# A Modular Detection System for Smart Cities: Integrating Monocular and LiDAR Solutions for Scalable Traffic Monitoring

Javier Borau-Bernad[1][a], Álvaro Ramajo-Ballester[1][b], José María Armingol Moreno[1][c] and Araceli Sanchis de Miguel[2][d]

[1] *Intelligent Systems Lab, University Carlos III of Madrid, Av. de la Universidad, 30, Leganés, 28911, Madrid, Spain*

[2] *Control Learning and Systems Optimization Group, University Carlos III of Madrid, Av. de la Universidad, 30, Leganés, 28911, Madrid, Spain*

Keywords: 3D Object Detection, Intelligent Infrastructures, Smart Cities, Autonomous Driving, Deep Learning.

Abstract: As smart cities continue to develop, they require scalable and efficient traffic monitoring systems. This paper presents a modular detection system that switches between monocular and multimodal modes, depending on the available sensors. The monocular mode, based on the MonoLSS algorithm, offers a cost-effective vehicle detection solution using a single camera, ideal for simpler or low-budget setups. In contrast, the multimodal mode integrates camera and LiDAR data via the MVX-Net model, enhancing 3D accuracy in complex traffic scenarios. This dual-mode flexibility allows smart cities to adapt the system to their infrastructure and budgetary needs, ensuring scalability as urban demands evolve. Inference results demonstrate the superior accuracy of the multimodal approach in challenging environments while validating the efficiency of the monocular mode for simpler settings. Therefore, the modular detection system offers a flexible solution that optimizes both cost and performance, effectively addressing the varied requirements of smart city traffic management.

## 1 INTRODUCTION

As urban environments continue to grow and become more complex, the rise of smart cities represents a solution to make them more efficient, sustainable, and secure, improving the quality of life for their residents. Smart cities leverage cutting-edge technologies to create advanced systems that manage resources, optimize energy usage and enhance transportation systems (Zanella et al., 2014). Among these systems, intelligent vehicle detection plays a key role in improving traffic management, advancing autonomous vehicle capabilities and increasing road safety. However, many current detection solutions either lack sufficient data quality for real-time analysis or rely on the widespread adoption of autonomous vehicles in the traffic system (Jain et al., 2019). Addressing these limitations requires the development of scalable solutions that can deliver high quality traffic data analysis.

Intelligent infrastructures emerge as a key component for traffic monitoring by integrating sensors and analysis algorithms into urban elements such as traffic cameras and bridge-mounted surveillance systems. These infrastructure-based sensors offer a broader field of view and reduce occlusions between vehicles, allowing detection algorithms to significantly improve precision (Borau Bernad et al., 2024). The strategic placement of roadside sensors also plays a key role in improving detection accuracy and minimizing blind spots, particularly in urban areas where efficient coverage is essential(Owais and Shahin, 2022). This is particularly beneficial in high-risk areas, like intersections, where dense traffic and low visibility complicate effective monitoring.

Despite these advantages of intelligent infrastructures, the challenge of developing a flexible solution that adapts to different precision requirements appears in the research paradigm, to guarantee optimal performance across diverse urban settings. Monocular detection systems present themselves as an affordable and simple solution for 3D vehicle detection, but are challenged by their limitations in depth estimation and lighting conditions. In contrast, multimodal detection systems, which integrate LiDAR and monocular technologies, provide higher 3D accuracy but come with increased installation costs (Wang et al., 2023). The challenge in this context lies in designing a flexible solution that can satisfy different installation needs while balancing cost and performance.

[a] https://orcid.org/0009-0009-5623-1688
[b] https://orcid.org/0000-0001-9425-9408
[c] https://orcid.org/0000-0002-3353-9956
[d] https://orcid.org/0000-0002-1429-4092

Borau-Bernad, J., Ramajo-Ballester, Á., Moreno, J. M. A. and Sanchis de Miguel, A.
A Modular Detection System for Smart Cities: Integrating Monocular and LiDAR Solutions for Scalable Traffic Monitoring.
DOI: 10.5220/0013195200003941

In *Proceedings of the 11th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2025)*, pages 326-333
ISBN: 978-989-758-745-0; ISSN: 2184-495X

This paper presents a modular detection system designed to operate in two distinct modes, offering flexibility and adaptability to diverse urban infrastructures. The system can work as a monocular solution, using the MonoLSS algorithm (Li et al., 2024), which provides a cost-effective and simpler approach for vehicle detection. Alternatively, it can operate as a multimodal system, integrating both LiDAR and monocular data through the MVX-Net model (Sindagi et al., 2019) to enhance 3D accuracy in more demanding environments. Unlike previous approaches that rely on fixed monocular or multimodal configurations, this system introduces a novel dynamic mode selection based on sensor availability. This modular design enhances the scalability of smart infrastructures, enabling cities to deploy the solution based on their specific needs, budget and available sensor installations, making it versatile across different urban settings.

