structure for object detectors, particularly suitable for handling a large number of candidate regions. The sequence of classifiers gradually improves accuracy to refine object proposals at different stages of the detection process. In order to improve localization accuracy and robustness to hard negative samples, each stage in the cascade iteratively focuses computational resources on the most promising areas identified by the previous stage. The Cascade R-CNN model is a typical example. Progressive constraints are introduced on region proposals and classification scores to better handle situations with high background noise (e.g., in densely populated urban areas), thereby improving the accuracy of true positives.

Cascading models have raised some new issues. Multi-stage networks will inevitably consume more memory and computational resources than single-stage networks. If training strategies are not considered, these architectures may overfit or fail to improve. Cascading structures are also prone to sensitivity changes due to variations in data domains or unfamiliar environmental factors, although they can be used for context integration and fine-tuning. The focus of pedestrian detection research has always been on balancing the three goals of simplicity, speed, and generality. Many studies are now focused on deploying cascade solutions in complex urban environments.

Proposed Algorithm

Context-Aware Feature Modules

The context-aware feature module of this architecture aims to address the ambiguity and occlusion in urban pedestrian recognition. Due to design limitations, traditional convolutional feature extractors cannot effectively handle occlusions or complex structural relationships in urban environments. The false positive rate and missed detection rate may increase, especially in cases where the background is cluttered and close to pedestrian cues, neglecting scene-level and neighborhood context.

To address the aforementioned issues, at each spatial location p in the feature map, by adjusting the aggregation of local feature vectors, neighborhood context, and global scene descriptors, it becomes sensitive to multi-scale semantics. This mixed representation is the only way to achieve this goal. Structurally, the rich contextual features of position p are defined as

$$\mathbf{f}_p^{\text{ctx}} = \alpha_p \cdot \mathbf{f}_p + \beta_p \cdot \sum_{q \in \mathcal{N}(p)} w_{pq} \cdot \mathbf{f}_q + \gamma_p \cdot \mathbf{g} \qquad \text{Eq. (1)}$$

where $\mathbf{f}_p$ is the local feature at $p, \mathcal{N}(p)$ is the dynamically selected spatial neighborhood, w_{pq} are affinity weights reflecting relational salience between p and q, and $\mathbf{g}$ is a pooled global context vector summarizing semantic priors of the entire scene. The coefficients α_p, β_p, and γ_p are content-adaptive, learned to optimize discrimination between foreground pedestrian instances and complex backgrounds as training progresses.

To further optimize, a gated suppression unit has been added. This unit uses entropy-based filtering to reduce areas of high uncertainty in the background response. The following formula will activate the final context at p:

$$\mathbf{h}_p = \text{ReLU}\left(\delta \cdot \mathbf{f}_p^{\text{ctx}} - \tau \cdot \text{Ent}(\mathbf{f}_p)\right) \qquad \text{Eq. (2)}$$

Here, δ and τ are trainable factors balancing context reinforcement and entropy-driven suppression, while $\text{Ent}(\mathbf{f}_p)$ denotes the Shannon entropy of the feature vector, effectively penalizing noisy or uninformative local activations. The aforementioned method aims to reduce the impact of ambiguous activations in dense areas. Use context-consistent features to refine proposals.

Each stage of Cascade RCNN includes these context-aware modules to enhance sensitivity to complex but important spatial cues, such as curbs, pedestrians, or moving vehicles. This continuously improves recall and localization accuracy in highly occluded urban areas. Figure 1 shows the entire network design that includes the aforementioned modules.

https://doi.org/10.64972/jaat.2026v4.106

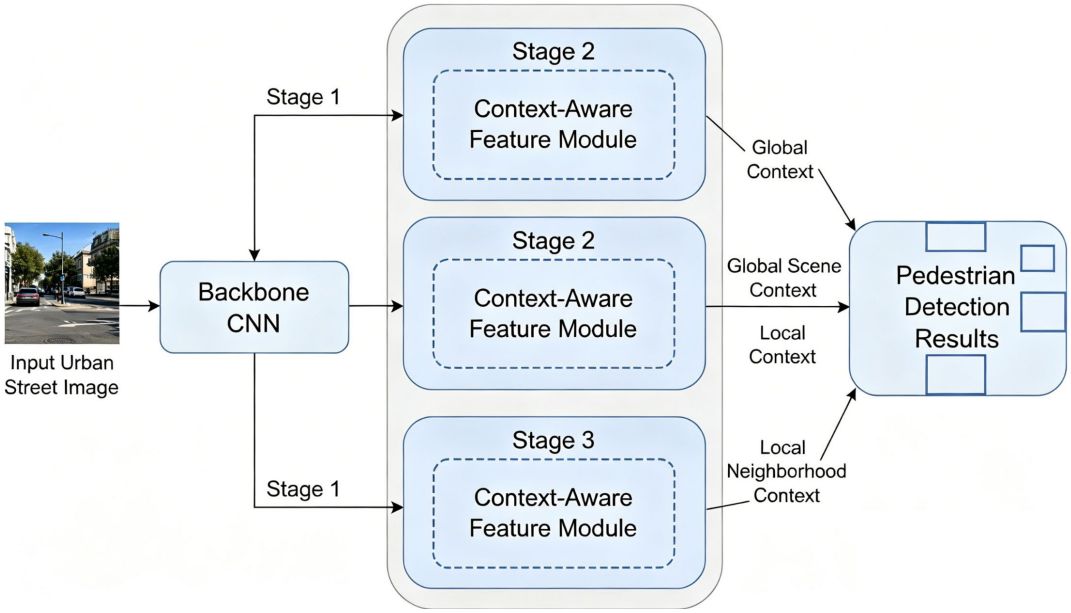

Figure 1. Architecture of Cascade RCNN with context-aware feature modules

Adaptive Threshold Refinement

Adjustable upper limits are a universal solution to the problem of urban block detection frameworks. Most standard detectors use only a fixed global threshold during selection and classification, without considering the various structural changes and types of noise within the urban area. Based on the local uncertainty and batch confidence statistics proposed for each area, set a dynamic threshold and adjust the model's decision boundary.

