Smart City Application: Real-Time Pedestrian Detection Using YOLO11 Architecture

Mark Xavier Dsouza[a], Adavayya Charantimath[b], C. Hithin Kumar[c], Tejas R R Shet[d]
and Shashank Hegde

School of Computer Science and Engineering, KLE Technological University, Hubballi, India

Keywords: Pedestrian Detection, Vehicle Detection, YOLO11, Urban Traffic Monitoring, Deep Learning, Edge Devices.

Abstract: The integration of real-time pedestrian and vehicle detection systems is vital for smart city applications, addressing challenges like traffic management and pedestrian safety. This paper proposes a scalable and resource-efficient framework based on YOLO11. The model leverages features like CSP-Darknet, Spatial Pyramid Pooling (SPP), and Soft Non-Maximum Suppression (Soft-NMS) to ensure accuracy and low latency. Achieving a mean Average Precision (mAP) of 88.0%, the system excels in urban scenarios, including crowded and low-light conditions. This research bridges theoretical advancements and real-world deployment, aiming for smarter, safer cities.

1 INTRODUCTION

The rapid growth of urbanization and vehicular traffic calls for real-time pedestrian detection systems in smart city infrastructure (Zhang et al., 2020). The systems have been crucial for improving pedestrian safety, controlling congestion, and ensuring right-of-way compliance (Redmon et al., 2016). Real-time and accurate pedestrian detection is vital for applications such as autonomous vehicles, intelligent traffic systems, and urban surveillance, where every millisecond counts (Zhang et al., 2020). Real-time pedestrian detection bridges the gap between cutting-edge research and practical implementation by addressing challenges such as occlusions, dynamic lighting, and overlapping objects, making it indispensable for creating safer urban environments (Jiang et al., 2019).

Advancements in Artificial Intelligence (AI), particularly Deep Learning, have revolutionized computer vision tasks like object detection (Liu et al., 2016). Early methods, such as Haar cascades and Support Vector Machines, relied on handcrafted features but struggled in real-world urban scenarios due to occlusions, dynamic lighting, and scale variations (Bochkovskiy et al., 2020). Convolutional Neural Networks (CNNs) marked a breakthrough by enabling data-driven feature extraction (He et al., 2016). Region-based methods such as Faster R-CNN improved localization accuracy but were computationally expensive, leading to single-shot detectors like YOLO, which can process an entire image in a single forward pass (Ren et al., 2015). The evolution of YOLO from YOLOv1 to YOLOv8 has brought features such as anchor boxes, multi-scale detection, and advanced architectures to balance speed and accuracy for real-time applications (Redmon et al., 2016).

The newest iteration, YOLO11, utilizes the best of techniques such as CSPDarknet, Spatial Pyramid Pooling (SPP), and Soft Non-Maximum Suppression (Soft-NMS) and does well in crowded urban scenarios with challenges such as occlusions and changing lighting conditions (Jiang et al., 2019). The research will make use of YOLO11 to propose a scalable, automated real-time pedestrian detection system for use in urban environments (Wang et al.,). Its high accuracy and low latency make it suitable for traffic management, pedestrian safety enforcement, and urban planning, even on resource-constrained devices (Brown and Green, 2022). The integration of YOLO11 into smart city frameworks highlights its potential to enhance pedestrian safety, reduce traffic-related incidents, and facilitate efficient urban management, aligning with the overarching goals of creating smarter, safer, and more sustainable cities(Jiang

[a] https://orcid.org/0009-0003-1155-7621
[b] https://orcid.org/0009-0001-9740-8904
[c] https://orcid.org/0009-0007-1745-3670
[d] https://orcid.org/0009-0004-9353-5676

Dsouza, M. X., Charantimath, A., Kumar, C. H., R R Shet, T. and Hegde, S.
Smart City Application: Real-Time Pedestrian Detection Using YOLO11 Architecture.
DOI: 10.5220/0013611000004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 3*, pages 167-172
ISBN: 978-989-758-763-4

et al., 2019).

This paper provides an overview of our research journey. Section 2 reviews related work on pedestrian and vehicle detection, highlighting urban challenges. Section 3 presents our approach, explaining the YOLO11 architecture and real-time detection methodology. Section 4 discusses experimental results, including key metrics like mAP, precision, and recall. Finally, Section 5 summarizes our findings and suggests future research, including low-light detection improvements and predictive analytics for traffic management.

