CEA Comité Español de Automática
XLVI JORNADAS AUTOMÁTICA CARTAGENA 2025
UNIVERSIDADE DA CORUÑA

Jornadas de Automática

Sistema Escalable de Detección de Vehículos para Infraestructuras Inteligentes

Borau Bernad, Javier[a,*], Armingol Moreno, José María[a], Sanchis de Miguel, Araceli[b]

[a]*Laboratorio de Sistemas Inteligentes, Universidad Carlos III de Madrid, Avda de la Universidad, 30, 28911, Leganés, Madrid, España.*
[b]*Laboratorio de Control, Aprendizaje y Optimización de Sistemas, Universidad Carlos III de Madrid, Avda de la Universidad, 30, 28911, Leganés, Madrid, España.*

To cite this article: Borau-Bernad, J, Armingol Moreno, J.M., Sanchis de Miguel, A. 2025. Scalable Vehicle Detection System for Intelligent Infrastructures. Jornadas de Automática, 46.
https://doi.org/10.17979/ja-cea.2025.46.12076

Resumen

El aumento de la población en las ciudades ha creado la necesidad de gestionar de una forma eficiente los sistemas de transporte para mejorar la seguridad vial, optimizar el tráfico y reducir su impacto medioambiental. Las infraestructuras inteligentes equipadas con tecnologías de detección emergen como la principal solución para la monitorización del tráfico, pero enfrentan desafíos relacionados con costes, precisión y complejidad de instalación. En este contexto, este artículo presenta un sistema escalable de detección tridimensional de vehículos que integra tres modalidades de detección: monocular, LiDAR y multimodal (LiDAR y cámara RGB). El sistema propuesto selecciona automáticamente el modo de operación en función de los sensores instalados en cada infraestructura. Los resultados obtenidos muestran que esta solución modular permite optimizar el equilibrio entre coste y precisión para facilitar la implementación progresiva según las necesidades específicas de cada entorno urbano.

Palabras clave: Sistemas Inteligentes de Transporte, Machine Learning, Integración de sensores y percepción, Percepción y detección.

Scalable Vehicle Detection System for Intelligent Infrastructures

Abstract

The growing urban population has increased the need for efficient transportation systems that enhance road safety, optimize traffic flow and reduce environmental impact. Intelligent infrastructures equipped with sensing technologies have emerged as a key solution for traffic monitoring; however, they still face challenges related to cost, accuracy and installation complexity. This article presents a scalable 3D vehicle detection system that supports three detection modes: monocular, LiDAR and multimodal (LiDAR combined with an RGB camera). The proposed system automatically selects the most suitable mode based on the sensors available in each infrastructure. Experimental results show that this modular approach effectively balances cost and performance, enabling flexible and progressive deployment according to the specific needs of each urban environment.

Keywords: Intelligent transportation systems, Machine Learning, Sensor integration and perception, Perception and sensing.

1. Introducción

El aumento de la población mundial y su concentración en núcleos urbanos han impulsado la necesidad de gestionar de forma eficiente los sistemas de transporte en las ciudades. En este contexto, las infraestructuras inteligentes emergen como una de las principales herramientas para mejorar la seguridad vial, optimizar el tráfico y reducir su impacto medioambiental (Zanella et al., 2014). Estas infraestructuras integran tecnologías de detección y análisis de datos que permiten monitorizar el entorno urbano de forma precisa gracias a su elevado punto de vista (Borau Bernad et al., 2024), facilitando la toma de decisiones en tiempo real y el desarrollo de soluciones de movilidad más avanzadas.

*Autor para correspondencia: jborau@pa.uc3m.es

Tabla 1: Comparativa de tecnologías de detección para infraestructuras inteligentes según costes, prestaciones y requisitos de instalación.

Tecnología	Coste	Precisión 3D	Rendimiento con poca luz	Rendimiento en clima adverso	Instalación	Información visual
Monocular	Bajo	Media	Bajo	Bajo	Muy fácil	Sí
LiDAR	Alto	Alta	Alto	Medio	Media	No
Multimodal	Muy alto	Muy alta	Alto	Alto	Compleja	Sí

Sin embargo, las infraestructuras inteligentes actualmente enfrentan importantes desafíos de escalabilidad, coste y adaptabilidad. Los sistemas basados en cámaras monoculares ofrecen una solución económica y sencilla de implementar a costa de su limitado rendimiento en escenarios complejos o situaciones de visibilidad adversas (Dong et al., 2025). Asimismo, los sistemas que integran sensores LiDAR proporcionan una mayor precisión y robustez, pero también implican un incremento significativo en los costes y complejidad de instalación y mantenimiento (Wang et al., 2023).