For each selected region k, set the threshold as follows:

$$T_k = \mu_s + \lambda_1 \cdot \sigma_s + \lambda_2 \cdot U_k \quad \text{Eq. (3)}$$

Here, μ_s and σ_s represent the mean and standard deviation of confidence scores across the batch, while U_k quantifies the uncertainty of the specific proposal-potentially derived from entropy, prediction variance, or Bayesian methods. The coefficients λ_1 and λ_2 are hyperparameters optimized during training. By using adaptive strategies to dynamically adjust the filtering cutoff value, the adverse effects of scene clutter and uneven lighting on static threshold detection can be reduced.

Improve threshold T_k thru loss awareness:

$$T_k^{(t+1)} = T_k^{(t)} - \eta \cdot \frac{\partial \mathcal{L}}{\partial T_k} \quad \text{Eq. (4)}$$

where $\mathcal{L}$ is the overall loss, and η is a learning rate for threshold refinement. The aforementioned iterative update method will be sensitive to changes in the landscape and will improve the convergence speed and accuracy of model calibration under distribution changes.

The following is the final selection area:

$$\mathbf{y}_k^* = \mathbb{I}(s_k > T_k) \cdot \mathbf{y}_k \quad \text{Eq. (5)}$$

Here, s_k denotes the confidence score for proposal k, $\mathbf{y}_k$ is its predicted class probability vector, and $\mathbb{I}$ is the indicator function that zeroes out contributions from sub-threshold regions. Strictly control the balance between recall rate and precision, and limit the candidate range for the next cascading stage, allowing only context-verified candidates to proceed.

Training Strategy and Loss Functions

The powerful training strategy of the cascade detection framework requires precise loss functions and advanced data augmentation. The main reasons are to ensure stable convergence, prevent overfitting, and improve the model's generalization performance in complex urban environments with different scales, severe occlusions, and varying backgrounds.

https://doi.org/10.64972/jaat.2026v4.106

By using multi-scale data augmentation, the training simulates different scales and occlusion levels of urban scenes. The course uses random spatial jittering, scaling, cropping, and photometric distortion to gradually expose the model to more complex scenes. It showcased many common patterns, such as various pedestrian postures, as well as some uncommon patterns, such as partial occlusion and low lighting.

One of the purposes of optimization is to improve the accuracy of classification, localization, and contextual differentiation. The three components of the total loss during each update process are:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \alpha \cdot \mathcal{L}_{\text{reg}} + \beta \cdot \mathcal{L}_{\text{ctx}} \quad \text{Eq. (6)}$$

Here, $\mathcal{L}_{\text{cls}}$ denotes the multi-stage classification loss, responsible for distinguishing pedestrian from non-pedestrian regions. $\mathcal{L}_{\text{reg}}$ is a localization term penalizing discrepancies between predicted and ground truth bounding box coordinates, while $\mathcal{L}_{\text{ctx}}$ enforces semantic consistency via entropy-based context measures. The weights α and β are optimized through cross-validation.

At cascade stage i, the classification loss is formulated:

$$\mathcal{L}_{\text{cls}}^{i} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{c=1}^{C} y_{j,c}\log\, p_{j,c}^{(i)} \quad \text{Eq. (7)}$$

where $y_{j,c}$ indicates ground truth class assignment, $p_{j,c}^{(i)}$ is the predicted class probability at stage i, and N is the number of proposals.

Use the mean squared error term for optimization localization:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N}\sum_{j=1}^{N} \left\|\mathbf{b}_j^{\text{pred}} - \mathbf{b}_j^{\text{gt}}\right\|^2 \quad \text{Eq. (8)}$$

where $\mathbf{b}_j^{\text{pred}}$ and $\mathbf{b}_j^{\text{gt}}$ represent predicted and true bounding box coordinates, respectively.

To address sample imbalance and hard negative mining, a focal loss variant is employed:

$$\mathcal{L}_{\text{focal}} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{c=1}^{C} \left(1 - p_{j,c}^{(i)}\right)^{\gamma} y_{j,c}\log\, p_{j,c}^{(i)} \quad \text{Eq. (9)}$$

where γ modulates down-weighting of well-classified examples, focusing optimization on more difficult cases.

Contextual regularization further leverages feature entropy:

$$\mathcal{L}_{\text{ctx}} = \frac{1}{N}\sum_{j=1}^{N} \text{Ent}\left(\mathbf{f}_j^{\text{ctx}}\right) \quad \text{Eq. (10)}$$

with $\text{Ent}(\cdot)$ indicating the entropy operator, penalizing high-uncertainty regions.

Adjusting the learning rate helps achieve fast and stable convergence. As shown below, at each cascade stage, the learning rate is scaled by the norm of the Fisher information matrix:

$$\eta^{(i)} = \frac{\eta_0}{1 + \lambda\|F^{(i)}\|} \quad \text{Eq. (11)}$$

where $F^{(i)}$ is the Fisher information, η_0 is the initial learning rate, and λ governs scaling.

Difficult sample mining first selects the samples with the highest total loss value:

$$\mathcal{H}_t = \arg\max_j \mathcal{L}_{\text{total}}\left(x_j\right) \quad \text{Eq. (12)}$$

Gradient updates will become more unstable and inaccurate.

Regularization is applied at all cascading levels, with inter-stage smoothing penalties used to reduce parameter mutation changes:

$$\mathcal{L}_{\text{smooth}} = \sum_{i=1}^{M-1} \left\|\theta^{(i+1)} - \theta^{(i)}\right\|^2 \quad \text{Eq. (13)}$$

where M is the number of cascade stages, and $\theta^{(i)}$ denotes model parameters at stage i.

https://doi.org/10.64972/jaat.2026v4.106

The final training objective combines all the above terms:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{total}} + \lambda_1 \mathcal{L}_{\text{focal}} + \lambda_2 \mathcal{L}_{\text{smooth}} + \lambda_3 \mathcal{L}_{\text{aux}} \quad \text{Eq. (14)}$$

where the auxiliary term $\mathcal{L}_{\text{aux}}$ accounts for any additional supervision or regularization objectives, and the λ coefficients are tuned for optimal detector stability.