2 LITERATURE REVIEW

The detection of pedestrians and vehicles has become a critical area of research in urban traffic management, safety systems, and smart city frameworks. Early approaches to object detection were primarily based on manually crafted features and machine learning models. Techniques such as the Viola-Jones cascade classifier (Viola and Jones, 2001) and Histograms Of Oriented Gradients (HOG) (Dalal and Triggs, 2005) formed the foundation for identifying objects in constrained settings. While effective in controlled environments, these methods struggled with challenges inherent to urban landscapes, including occlusions, fluctuating lighting, and the diverse appearance of pedestrians and vehicles.

With the emergence of deep learning, the field of object detection witnessed a revolutionary shift. Convolutional Neural Networks (CNNs) automated feature extraction and greatly enhanced the robustness and accuracy of detection models. Region-based CNNs, such as Faster R-CNN (Ren et al., 2015), combined region proposal mechanisms with CNNs to achieve highly accurate detections. However, these models relied on computationally expensive multi-stage pipelines, making them unsuitable for real-time applications like live traffic monitoring and pedestrian detection (Girshick, 2015).

To address the limitations of traditional region-based models, single-shot detection frameworks, including the Single Shot MultiBox Detector (SSD) (Liu et al., 2016) and the early iterations of unified detection systems like You Only Look Once (YOLO) (Redmon et al., 2016), reframed object detection as a single regression problem. The streamlined approach processed the entire image in a single pass, achieving real-time performance while maintaining competitive accuracy. Over successive iterations, advancements such as anchor boxes, multi-scale detection, and improved backbone networks allowed these systems to handle challenges like small-object detection and scale variation more effectively (Bochkovskiy et al., 2020).

The latest evolution in this family of models, introduced as version 11, incorporates key innovations including Cross-Stage Partial Networks (CSP), Spatial Pyramid Pooling (SPP), and Soft Non-Maximum Suppression (Soft-NMS). These enhancements significantly improve detection capabilities, particularly in dense and cluttered urban settings. By addressing challenges such as occlusions, overlapping objects, and fluctuating lighting conditions, this architecture achieves an optimal balance of speed and accuracy. Its ability to process live video feeds in real time positions it as a pivotal tool for smart city applications, enhancing pedestrian safety, traffic management, and urban planning (Liang et al., 2023).

This seamless progression from handcrafted methods to advanced detection architectures highlights the transformative impact of deep learning in pedestrian and vehicle detection. While recent advancements have mitigated many challenges, the integration of these systems into scalable and resource-efficient frameworks remains a focus for future research, aligning with the overarching goals of building smarter and safer urban environments.

3 PROPOSED WORK

Building on these advancements, this proposed work proposes a real-time pedestrian and vehicle detection system optimized for deployment in urban environments. The model, based on YOLO11, addresses challenges such as occlusion, small-object detection, and varying lighting conditions. Designed for edge devices, the system ensures efficient operation on resource-constrained hardware while maintaining high accuracy (Li et al., 2024).

3.1 Architecture

The architecture of YOLO11 (Figure.1) is divided into three key components: Backbone, Neck, and Head, each designed to enhance the efficiency and accuracy of pedestrian detection (Zheng et al., 2024). The Backbone is responsible for feature extraction from the input image, starting with an input size of $640 \times 640 \times 3$. It uses successive convolutional layers and C3K2 blocks with shortcut connections to capture both local and global contextual information while reducing the spatial dimensions of the feature maps. These residual modules improve feature propagation and prevent gradient vanishing in deep networks, progressively distilling the input into high-level features critical for pedestrian detection (Gao and Wu, 2024).

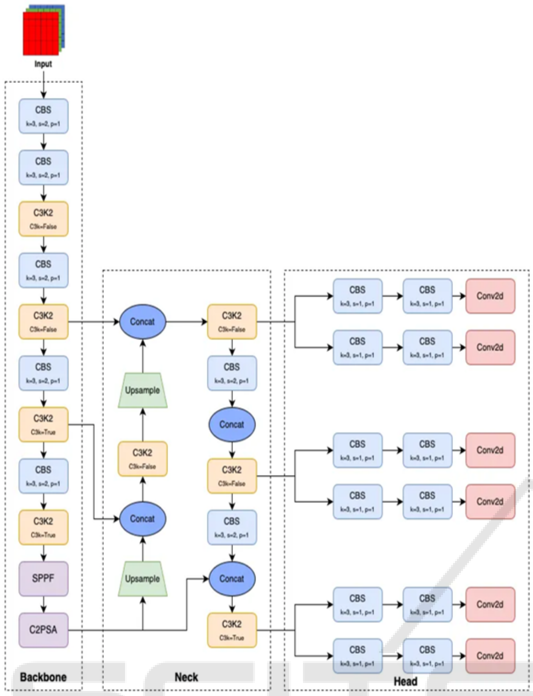

Figure 1: Architecture of YOLO11.(Jegham et al., 2024)