## 2 RELATED WORKS

The growing interest in developing technologies for smart cities has fueled the advancement of intelligent traffic monitoring systems. While monocular methods provide an affordable and simple solution, they struggle with depth estimation errors and rely heavily on certain visibility conditions for optimal performance. In contrast, LiDAR and multimodal systems offer improved detection accuracy and perform well under low visibility conditions, though at a higher system cost (Arnold et al., 2019). This section reviews the key advancements in vehicle detection systems, highlighting state-of-the-art models, their limitations and the need for a flexible modular solution that can harness the strengths of each approach based on the specific needs of intelligent infrastructure.

Monocular detection systems have emerged as an attractive solution for urban data acquisition due to their affordability and ease of deployment. These systems consist of a single-camera setup that captures 2D images, which are processed by deep learning algorithms to obtain three-dimensional vehicle information. The simplicity of monocular systems makes them suitable for cities with limited installation budgets, as they only require conventional calibrated camera setups and a computational unit to process the information. This cost-effectiveness and their easy integration into existing infrastructure make monocular systems an interesting tool for large-scale deployment in intelligent traffic monitoring solutions.

During the last decade, several deep learning models have been developed to improve existing monocular detection algorithms. The SMOKE model (Liu et al., 2020) is known for its efficiency, speed and simplicity, performing 3D inference in a single stage. On the other hand, PGD algorithm (Wang et al., 2022) enhances depth estimation by combining probabilistic and geometric methods to improve detection accuracy. MonoCon (Liu et al., 2022) leverages additional contextual information during training, improving detection accuracy in crowded environments. Lastly, MonoLSS introduces a Learnable Sample Selection (LSS) module, optimizing the use of features during 3D property learning and uses MixUp3D augmentation technique to enhance training data, making the algorithm highly robust against occlusions.

Despite these advancements, monocular detection systems still face important limitations, mainly due to errors during depth estimation and their sensitivity to non-optimal environmental conditions. The process of inferring three-dimensional information from 2D images causes the models to have significant errors, which has a direct impact on the final accuracy of the algorithm. In addition, monocular systems are particularly affected by low visibility situations, such as adverse weather conditions, as well as low light during night or direct sunlight hitting the camera lens during sunrise or sunset, notably reducing their capabilities and precision (Qian et al., 2022). These limitations establish the need for integration of more robust solutions, leading to the exploration of LiDAR-based detection systems, which offer enhanced depth perception and accuracy but at a higher cost.

LiDAR-based systems overcome many of the issues of monocular approaches by inferring detection from 3D point clouds generated by laser pulses, resulting in better depth and position accuracy regardless of light conditions, making them more reliable in complex urban environments. Notable LiDAR-based models include SECOND (Yan et al., 2018), which uses sparse convolutions to efficiently process LiDAR point clouds and PointPillars (Lang et al., 2019), a method that organizes point clouds into vertical columns for faster processing. Another advanced approach is VoxelNet (Zhou and Tuzel, 2018), which divides the point cloud space into 3D voxels and extracts features for object detection.

However, LiDAR systems face significant challenges that limit their widespread adoption in Smart Cities. These include high sensor costs, complex installation and maintenance requirements, such as frequent calibration and cleaning, and reduced robustness in adverse weather conditions like heavy rain or snow (Owais, 2024). Table 1 provides a comparative analysis of monocular camera systems and LiDAR-based systems, highlighting the advantages and disadvantages of each sensor type.