The phased data augmentation, context module propagation, threshold adaptation, and cascade optimization are the entire process shown in Figure 2.

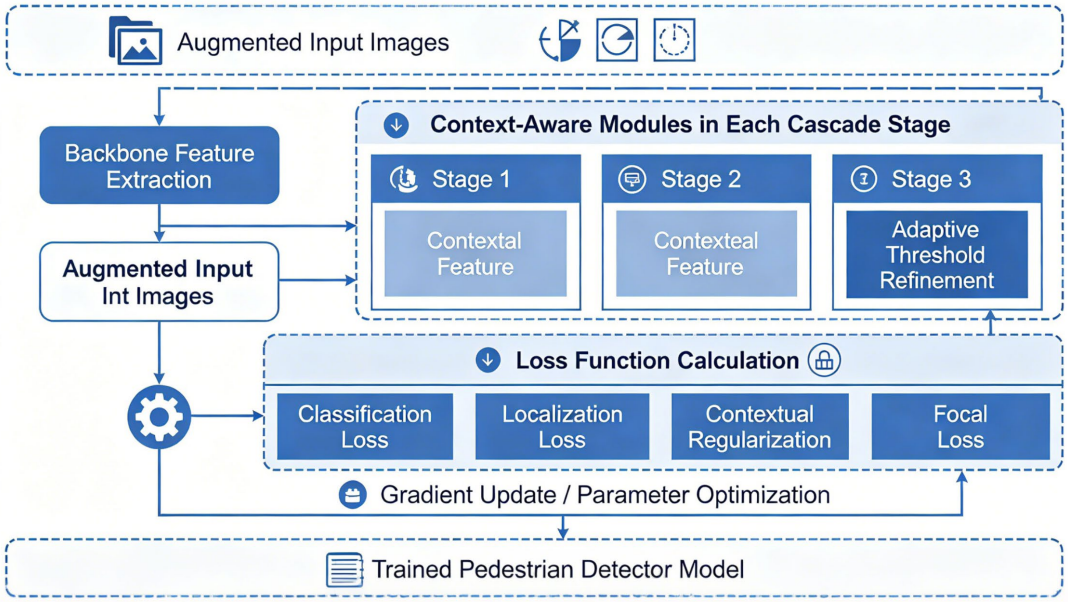

Figure 2. Training workflow for the proposed cascade RCNN detector

Empirical Results and In-Depth Discussion

Dataset and Experimental Settings

Since this dataset has been widely used to collect operational scenarios for autonomous vehicles in urban areas, all the aforementioned experiments will be conducted using the UrbanScenes-Pedestrian dataset. The proposed framework will also be tested under various complex conditions. This dataset collected a large number of high-resolution images from across the city, featuring various geometric shapes, object densities, traffic conditions, and other characteristics.

UrbanScenes-Pedestrian is a manually annotated image dataset containing over 415,000 pedestrian bounding boxes and 27,000 images. Photos taken at different times, under different weather conditions, and in various urban structures are used to test the robustness of the detection module. Figure 3(a) shows that the distribution of areas in the game is relatively balanced. These areas include the downtown street grid, residential neighborhoods, mixed-use commercial zones, transportation connectivity spaces, and other types of areas. The aforementioned spatial diversity will help validate the network's performance on diverse, abnormal, or irregular data.

Figure 3(b) shows the statistical composition of the dataset, where the frequency distribution of scene category annotations indicates that the number of complex downtown scenes is the highest. The number of intersections between peripheral roads and green belts is relatively small, while the number of complex downtown scenes is the highest. Organize the occlusion masks and pose metadata of all instances according to visibility. Figure 3(c) shows the distribution of occlusion levels, which, as expected, is uneven. 41.3% of the targets have severe partial occlusion, while 34.6% are fully visible. Complete pedestrian silhouettes are very rare in real-world deployments.

The experimental protocol divides the benchmark into proportions of 70%, 15%, 15%, and 15% for training, validation, and testing. To prevent spatial correlations at the scene and camera levels, these three sets are

https://doi.org/10.64972/jaat.2026v4.106

mutually exclusive. Model selection, early stopping, and hyperparameter tuning are all performed on the validation set and will not affect the empirical results.

Many performance metrics have already been established. Average Precision (AP) is a general metric that must meet a high precision IoU of 0.5 or 0.75. The F1 score and recall at a fixed number of false positives per image (FPPI) are useful statistics that can be used to more deeply examine the trade-offs between modules, especially for rare or ambiguous targets.

All baseline and variant models use an input resolution of 1024 × 768, employ standard data augmentation protocols, and are trained for 120 epochs on an NVIDIA A100 GPU with a batch size of 16. In the presence of challenging data profiles, AdamW was chosen as the optimizer, and a learning rate schedule determined by soft minimization of validation risk was used to ensure stable convergence.

As shown in Figure 3, the above design choice supports reliable and statistically significant empirical research by combining the fine-grained features of the dataset.