Las diferentes necesidades y recursos disponibles en las infraestructuras urbanas plantean la necesidad de diseñar soluciones de detección de vehículos que sean flexibles y adaptables. Estos sistemas deben ser capaces de optimizar su funcionamiento dependiendo de los sensores instalados y las necesidades de cada entorno urbano.

En este artículo, se presenta un sistema escalable de detección de vehículos que alterna entre los modos monocular, LiDAR y multimodal (LiDAR y cámara RGB), seleccionando automáticamente entre ellos en función de los sensores disponibles en la infraestructura. Por un lado, se ha seleccionado el modelo MonoLSS (Li et al., 2024) para el modo monocular, con el objetivo de ofrecer una solución económica y capaz de operar de manera eficiente en entornos menos exigentes. Por otro lado, se utiliza el modelo SECOND (Yan et al., 2018) en el modo LiDAR para aquellas implementaciones que requieran un mayor nivel de precisión. Finalmente, el modo multimodal, basado en el modelo MVX-Net (Sindagi et al., 2019), se beneficia de la eficacia de integrar sensores LiDAR y cámaras RGB, mejorando la robustez del sistema ante oclusiones, condiciones adversas de iluminación o escenas con alta complejidad visual. Este sistema flexible permitirá a las infraestructuras inteligentes integrar tecnologías avanzadas de detección 3D de vehículos de forma progresiva, adaptándose a las necesidades específicas de cada entorno urbano y optimizando el equilibrio entre coste, rendimiento y escalabilidad.

2. Comparativa de Tecnologías de Detección

Actualmente, las infraestructuras inteligentes pueden equiparse con diferentes sensores para la detección de vehículos, cada uno con sus propias características, costes y limitaciones. Las principales tecnologías utilizadas en el ámbito de la detección 3D de objetos son las cámaras monoculares, los sensores LiDAR y la combinación multimodal de cámara y LiDAR. Esta sección presenta un análisis comparativo de estas tecnologías, teniendo en cuenta los factores clave que condicionan su adopción según las necesidades del entorno urbano y los recursos disponibles.

La detección basada en el uso de cámaras monoculares destaca por su bajo coste de implementación, facilidad de instalación y compatibilidad con infraestructuras ya existentes. Este tipo de sistema requiere únicamente una cámara calibrada y una unidad de procesamiento de datos, lo que permite su integración a gran escala sin necesidad de modificar el entorno urbano. Además, gracias a los avances en Deep Learning en los últimos años, los algoritmos de visión monocular han mejorado significativamente su capacidad para estimar las propiedades tridimensionales de los objetos a partir de imágenes 2D (Liu et al., 2020), (Liu et al., 2021).

Sin embargo, su principal desventaja radica en la dificultad para estimar la profundidad con precisión, ya que la información tridimensional debe inferirse a partir de proyecciones en 2D. Este hecho limita su rendimiento en situaciones con oclusiones, baja visibilidad o escenarios complejos, así como en condiciones de iluminación adversas o durante la noche, donde la calidad de la imagen empeora drásticamente.

Los sistemas basados en LiDAR superan estas limitaciones realizando la detección a partir de nubes de puntos tridimensionales generadas por láser, alcanzando mayor precisión en la estimación de profundidad y posición de los objetos (Zimmer et al., 2022). Esta capacidad los hace especialmente eficaces en situaciones con baja visibilidad, iluminación insuficiente o elevada densidad de tráfico. A pesar de sus beneficios, el elevado coste del sensor dificulta su despliegue a gran escala debido a la importante inversión necesaria (Arnold et al., 2019). Además, al no proporcionar información visual, los sistemas LiDAR no permiten capturar información gráfica como matrículas o imágenes de vehículos para futuras aplicaciones de las Smart Cities.