Table 1: Comparison of Camera and LiDAR sensors under different environmental conditions, considering the impact of adverse conditions, color capabilities, installation complexity and cost (Zhang et al., 2023).

| Sensor type | Light rain | Heavy rain | Smog | Snow | Strong light | Color | Complexity | Cost |
|---|---|---|---|---|---|---|---|---|
| Camera | Mid | High | High | Mid | High | Yes | Low | Low |
| LiDAR | Low | Mid | Low | High | Low | No | High | High |

Multimodal detection systems combine the strengths of both cameras and LiDAR, integrating 2D image data with 3D point clouds to achieve more accuracy and robustness in vehicle detection. These systems can capture both the visual information needed for tasks like license plate recognition, while also benefiting from the depth and spatial accuracy provided by LiDAR. A good example of this approach is MVX-Net, which fuses camera and LiDAR data to generate highly accurate 3D object detections. By combining the complementary features of both sensors, multimodal systems overcome the limitations of using only image or LiDAR solutions, particularly useful in highly demanding traffic situations with poor visibility or occlusions. Despite these advantages, the integration of multiple sensors increases both the hardware costs and the complexity of the system, presenting challenges for widespread deployment in cost-sensitive urban infrastructures.

Monocular, LiDAR and multimodal systems each offer distinct advantages but also come with trade-offs in cost, complexity and performance. Monocular systems are affordable but struggle with depth accuracy, while LiDAR systems are precise but expensive and limited in capturing visual details like color or license plates and multimodal systems provide a balance but increase overall costs. To address these challenges, this paper proposes a modular detection system that operates in monocular and multimodal modes, allowing for flexible deployment based on urban infrastructure needs, balancing cost and accuracy.

# 3 OUR APPROACH: MODULAR DETECTION SYSTEM

To address the challenges resulting from varied traffic demands, this section introduces a modular detection system for intelligent infrastructures that can operate in both monocular and multimodal modes, depending on the inputs received from the sensors. The system dynamically selects the appropriate mode based on real-time data to maximize the effectiveness of the intelligent infrastructure. This solution provides the flexibility needed to adapt to different urban environments while balancing cost and performance requirements, guaranteeing scalability for smart cities.

## 3.1 System Architecture

The modular detection system is designed to efficiently handle both monocular and multimodal data inputs, enabling it to adapt to various urban infrastructure setups. The system autonomously selects between these modes based on the sensors available at the infrastructure setup, ensuring optimal operation regardless of the installed sensors and making it an adaptable solution for 3D data acquisition. This adaptability is critical for ensuring that diverse urban environments, with different levels of technological infrastructure, can still benefit from the detection system's advanced capabilities.

The system is built around two core components: one for monocular detection and another for multimodal detection, which are automatically selected by the system depending on the installed sensors. If only camera data is available, the system activates the monocular mode, processing the images with a monocular 3D detection model to perform 3D vehicle detection in cost-effective configurations. In contrast, when both camera and LiDAR are connected, the system switches to multimodal mode, leveraging enhanced detection capabilities for more complex situations. This transition from monocular to multimodal mode requires minimal infrastructure modifications. Existing monocular setups can be upgraded with LiDAR sensors and updating software to enable multimodal data processing.

The applications of each module vary based on the needs of the infrastructure environment. Monocular mode is well-suited for areas with budget constraints or less complex traffic, such as smaller cities or high-visibility intersections where camera-based systems can guarantee accurate vehicle detection. On the other hand, multimodal mode is designed for more demanding situations, integrating LiDAR and camera data to provide enhanced detection in complex traffic conditions. This mode is ideal for busy intersections, environments with poor visibility, or locations frequently affected by bad weather conditions.

Additionally, the flexibility and ease of installation of the modular system provide significant advantages for smart cities development. Monocular setups can be integrated into existing infrastructure cameras, allowing for affordable deployment and scalability over time. As cities evolve and develop, the system can easily transition from monocular to multimodal mode without manual intervention, enabling more advanced detection capabilities without requiring complete infrastructure redesign. This scalability makes the system a sustainable long-term solution, allowing cities to gradually enhance their traffic monitoring capabilities as resources and infrastructure allow. The system's modular design ensures that new technologies, like AI-based traffic management platforms, advanced sensor fusion algorithms or next-generation LiDAR and high resolution camera systems, can be incorporated with minimal disruption, allowing for smooth upgrades as sensor technology advances. Figure 1 illustrates the overall system architecture and the module selection depending on the input available from the infrastructure installed sensors.