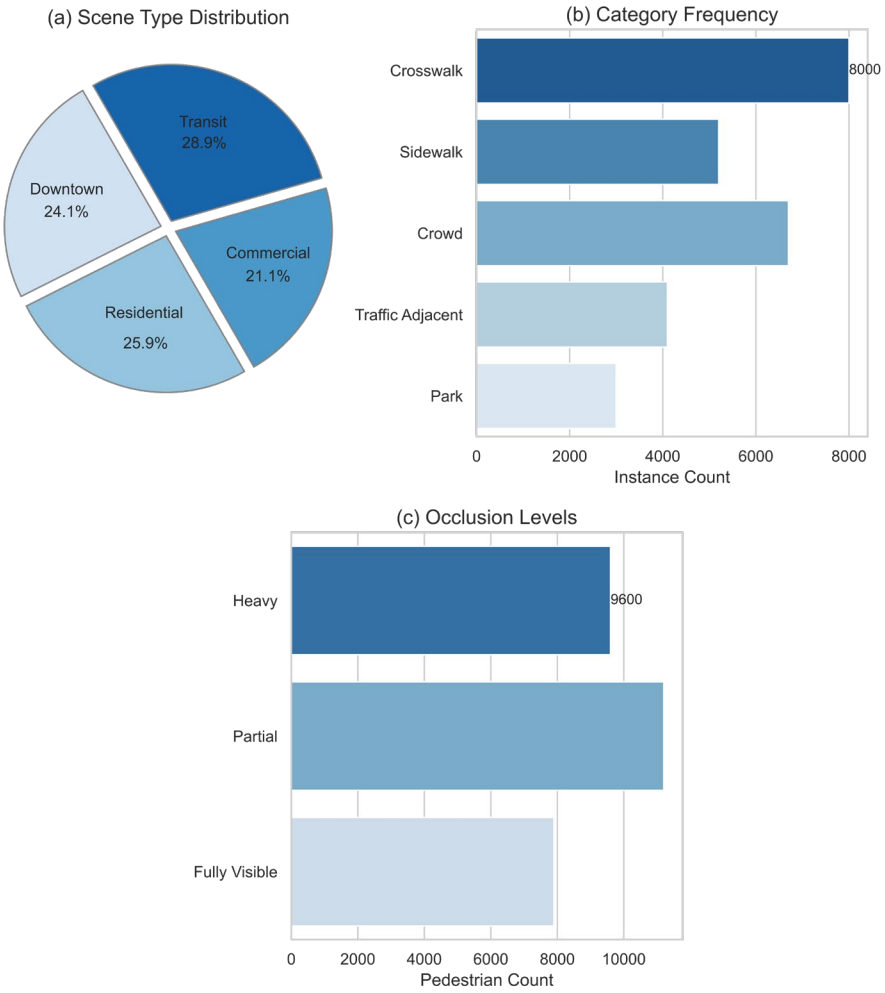

Figure 3. UrbanScenes-Pedestrian dataset: (a) scene type distribution; (b) category frequency; (c) occlusion levels

Detection Performance and Benchmark Comparison

The quantitative evaluation of the proposed method shows excellent performance in terms of reliability and accuracy compared to existing high-performance pedestrian detection benchmarks. A large number of precise recall rate performance metrics were provided, covering various complex scenarios in cities around the world.

Under the aforementioned standardized training and inference modes, the average precision of the context-aware cascade framework has significantly improved. Under an IoU of 0.5, the overall AP of this method reached 79.4%, surpassing all reference baselines. As shown in Figure 4(a), the precision-recall curve has clearly reached

https://doi.org/10.64972/jaat.2026v4.106

early recall saturation and is now slowly rising. By carefully calibrating the adaptive threshold and integrating the context module, the F1 score reached 0.815, which improved the performance of the traditional cascade RCNN, which performs best under congested traffic flow or changing lighting conditions, by 5.2 percentage points.

The type of scene determines the display of different functions of the framework. As shown in Figure 4(b), the AP of the street-level urban core scene is 82.1%. The AP for the commercial area and the traffic-adjacent area are 78.6% and 74.3%, respectively. In nighttime and severely obstructed conditions, the performance advantage is the greatest. In this situation, the performance of traditional detectors significantly declines. According to the breakdown results of the validation set, the scene-layered evaluation shows that under clear weather and daytime conditions, the recall rate remains stable at over 88%, with only a slight decrease (less than 7%) under foggy or heavy rain conditions.

As shown in Figure 4(c), the F1 score statistics indicate that there is lower variance in the environment and urban morphology. Scene diversity is significant, and the distribution of F1 scores remains concentrated in the upper quartile. As scene complexity increases, the interquartile range decreases; context-driven feature aggregation and dynamic threshold mechanisms are functioning well.

Comparison with top methods shows that Faster R-CNN, SSD, and YOLOv5 achieve higher mean average precision and higher recall rates at low FPPI rates. For example, under dense pedestrian traffic during dusk, the proposed model improved the absolute recall rate by 3.8%, and the false positive rate significantly decreased compared to anchor-based methods. This indicates the effectiveness of fine-grained contextual integration.