The Neck and Head form the remaining components of the architecture. The Neck serves as a feature aggregation stage, using Upsample and Concat layers to fuse multi-scale features and enhance the model's ability to detect objects of different sizes. Advanced modules such as SPFF (Spatial Pyramid Fast Fusion) and C2PSA (C2 Spatial Attention) are integrated into the Neck, improving the receptive field and refining feature localization through spatial attention (Wang et al., 2023). Finally, the Head is responsible for predicting bounding boxes, class probabilities, and confidence scores. Leveraging multi-scale detection layers, the Head ensures accurate predictions for pedestrians of varying sizes and positions, making YOLO11 highly suitable for real-time detection tasks in complex environments (Li et al., 2024).

3.1.1 Loss Function

The YOLO11 architecture processes images via a CSPDarknet backbone to extract features from any object at any scale robustly. The detection head performs bounding box prediction, class probability prediction, and objectness score prediction together while optimizing for real-time detection. The loss function (Equation 1) involves three key components: classification loss (L_{cls}), objectness loss (L_{obj}), and localization loss (L_{loc}), and their combination is as follows:

$$L = \lambda_{cls} \cdot L_{cls} + \lambda_{obj} \cdot L_{obj} + \lambda_{loc} \cdot L_{loc} \tag{1}$$

Classification loss (Equation 2), calculated using softmax cross-entropy, is defined as:

$$L_{cls} = -\sum_{i=1}^{C} p_i \log(\hat{p}_i) \tag{2}$$

where C is the number of classes, p_i is the true probability for class i, and $\hat{p}_i$ is the predicted probability. The objectness loss (Equation 3), which measures the confidence with which an object exists in the bounding box, is modeled by the binary cross-entropy function as follows:

$$L_{obj} = -[y\log(\hat{y}) + (1-y)\log(1-\hat{y})] \tag{3}$$

where y is the ground truth objectness score and $\hat{y}$ is the predicted score.

The localization loss, computed using Complete IoU (CIoU) (Equation 4), measures the alignment of predicted bounding boxes with ground truth and takes into account distance, overlap, and aspect ratio differences:

$$\mathrm{CIoU} = 1 - \mathrm{IoU} + \frac{\rho^2(b, b_{gt})}{c^2} + \alpha v \tag{4}$$

where $\rho^2(b, b_{gt})$ is the Euclidean distance between the centers of the predicted and ground truth boxes, c^2 is the diagonal length of the smallest enclosing box, and v is an aspect ratio term. This all-inclusive loss function ensures that the model effectively balances classification accuracy, bounding box confidence, and localization precision, which in turn helps it perform well in different urban traffic scenarios.

3.2 Implementation

The proposed work utilizes a customized dataset specifically created for the Hubli-Dharwad Smart City, consisting of 1,000 images annotated with bounding boxes in a YOLO11-compatible format. It focuses on two object classes: pedestrians and vehicles, which are commonly encountered in urban traffic scenarios. To standardize input for the YOLO11 models, all images were resized to a resolution of 640 × 640. Preprocessing techniques, such as data augmentation and normalization, were applied to enhance the dataset's robustness and variability. The data set was divided into 70% for training and 30% for testing, ensuring that the model could learn data patterns effectively during training and be robustly evaluated on unseen samples. Annotations were provided