## 3.2 Monocular

The monocular detection component of the modular detection system is based on MonoLSS model (Li et al., 2024), an advanced and well-proven algorithm for 3D object detection using images. MonoLSS is based on CenterNet (Zhou et al., 2019) and first creates a feature map from the image inputs using a backbone DLA-34 (Yu et al., 2018). Then, it leverages a Learnable Sample Selection (LSS) module which identifies the most relevant features from the image to improve the precision of vehicle detection. This LSS module is based in Gumbel-Softmax and a relative distance sample divider and uses a warm-up schedule to improve training stability. In addition, MonoLSS introduces a new data augmentation technique named MixUp3D, designed to simulate spatial overlap of objects without introducing depth ambiguity, which could cause errors in 3D parameter calculations.

The monocular detection module extracts 3D information, including depth, dimensions and orientation of vehicles, using the MonoLSS model. To achieve accurate 3D localization, the system relies on the camera calibration parameters to transform from image to real-world coordinates. This approach ensures consistent performance in real-time traffic situations, providing reliable 3D information even with a single camera, making it ideal for low-cost deployments in less complex traffic environments.

## 3.3 Multimodal

The multimodal detection branch of the modular detection system is based on the MVX-Net model (Sindagi et al., 2019). This algorithm combines LiDAR and monocular image data to extract accurate 3D information, reducing false negatives and false positives compared to other 3D detection approaches. MVX-Net proposes two data fusion techniques, VoxelFusion and PointFusion, to integrate information from both sensor types at different detection stages. For the implementation of MVX-Net into the modular system, PointFusion was selected due to its slightly better performance compared to VoxelFusion.

PointFusion is an early fusion technique that first uses a 2D detection network to extract the feature maps from the images. Then, the 3D points are projected onto the image using the camera calibration matrix and the image feature of each location is concatenated to the correspondent LiDAR point. These combined features are processed through Voxel Feature Encoding (VFE) layers of VoxelNet (Zhou and Tuzel, 2018), allowing the neural network to learn from LiDAR and image data. This early fusion method provides richer information resulting in higher accuracy during 3D object detection.

The multimodal detection system using MVX-Net combines LiDAR and image data to utilize the strengths of both modalities. This fusion approach ensures improved object detection in complex urban scenarios, where precision is essential to guarantee citizen safety and enhance traffic management.

# 4 EXPERIMENTS

## 4.1 Dataset

For training and implementing the algorithms that set up the modular detection system, it is necessary to select a dataset that includes both image and LiDAR data and which mirrors the design of the future application, in this case, infrastructure-based data. The DAIR-V2X dataset (Yu et al., 2022) satisfies these requirements, providing data from LiDAR and RGB cameras captured from vehicles and roadside infrastructures. In this work, the infrastructure part of the dataset, DAIR-V2I, was selected.

The subset originally includes 10,000 images captured with RGB 1920x1080 cameras, along with point cloud data from LiDAR sensors which offer a horizontal field of view (FOV) of 100° and a range of up to 280 meters. The dataset contains almost 500,000 labeled annotations, including bounding boxes, object classes and calibration files that provide camera intrinsic parameters and camera-LiDAR calibration information. During model training, 7,058 images and corresponding LiDAR data from the dataset were used, split into 80% for training, 10% for validation and 10% for testing, ensuring a diverse data distribution that enables the model to generalize correctly.

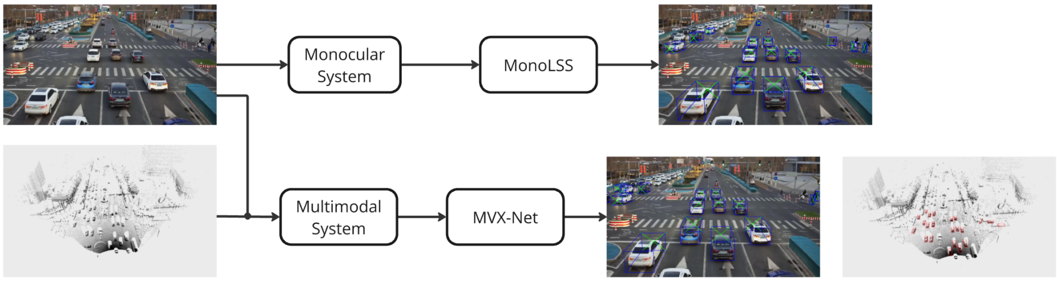

Figure 1: Overview of the modular detection system, illustrating the two core components for monocular and multimodal detection. The system dynamically selects the detection mode based on the available sensor inputs, enabling flexible and scalable deployment in different urban environments.