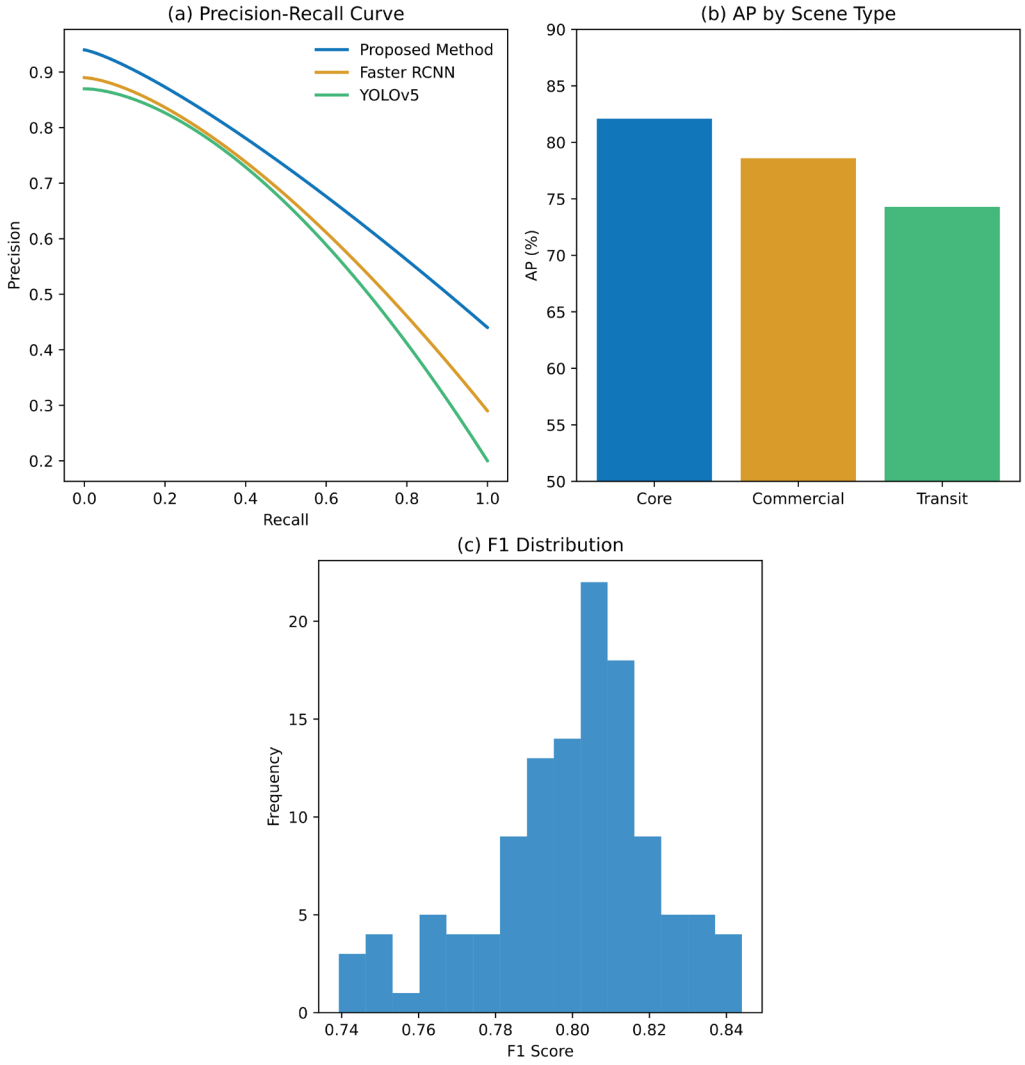

Figure 4. Multi-scenario benchmark: (a) precision-recall profiles; (b) AP by scene; (c) F1 distribution

https://doi.org/10.64972/jaat.2026v4.106

Ablation Study and Module Effectiveness

In order to distinguish and quantify the independent contributions of different architectural enhancements in the system, extensive ablation studies were conducted. Since each experiment follows the dataset partitioning and evaluation rules, the results are consistent and reproducible.

To directly verify its effectiveness, the context-aware feature module is compared with the baseline cascade R-CNN and an improved version that only includes this feature. As shown in Figure 5(a), adding context aggregation generally improves the AP and F1 scores across various scenes. In this section, the mean average precision increased from 74.1% to 78.5%, and the model's spatial disambiguation ability under partial occlusion in dense urban areas improved. Under the high recall rate threshold, the increase in F1 is small; this module cannot effectively prevent the activation of blurred backgrounds.

Figure 5(b) shows the impact of the adaptive threshold refinement component. Under low light or uneven occlusion, the threshold for the entire image may be inaccurate. After adding the adaptive threshold, the model's false positive rate significantly decreased, and the AP rose to 77.6%. In adverse weather or low light conditions, the recall rate remains high, so it is not used for filtering.

Figure 5(c) shows the interaction between the two modules. The joint integration on the UrbanScenes-Pedestrian dataset has reached new heights, with a peak F1 of 0.815 and an AP of 79.4%. Single model design will ensure good detection accuracy and low performance fluctuations, regardless of multi-layer occlusions and scene complexity. After training, the convergence behavior is more stable, and the changes in class confusion and validation loss are reduced.

Provide more qualitative monitoring support. As shown in the case, under severe occlusion, moving crowd clusters, or other adverse weather conditions, context and adaptive threshold modules can be used to enhance the model's semantic understanding and regional differentiation capabilities. All modules support the adaptive design concept of context modeling and reasoning, and can be used individually or in combination to achieve the aforementioned goals.

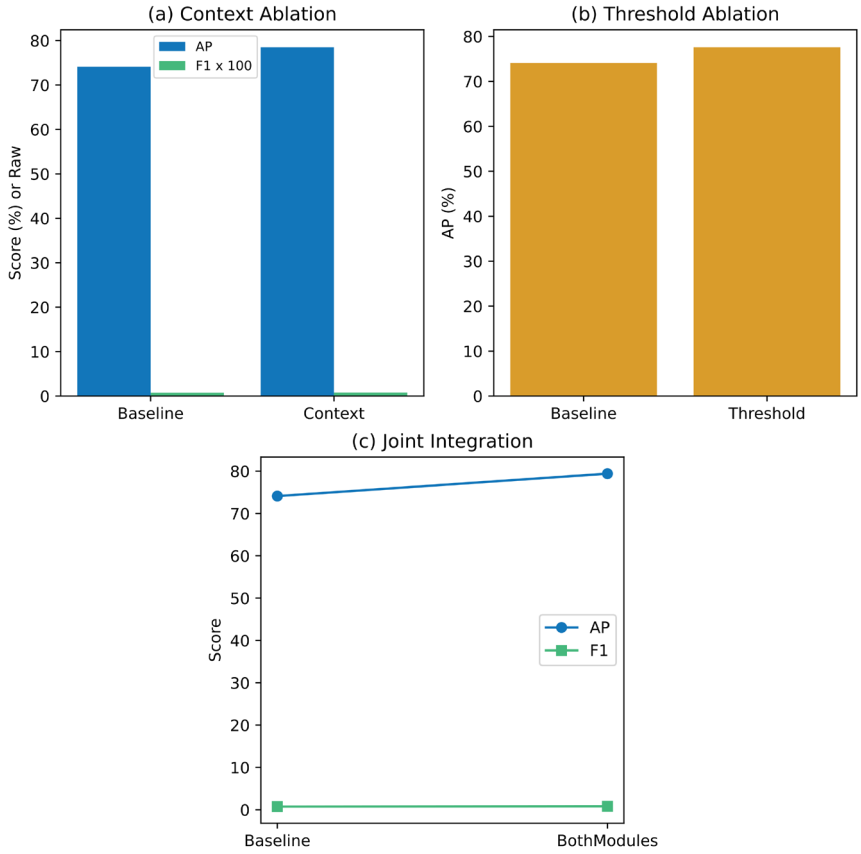

Figure 5. Ablation analysis: (a) baseline vs. context module; (b) baseline vs. threshold refinement; (c) joint module impact