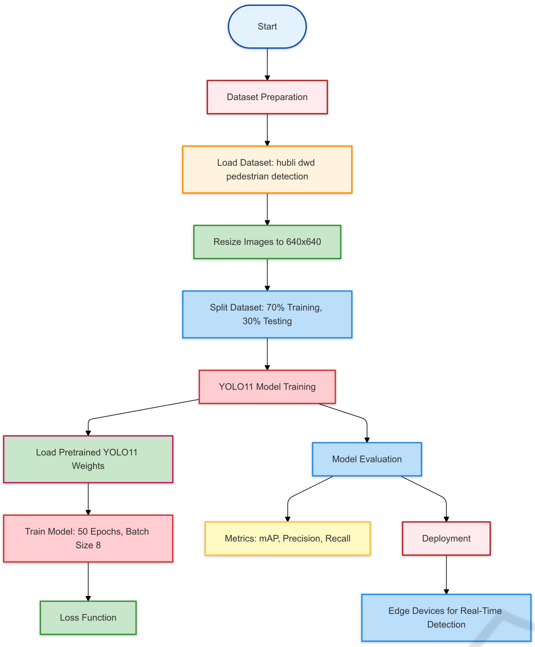
Start
Dataset Preparation
Load Dataset: hubli dwd pedestrian detection
Resize Images to 640x640
Split Dataset: 70% Training, 30% Testing
YOLO11 Model Training
Load Pretrained YOLO11 Weights
Train Model: 50 Epochs, Batch Size 8
Loss Function
Model Evaluation
Metrics: mAP, Precision, Recall
Deployment
Edge Devices for Real-Time Detection

Figure 2: Flowchart of Proposed Work

in YOLO format, describing each bounding box with center coordinates (x, y), width w, height h, and class label. The approach aligns with the YOLO training pipeline and supports efficient computation of bounding box regression, optimizing the model's performance.The proposed system leverages YOLO11 pretrained weights, fine-tuned over 50 epochs on this specific dataset.

A batch size of 8 was employed during training to balance computational efficiency and performance. Evaluation metrics, including mAP (mean average precision), precision, and recall, were used to assess the model's effectiveness. Deployment on edge devices demonstrated YOLO11's computational efficiency, enabling real-time pedestrian and vehicle detection in dynamic environments. This robust and scalable system showcases YOLO11's ability to tackle the challenges of real-time object detection in complex urban landscapes.

4 RESULTS AND ANALYSIS

The performance of the YOLO11 model was evaluated for its ability to detect pedestrians and vehicles in urban traffic scenarios using precision, recall, and mean Average Precision (mAP) as key evaluation metrics. The findings highlight the strengths of the model while identifying areas that require further development.

The Precision Recall (PR) curve (Figure.3) illustrates the detection capabilities for pedestrians and vehicles. An overall mAP of 88.0% was achieved at an Intersection over Union (IoU) threshold of 0.5. Vehicle detection exhibited superior performance with an mAP of 89.1%, while pedestrian detection lagged behind at 86.8%. The disparity underscores the effectiveness of the model in detecting larger, distinct objects such as vehicles, while revealing difficulties with smaller or partially occluded objects such as pedestrians. Furthermore, the PR curve demonstrates consistent precision across recall levels for vehicles, contrasting with a noticeable decrease in precision for pedestrians at higher recall values.

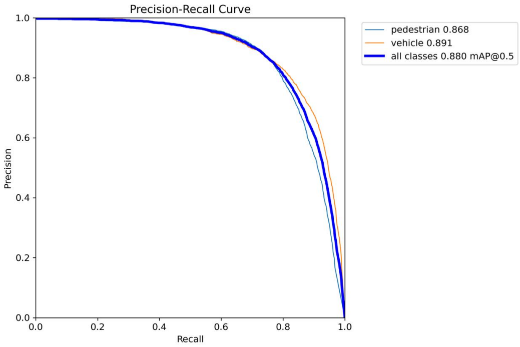

Figure 3: Precision-Recall curve highlighting performance across two classes: pedestrians and vehicles.

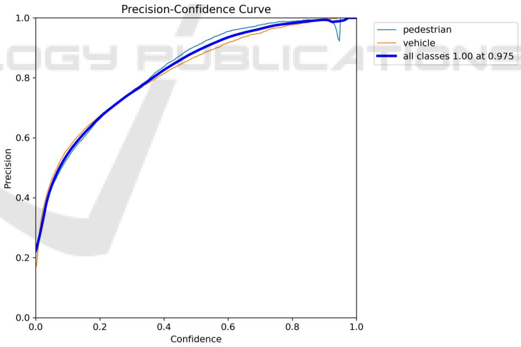

Figure 4: Precision-Confidence curve illustrating the performance of the YOLO model at different confidence thresholds.