## 4.2 Implementation Details

The algorithms of the modular detection system were implemented and trained on a single server equipped with two NVIDIA GeForce GTX 3090 Ti.

The MonoLSS model was implemented using the RoadVision3D library (Borau Bernad, 2024), which is a custom library developed in PyTorch designed specifically for 3D vehicle detection and traffic monitoring applications. The images from the DAIR dataset were processed in their original resolution of 1920x1080 pixels and then downsampled by a factor of 4 during feature extraction. The data augmentation techniques applied include random flipping with a 50% probability, random cropping with a 50% probability, a maximum scaling factor of 0.4 and image center shifting of 0.1. Additionally, the MixUp3D augmentation technique was applied with a 50% probability to simulate overlapping objects. To evaluate the impact of augmentation techniques, we conducted an ablation study comparing training with augmentation techniques against training without augmentation for underrepresented classes like Pedestrians and Cyclists. Augmentation techniques significantly improved detection accuracy, increasing AP40 for Pedestrians and Cyclists by 3.6% and 2%, respectively, under moderate difficulty. MonoLSS training was conducted over 150 epochs, with validation at intervals every 10 epochs, using the Adam optimizer. The initial learning rate was set to 0.001, with a warmup phase during the initial 5 epochs, followed by a decay rate of 0.1 at epochs 90 and 120.

For the multimodal detection branch, the MVX-Net model was trained using the MMDetection3D (MMDetection3D Contributors, 2020) library, based in PyTorch, which allows for the integration of LiDAR and image data to enhance 3D object detection. MVX-Net was trained using a voxel size of [0.05, 0.05, 0.1] and a point-cloud range from [0, -40, -3] to [100, 40, 1], with images loaded at their original resolution of 1920x1080. The data augmentation techniques included random resizing of images between (1536, 864) and (2304, 1296), random flipping with a 50% probability and random shuffling of point cloud data. The training lasted for 60 epochs with an AdamW scheduler, validating every 5 epochs and used a cosine annealing learning rate that decayed from 0.003 to $3 \times 10^{-6}$.

## 4.3 Evaluation Metrics

The evaluation of the 3D detection models was conducted using the Average Precision (AP) metric, commonly used in algorithm comparison on the KITTI dataset (Geiger et al., 2012), a standard benchmark for evaluating 3D object detection models. For comparing the performance of the trained MonoLSS and MVX-Net algorithms, we used the $\text{AP}_{40}$ metric, which classifies ground truth objects into three levels of difficulty: easy, moderate and hard, based on factors such as occlusions, pixel size of bounding boxes and truncation. $\text{AP}_{40}$ is well suited for 3D traffic monitoring applications as it combines recall and precision into a single metric providing a reliable measure of overall detection performance across different scenarios.

The Intersection over Union (IoU) metric is used to determine the overlap between predicted and ground truth bounding boxes. The IoU filters detections, ensuring that only those meeting a minimum overlap threshold are considered in the accuracy calculation. Therefore, the Average Precision at 40 recall points ($\text{AP}_{40}$) is calculated using the following equation:

$$AP_{R_N} = \frac{1}{N} \sum_{r \in R} P(r), \tag{1}$$

where $R = [r_0, r_0 + \frac{r_1 - r_0}{N-1}, r_0 + \frac{2(r_1 - r_0)}{N-1}, \ldots, r_1]$ defines $N$ equally spaced recall values between $r_0 = 0$ and $r_1 = 1$, and $P(r)$ is the maximum precision observed for any recall value greater than or equal to $r$.

## 4.4 Results

This section presents the performance evaluation of the modular detection system, examining both the monocular and multimodal modules. Each model was tested using the $AP_{40}$ metric for the detection of three object classes. The evaluation uses Intersection over Union thresholds of 0.70 for Car and 0.50 for Pedestrian and Cyclist to determine true positive detections. The results, shown in Tables 2 and 3, provide a comprehensive overview of how the models handle different difficulty levels, demonstrating the robustness of each approach in dynamic urban environments.