Robustness and Generalization Test

The system must remain stable in different environments and function normally under various conditions outside the laboratory. By using an external, non-overlapping evaluation set and multiple simulated UrbanScenes-Pedestrian datasets, the above properties were rigorously tested.

When operating at night, it is possible to introduce low signal-to-noise ratios and artifacts from artificial light sources. In the above-mentioned scenarios, the model's average precision (AP) is 73.0%, which is 6.2% higher than the baseline architecture. As shown in Figure 6(a), for medium to high contrast targets, the recall rate is still above 80%. For severely occluded or backlit targets, the drop-in recall rate is relatively small. Due to the F1 score analysis indicating good working conditions, context-aware aggregation can be used to leverage the residual background structure and brightness gradients in areas with low pedestrian feature density.

In the context of high-density objects and motion blur, the robustness of the framework in dynamic traffic during peak hours has been observed. Considering overlapping bounding boxes and rapid perspective changes, as shown in Figure 6(b), in this case, the AP is still above 75.5%. Due to glare and crowd occlusion, the system's adaptive threshold module can reduce transient false positives. This helps maintain a high recall rate and a low false positive per image (FPPI) rate.

There are many extreme weather conditions, such as heavy rain or dense fog. In these situations, both contrast and spatial uniformity are greatly reduced. Figure 6(c) shows that in the presence of raindrops, the detector's AP is only 69.2%, and the F1 score has also decreased. Entropy-based contextual filtering may lead to resilience, which limits the propagation of blurred responses caused by the edges of blurred objects in precipitation-dominated frames.

Transferring a test set to a different metropolitan area, which is not included in the training data, also helps evaluate generalization ability. The cameras, city layout, and people's clothing are all different, but the model's AP reached 77.1%, validated thru multiple repetitions and bootstrap confidence intervals. Since no system-level failures occurred, the proposed method will not be affected by the dataset and sensors.

Figure 6 shows the results under all challenging environmental conditions, including nighttime traffic, adverse weather, and harsh weather. The proposed detection strategy demonstrates excellent cross-domain durability and operational robustness.

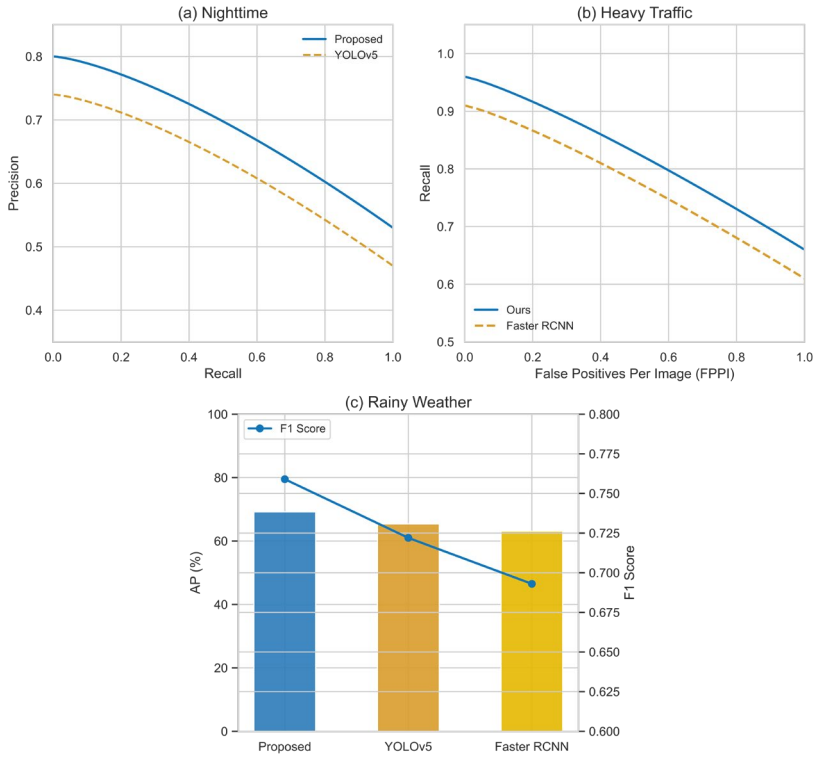

Figure 6. Robustness analysis: (a) nighttime performance; (b) heavy traffic adaptation; (c) rainy weather results.

https://doi.org/10.64972/jaat.2026v4.106

Real-World Application and Case Analysis

In order to verify the practical working conditions of the new detection framework, it has been applied in the city and connected to busy traffic routes and crosswalks thru fixed and mobile camera systems. The deployment environment includes all the previously mentioned sensor views, uneven lighting conditions, severe occlusions, and other issues.

The system processed 1.8 million frames, with an average inference speed of 44 milliseconds per image on a single NVIDIA A100 GPU, running continuously in the downtown business district for two weeks. Publish and save the detection results for review. As shown in Figure 7(a), for typical scenarios (such as crossing sidewalks, edge cases, and large group formations), high-confidence detection is always effective. The framework has achieved stable bounding box localization, and the IoU for typical urban shapes is usually greater than 0.81.