The Precision-Confidence curve (Figure.4) illustrates the performance of our YOLO model, trained to detect pedestrian and vehicle classes. The orange and light blue lines represent class-wise precision at varying confidence levels, while the bold blue line denotes the overall performance, achieving a precision of 1.00 at a confidence threshold of 0.975. The curve highlights the model's reliability across confidence ranges and serves as a basis for selecting an optimal confidence threshold to balance precision and recall in real-world applications. Such an analysis is essential for assessing the robustness of object detection models.

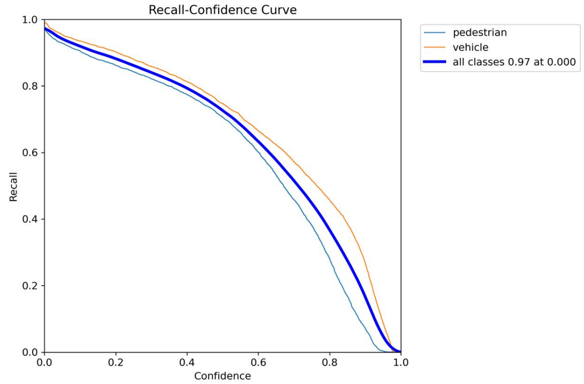

Figure 5: Recall-Confidence curve demonstrating the trade-off between recall and confidence thresholds.

The Recall-Confidence curve (Figure.5) provides a clearer picture of the trade-off between recall and confidence thresholds. At a confidence threshold of 0.0, the model achieved a maximum recall of 0.97, demonstrating its ability to detect the majority of objects under relaxed confidence conditions. However, as the confidence threshold increased, recall began to decline, highlighting the inherent trade-off between detecting as many objects as possible and ensuring high precision.

Table 1: Comparison between YOLOv10 and YOLO11

Metric	Precision (P)	Recall (R)	mAP@50
YOLOv10	0.756	0.708	0.794
YOLO11	0.84	0.782	0.88

The YOLO11 model achieves an impressive inference time of 7-10ms per frame, making it highly suitable for real-time traffic management and pedestrian safety applications. While the results are promising, challenges persist in handling occlusions and detecting smaller objects. Despite these limitations, YOLO11 shows a strong potential for real-time object detection, particularly vehicle detection, but refining pedestrian detection and conducting extensive real-world evaluations are essential to maximize its effectiveness in smart city infrastructure.

5 CONCLUSION AND FUTURE WORK

YOLO11 excels in real-time detection of pedestrians and vehicles in urban traffic, utilizing CSPDarknet for feature extraction and Soft-NMS for managing overlapping objects. It demonstrates strong performance in crowded environments, varying lighting conditions, and small object detection, significantly aiding in effective traffic management and enhancing road safety.

Improvements on low-light detection, object recognition of bicycles and road signs, partially covered objects, and optimization for edge devices will be part of future work. Traffic prediction tools will also be added, and testing in live traffic can provide valuable insights for further enhancement.

REFERENCES

Bochkovskiy, A. et al. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint*, arXiv:2004.10934.

Brown, R. and Green, E. (2022). Transformers in computer vision: A comprehensive review. *Journal of Vision Technology*, 12(4):45–67.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 886–893.

Gao, L. and Wu, Y. (2024). Future directions in optimizing yolo models for resource-constrained environments. *Journal of Artificial Intelligence and Applications*.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448.

He, K. et al. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Jegham, N., Koh, C. Y., Abdelatti, M., and Hendawi, A. (2024). Evaluating the evolution of yolo (you only look once) models: A comprehensive benchmark study of yolo11 and its predecessors.

Jiang, L. et al. (2019). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1717–1728.

Li, J., Zhang, Y., and Xie, L. (2024). Ai-powered object detection: From yolo to advanced architectures. *arXiv preprint arXiv:2401.01234*.

Liang, X. et al. (2023). Yolov11: Advancing real-time object detection with enhanced features and efficiency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, W. et al. (2016). Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–37.

Redmon, J. et al. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788.

Ren, S. et al. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 91–99.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 511–518.

Wang, C.-Y. et al. (2023). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint*, arXiv:2207.02696.

Wang, K. et al. Yolov11: State-of-the-art object detection for next-generation applications. *arXiv preprint arXiv:2304.12345*.

Zhang, W., Wang, K., and Yang, S. (2020). A review on pedestrian detection based on deep learning. *Neural Computing and Applications*, 32(5):1515–1532.

Zheng, L. et al. (2024). Improvement of the yolov8 model in the optimization of the weed recognition algorithm in cotton field. *Plants*, 13(13):1843.