Table 2: Results of MonoLSS Model on $AP_{40}$ 3D detection for Car, Pedestrian and Cyclist classes at different difficulty levels (Easy, Moderate and Hard), with IoU thresholds of 0.70 for Car and 0.50 for Pedestrian and Cyclist.

| Class | Easy | Moderate | Hard |
|---|---|---|---|
| Car | 62.24 | 52.16 | 52.08 |
| Pedestrian | 15.75 | 14.70 | 14.60 |
| Cyclist | 39.44 | 21.04 | 21.35 |

Table 3: Results of MVX-Net Model on $AP_{40}$ 3D detection for Car, Pedestrian and Cyclist classes at different difficulty levels (Easy, Moderate and Hard), with IoU thresholds of 0.70 for Car and 0.50 for Pedestrian and Cyclist.

| Class | Easy | Moderate | Hard |
|---|---|---|---|
| Car | 69.13 | 56.74 | 56.78 |
| Pedestrian | 63.47 | 59.14 | 59.03 |
| Cyclist | 61.49 | 29.36 | 31.04 |

The results show that both MonoLSS and MVX-net perform correctly for most object classes and difficulty levels, proving accurate 3D detection capabilities of the modular system. However, the MonoLSS model shows limitations in detecting pedestrians (14.70%) and cyclists (21.04%) at moderate difficulty, showcasing the limited reliability of monocular detection systems for less visible objects. Despite this, MonoLSS remains accurate for vehicle detection, making it suitable for low-cost implementations where cars are the primary concern. Such setups are appropriate in environments with minimal pedestrian or cyclist traffic, such as highways or districts farther from the city center. On the other hand, MVX-Net shows increased performance in all classes due to its multimodal input source, which highlights the advantage of incorporating LiDAR sensors into intelligent infrastructures. Table 4 shows these accuracy improvements, particularly for pedestrians and cyclists.

Table 4: Comparison of $AP_{40}$ 3D detection performance for Car, Pedestrian and Cyclist classes at an IoU threshold of 0.70 for Car and 0.50 for Pedestrian and Cyclist between MonoLSS and MVX-Net at Moderate difficulty level. This comparison highlights the key role of multimodal detection in improving reliability, especially for smaller objects like pedestrians and cyclists.

| Class | MonoLSS | MVX-Net | Diff. |
|---|---|---|---|
| Car | 52.16 | 56.74 | +4.58 |
| Pedestrian | 14.70 | 59.14 | +44.44 |
| Cyclist | 21.04 | 29.36 | +8.32 |

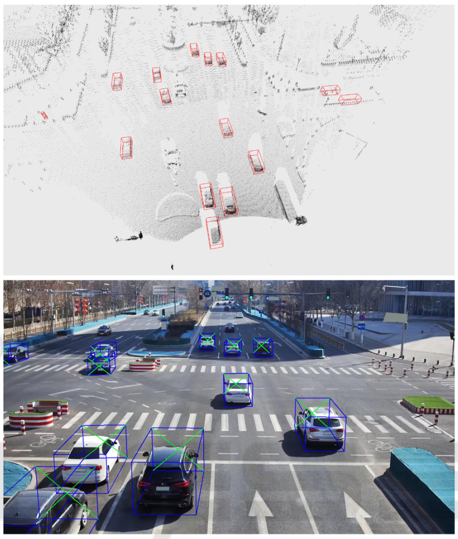

Figure 2: Inference results of the MVX-Net model, illustrating 3D object detection using LiDAR and monocular image data, with results projected onto the point cloud and image. This figure showcases MVX-Net's accuracy and robustness in vehicle detection using multimodal inputs.

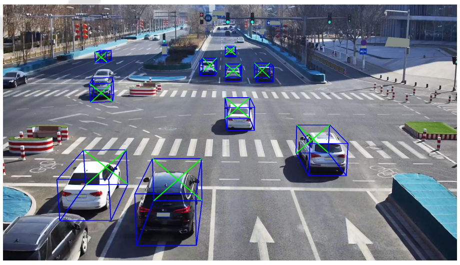

Figure 3: Inference results of the MonoLSS model, demonstrating 3D object detection capabilities using only monocular image data. This figure showcases MonoLSS's reliability in vehicle detection, making it suitable for cost-sensitive scenarios.