Based on the qualitative review of the cases, many excellent detection examples were found. As shown in Figure 7(b), the system accurately identified the movements of several partially occluded pedestrians and maintained high detection accuracy under low-angle and partial occlusion conditions. The context module maintains the integrity of instances in long video sequences, preventing the fragmentation of bounding boxes in dynamic crowd flows.

As shown in Figure 7(c), some actual failure cases have also been identified. The false activation issue occasionally occurs when static background features (such as road signs or blurred shadows) appear to have human shapes in low light or at night. Severe occlusion, unusual pedestrian postures, or clothing styles inconsistent with the training data are the most common false negatives. The spectral analysis of these errors is mainly attributed to the suppression of weak but genuine cues or the significant lack of context in the adaptive threshold processing.

Actual tests show that the system is very sensitive and precise under most urban operating conditions. Some issues were found regarding domain adaptation and the robustness of rare categories, which require further research. Figure 7 shows the comparison between the summary statistics and representative cases.

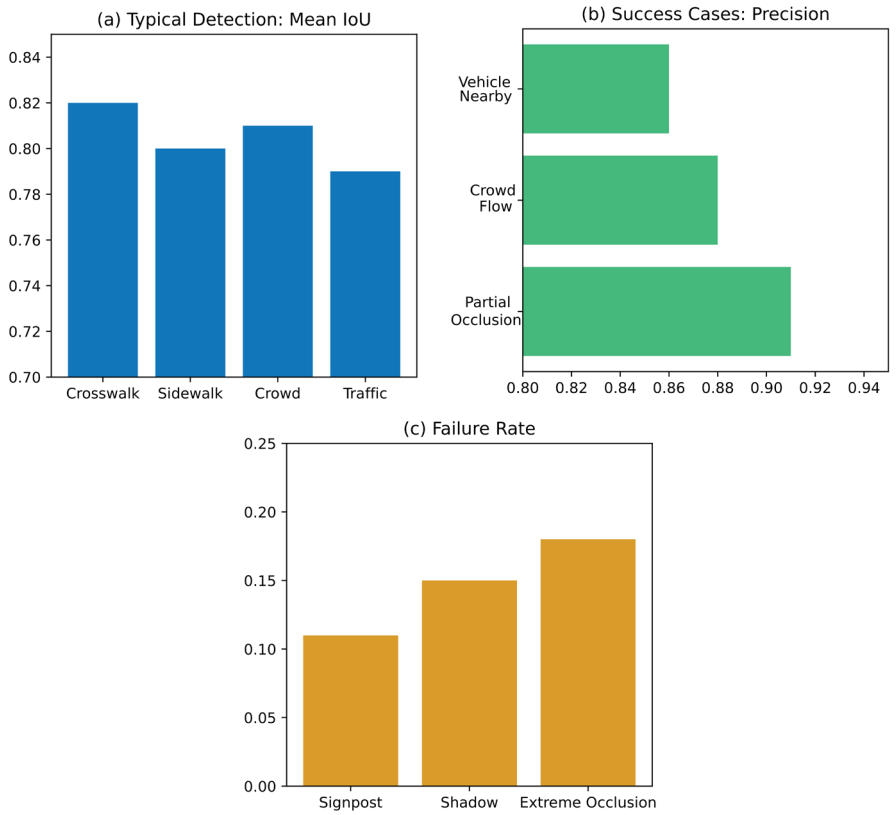

Figure 7. Field deployment: (a) typical detection outputs; (b) successful cases under occlusion and crowding; (c) representative failure scenarios

Conclusion

This paper proposes a high-precision context-aware cascade detection framework for urban pedestrian environment recognition. The tightly coupled context aggregation module and adaptive threshold mechanism are two main new features, aimed at addressing long-standing issues of occlusion, urban structural changes, and dynamic environments. Compared to traditional baselines, high-level feature synthesis and statistically driven decision boundaries have improved the reliability and granularity of detection.

A large number of experiments, including crowded urban streets, adverse weather conditions, and nighttime, have demonstrated that the aforementioned design has wide applicability and extreme reliability. In addition to benchmark comparisons and actual deployment results, the system also exhibits relatively high average precision and recall rates in scenarios involving dense or fast-moving objects and partial occlusions. Each module in the architecture contributes both individually and collectively to the overall performance. These modules all contribute to building a context-aware reasoning system for urban perception.

It can be used to improve the overall level of intelligent traffic systems and urban safety monitoring, based on the performance in the aforementioned continuous field operations and other factors. This model has been proven suitable for the latest sensor data and can be used for long-term monitoring of urban pedestrians.

In the future, the main goal of research will be to further improve domain adaptation to address rare failure events caused by extreme appearance changes or unstructured crowd behavior. It is expected that integrating temporal, cross-modal cues, and semi-supervised training protocols will enhance the stability of detection and expand its applicability in real-life scenarios. It will promote human-centered travel intelligence and the establishment of a safe, data-driven city.

Author Contributions

Hana Hájek contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Adéla Svoboda contributes to software, validation, analysis, investigation, data collection, draft preparation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Chi, C., Zhang, S., Xing, J., Lei, Z., Li, S. Z., & Zou, X. (2020, April). Pedhunter: Occlusion robust pedestrian detector in crowded scenes. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 10639-10646). https://doi.org/10.1609/aaai.v34i07.6690

[2] Li, Z., Chen, H., Biggio, B., He, Y., Cai, H., Roli, F., & Xie, L. (2024). Toward effective traffic sign detection via two-stage fusion neural networks. IEEE Transactions on Intelligent Transportation Systems, 25(8), 8283-8294. https://doi.org/10.1109/TITS.2024.3373793

[3] Tang, S., Zhou, Y., Li, J., Liu, C., & Shi, J. (2024). Attention-Guided Sample-Based Feature Enhancement Network for Crowded Pedestrian Detection Using Vision Sensors. Sensors, 24(19), 6350. https://doi.org/10.3390/s24196350

[4] Yang, H. (2024, November). Small Object Detection Based on Self-attention and Multi-scale Feature Fusion. In International Conference on Information Processing and Network Provisioning (pp. 239-253). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-95-1334-5_20

[5] Xiao, Y., Zhou, K., Cui, G., Jia, L., Fang, Z., Yang, X., & Xia, Q. (2021). Deep learning for occluded and multi-scale pedestrian detection: A review. IET Image Processing, 15(2), 286-301. https://doi.org/10.1049/ipr2.12042Digital Object Identifier

[6] Ruan, J., Cui, H., Huang, Y., Li, T., Wu, C., & Zhang, K. (2023). A review of occluded objects detection in real complex scenarios for autonomous driving. Green energy and intelligent transportation, 2(3), 100092. https://doi.org/10.1016/j.geits.2023.100092

[7] Zou, J., Zheng, H., & Wang, F. (2023). Real-Time target detection system for intelligent vehicles based on multi-source data fusion. Sensors, 23(4), 1823. https://doi.org/10.3390/s23041823

[8] Li, F., Li, X., Liu, Q., & Li, Z. (2022). Occlusion handling and multi-scale pedestrian detection based on deep learning: A review. IEEE Access, 10, 19937-19957. https://doi.org/10.1109/ACCESS.2022.3150988

[9] Zhao, X., Dou, X., Zheng, J., & Zhang, G. (2025). A Lightweight Multi-Stage Visual Detection Approach for Complex Traffic Scenes. Sensors, 25(16), 5014. https://doi.org/10.3390/s25165014

[10] Chen, Y., Wu, Y., Cui, X., Li, Q., Liu, J., & Niu, W. (2024). Reflective adversarial attacks against pedestrian detection systems for vehicles at night. Symmetry, 16(10), 1262. https://doi.org/10.3390/sym16101262

[11] Sun, Z., Wei, G., Fu, W., Ye, M., Jiang, K., Liang, C., ... & Mukherjee, M. (2024). Multiple pedestrian tracking under occlusion: A survey and outlook. IEEE Transactions on Circuits and Systems for Video Technology, 35(2), 1009-1027. https://doi.org/10.1109/TCSVT.2024.3481425

[12] Ling, H., Wang, Z., Li, P., Shi, Y., Chen, J., & Zou, F. (2019). Improving person re-identification by multi-task learning. Neurocomputing, 347, 109-118. https://doi.org/10.1016/j.neucom.2019.01.027

[13] Song, Y., & Lu, Y. (2025). A Review of Unmanned Visual Target Detection in Adverse Weather. Electronics, 14(13), 2582. https://doi.org/10.3390/electronics14132582

[14] Ni, P., Li, X., Kong, D., Wei, K., & Hu, Y. (2023). Scene-adaptive 3-D semantic segmentation method based on multiphase edge enhancement for intelligent vehicles. IEEE Sensors Journal, 23(24), 31471-31482. https://doi.org/10.1109/JSEN.2023.3329389

[15] Lei, M., Song, Y., Zhao, J., Wang, X., Lyu, J., Xu, J., & Yan, W. (2022). End-to-end network for pedestrian detection, tracking and re-identification in real-time surveillance system. Sensors, 22(22), 8693. https://doi.org/10.3390/s22228693

[16] Song, F., & Li, P. (2023). YOLOv5-MS: Real-time multi-surveillance pedestrian target detection model for smart cities. Biomimetics, 8(6), 480. https://doi.org/10.3390/biomimetics8060480

[17] Yang, R., Yan, Z., Yang, T., Wang, Y., & Ruichek, Y. (2023). Efficient online transfer learning for road participants detection in autonomous driving. IEEE Sensors Journal, 23(19), 23522-23535. https://doi.org/10.1109/JSEN.2023.3305592

[18] Gao, H., Huang, S., Li, M., & Li, T. (2024). Multi-scale structure perception and global context-aware method for small-scale pedestrian detection. IEEE Access, 12, 76392-76403. https://doi.org/10.1109/ACCESS.2024.3406968

[19] Spaulding, A., & Middleton, D. (2003). Optimum reception in an impulsive interference environment-Part I: Coherent detection. IEEE Transactions on Communications, 25(9), 910-923. https://doi.org/10.1109/TCOM.1977.1093943

[20] Gao, C., Zhao, F., Zhang, Y., & Wan, M. (2024). Research on multitask model of object detection and road segmentation in unstructured road scenes. Measurement Science and Technology, 35(6), 065113. https://doi.org/10.1088/1361-6501/ad35dd

[21] Xiao, Y., Zhou, K., Cui, G., Jia, L., Fang, Z., Yang, X., & Xia, Q. (2021). Deep learning for occluded and multi-scale pedestrian detection: A review. IET Image Processing, 15(2), 286-301. https://doi.org/10.1049/ipr2.12042Digital Object Identifier

[22] Yuan, W. (2025, June). Graph neural network-based multimodal sensor fusion for robust autonomous driving perception. In Second International Conference on Intelligent Transportation and Smart Cities (ICITSC 2025) (Vol. 13682, pp. 26-35). SPIE. https://doi.org/10.1117/12.3073578

[23] Huang, X., Zeng, L., Liang, H., Li, D., Yang, X., & Zhang, B. (2024). Comprehensive walkability assessment of urban pedestrian environments using big data and deep learning techniques. Scientific Reports, 14(1), 26993. https://doi.org/10.1038/s41598-024-78041-x

[24] Wu, P., Zhang, Z., Peng, X., & Wang, R. (2024). Deep learning solutions for smart city challenges in urban development. Scientific Reports, 14(1), 5176. https://doi.org/10.1038/s41598-024-55928-3

[25] Sun, C., Zhang, R., Lu, Y., Cui, Y., Deng, Z., Cao, D., & Khajepour, A. (2023). Toward ensuring safety for autonomous driving perception: Standardization progress, research advances, and perspectives. IEEE Transactions on Intelligent Transportation Systems, 25(5), 3286-3304. https://doi.org/10.1109/TITS.2023.3321309