In spite of the increased performance of MVX-Net due to the integration of LiDAR, MonoLSS remains a viable and cost-effective solution for simpler applications due to its reliability in vehicle detection. Figures 2 and 3 illustrate the inference results of both models in real-world conditions. These images highlight their performance in detecting vehicles as the primary objective in traffic monitoring scenarios.

### 4.5 Performance Evaluation Under Challenging Conditions

Despite the overall satisfactory results obtained by both models, it is insightful to delve into possible traffic scenarios where the multimodal system might perform better than the monocular, benefiting from the installed LiDAR sensors. One possible scenario involves high-density traffic situations, with occlusions and a large number of vehicles, making detection difficult for a single camera system. This scenario is illustrated in Figure 4, where the multimodal system is correctly detecting the vehicles while the monocular one struggles and is not capable of locating all cars going through the intersection.

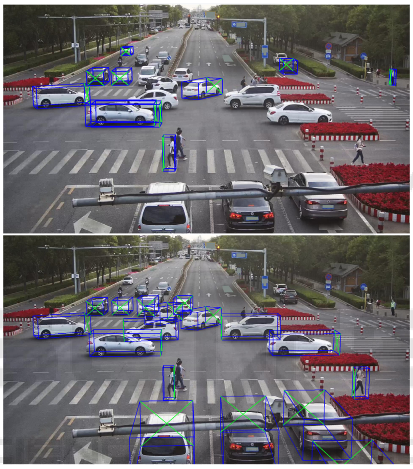

Figure 4: Inference results of the MonoLSS model (top) and MVX-Net (bottom), illustrating the differences in detection capabilities under high-density traffic conditions.

Similarly, in low-light conditions such as night operations, the monocular approach is challenged due to limited visibility and low image quality. In contrast, the multimodal system shows its strength by using LiDAR data to maintain detection accuracy despite the lack of light. Figure 5 shows experimental results in this scenario, where low visibility significantly affects the monocular system's performance.

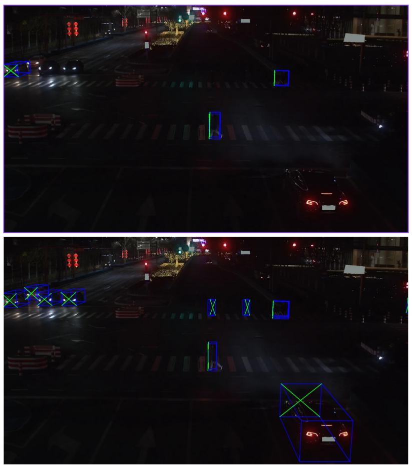

Figure 5: Inference results of the MonoLSS model (top) and MVX-Net (bottom), illustrating the differences in detection capabilities under low light traffic conditions, where image data alone is not enough for accurate detection.

## 5 CONCLUSIONS AND FUTURE WORKS

In this work, a modular detection system has been designed for smart cities, capable of operating in intelligent infrastructures equipped with either monocular or multimodal sensors. This modular configuration addresses the need for flexibility in traffic monitoring by offering a monocular solution, based on the MonoLSS algorithm, for simpler environments, while enhancing detection accuracy and robustness with the integration of LiDAR data through the multimodal MVX-Net model. This dual-mode design allows urban environments to adapt their traffic monitoring systems to different requirements and resource availability, balancing performance with cost efficiency.

The experiments conducted demonstrate that both algorithms are able to provide accurate 3D vehicle detection, ensuring effective traffic monitoring in diverse urban environments. The monocular module performs correctly and accurately in most scenarios while the multimodal configuration shows enhanced detection capabilities in more challenging situations, such as high density traffic or low light environments. Overall, the results validate the effectiveness of the modular approach in addressing a range of urban traffic monitoring needs, demonstrating that both the monocular and multimodal configurations work effectively under their respective conditions.

To advance this work, it is key to carry out real-world experiments on vehicle detection to validate the system's efficiency under dynamic and uncontrolled situations. Based on the results of these experiments, it will be important to improve the system's accuracy and robustness, particularly in classes and situations where it struggles. Additionally, future studies should incorporate datasets representing diverse real-world scenarios, including adverse weather, dense urban areas and rural regions, to enhance the generalizability of findings and validate the system under varied environmental and traffic conditions. Such tools would make it easier to identify bottlenecks, optimize traffic light timing, and enhance overall urban traffic flow, contributing to the development of smart cities. Furthermore, future research could also explore the application of the modular detection system beyond traffic monitoring, such as pedestrian flow analysis, public safety enhancement or intelligent parking detection.

In conclusion, the modular detection system developed in this work advances the development of smart cities by providing a scalable, flexible and adaptable solution for traffic monitoring. This approach not only meets current urban mobility demands but also sets a strong foundation for future innovations in the intelligent infrastructures field, ensuring that cities can address the emerging challenges in traffic safety, efficiency and sustainability.

## ACKNOWLEDGEMENTS

This research was funded by MCIN/AEI/10.13039/501100011033 grant numbers PID2021-124335OB-C21, PID2022-140554OB-C32 and PDC2022-133684-C31.

## REFERENCES

Arnold, E., Al-Jarrah, O. Y., Dianati, M., Fallah, S., Oxtoby, D., and Mouzakitis, A. (2019). A survey on 3d object detection methods for autonomous driving applications. *IEEE Transactions on Intelligent Transportation Systems*, 20(10):3782–3795.

Borau Bernad, J. (2024). RoadVision3D. https://github.com/jborau/RoadVision3D.

Borau Bernad, J., Ramajo-Ballester, Á., and Armingol Moreno, J. M. (2024). Three-dimensional vehicle detection and pose estimation in monocular images for smart infrastructures. *Mathematics*, 12(13):2027.

Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361.

Jain, N. K., Saini, R., and Mittal, P. (2019). A review on traffic monitoring system techniques. *Soft computing: Theories and applications: Proceedings of SoCTA 2017*, pages 569–577.

Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., and Beijbom, O. (2019). Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705.

Li, Z., Jia, J., and Shi, Y. (2024). Monolss: Learnable sample selection for monocular 3d detection. In *2024 International Conference on 3D Vision (3DV)*, pages 1125–1135.

Liu, X., Xue, N., and Wu, T. (2022). Learning auxiliary monocular contexts helps monocular 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1810–1818.

Liu, Z., Wu, Z., and T'oth, R. (2020). Smoke: Single-stage monocular 3d object detection via keypoint estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4289–4298.

MMDetection3D Contributors (2020). MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d.

Owais, M. (2024). Deep learning for integrated origin–destination estimation and traffic sensor location problems. *IEEE Transactions on Intelligent Transportation Systems*, 25(7):6501–6513.

Owais, M. and Shahin, A. I. (2022). Exact and heuristics algorithms for screen line problem in large size networks: Shortest path-based column generation approach. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):24829–24840.

Qian, R., Lai, X., and Li, X. (2022). 3d object detection for autonomous driving: A survey. *Pattern Recognition*, 130:108796.

Sindagi, V. A., Zhou, Y., and Tuzel, O. (2019). Mvx-net: Multimodal voxelnet for 3d object detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7276–7282.

Wang, T., Xinge, Z., Pang, J., and Lin, D. (2022). Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pages 1475–1485. PMLR.

Wang, Y., Mao, Q., Zhu, H., Deng, J., Zhang, Y., Ji, J., Li, H., and Zhang, Y. (2023). Multi-modal 3d object detection in autonomous driving: a survey. *International Journal of Computer Vision*, 131(8):2122–2152.

Yan, Y., Mao, Y., and Li, B. (2018). Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337.

Yu, F., Wang, D., Shelhamer, E., and Darrell, T. (2018). Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412.

Yu, H., Luo, Y., Shu, M., Huo, Y., Yang, Z., Shi, Y., Guo, Z., Li, H., Hu, X., Yuan, J., and Nie, Z. (2022). Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21361–21370.

Zanella, A., Bui, N., Castellani, A., Vangelista, L., and Zorzi, M. (2014). Internet of things for smart cities. *IEEE Internet of Things journal*, 1(1):22–32.

Zhang, Y., Carballo, A., Yang, H., and Takeda, K. (2023). Perception and sensing for autonomous vehicles under adverse weather conditions: A survey. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196:146–177.

Zhou, X., Wang, D., and Krähenbühl, P. (2019). Objects as points. *arXiv preprint arXiv:1904.07850*.

Zhou, Y. and Tuzel, O. (2018). Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.