La combinación de sensores LiDAR y cámara en sistemas de detección multimodal permite aprovechar las fortalezas de ambas tecnologías y compensar sus limitaciones individuales. Por un lado, el LiDAR aporta información tridimensional precisa y robusta en condiciones de visibilidad adversas. Por otro, la cámara proporciona información visual útil para tareas como la clasificación de vehículos o el reconocimiento de matrículas. La fusión de estas dos tecnologías permite mejorar el rendimiento y la fiabilidad del sistema de detección en escenarios complejos, con objetos pequeños, ocluidos o en movimiento (Huang et al., 2024). No obstante, esta solución conlleva, además del elevado coste asociado al sensor LiDAR, un aumento en la complejidad de instalación y mantenimiento, ya que requiere la calibración conjunta de ambos sensores y un procesamiento de datos más intensivo (Owais, 2024).

Cada uno de los sistemas analizados presenta ventajas importantes, pero también limitaciones que condicionan su instalación en las ciudades. En la Tabla 1 se resume de forma comparativa el comportamiento de cada tecnología en relación con aspectos clave como el coste, la precisión o su rendimiento en condiciones atmosféricas adversas (Zhang et al., 2023). Esta diversidad de características expone la necesidad de contar

con soluciones de detección que sean flexibles y escalables, capaces de adaptarse a los distintos sensores instalados. Para ello, este artículo propone un sistema que combina los distintos métodos de detección para ajustarse a las necesidades y capacidades reales de cada infraestructura.

3. Sistema Propuesto

El sistema propuesto está diseñado para proporcionar una solución flexible y capaz de adaptarse a las necesidades de cada infraestructura inteligente para la detección tridimensional de vehículos. Su arquitectura permite operar en tres modos distintos: monocular, LiDAR y multimodal. El sistema selecciona automáticamente el modo de operación en función de los sensores disponibles sin necesidad de intervención manual. Esta característica facilita su implementación y mantenimiento, permitiendo desplegarlo en entornos urbanos con diferentes niveles de equipamiento o incluso aplicar mejoras posteriores mediante la instalación de sensores adicionales. Del mismo modo, su diseño modular lo hace compatible con infraestructuras preexistentes que ya cuenten con cámaras instaladas, lo que facilita la adopción progresiva de sistemas de detección de tráfico inteligentes.

3.1. Arquitectura General

La arquitectura del sistema de detección se organiza en tres módulos independientes: monocular, LiDAR y multimodal, cada uno con su propio modelo de detección optimizado para su tipo de datos de entrada. Para coordinar estos tres módulos se incluye un componente de gestión de dispositivos que selecciona automáticamente el módulo en función de los dispositivos instalados. Además, la arquitectura modular es fácilmente actualizable con nuevos sensores o modelos de detección para mejoras del sistema. Gracias a esta estructura modular, el sistema puede mantenerse operativo y eficiente a medida que evolucionan las capacidades tecnológicas o las necesidades de las ciudades.

3.2. Modo Monocular

Para la detección mediante el modo monocular, se ha seleccionado el modelo MonoLSS por su eficiencia computacional y su buen comportamiento incluso en entornos con oclusiones o visibilidad limitada. Este modelo utiliza un backbone DLA-34 (Yu et al., 2018) que genera mapas de características a partir de imágenes RGB sobre los que se infieren las propiedades tridimensionales de los objetos. El algoritmo integra un módulo *Learnable Sample Selection* (LSS) diseñado para identificar las regiones más relevantes de la imagen y mejorar la precisión en la detección de vehículos parcialmente ocluidos. El modo monocular es recomendable en implementaciones con bajo presupuesto o cuando no es posible incorporar sensores adicionales. También, es útil en zonas con baja complejidad de tráfico, donde no se requiere un sistema de alta precisión, como áreas rurales o residenciales.

3.3. Modo LiDAR

En infraestructuras donde únicamente se dispone de LiDAR, se utiliza el modo basado en el modelo SECOND: *Sparsely Embedded Convolutional Detection*. Este modelo utiliza una arquitectura de convoluciones dispersas para reducir el coste computacional de la detección y aumentar la velocidad de procesamiento sin disminuir la precisión del algoritmo. Inspirado por el algoritmo VoxelNet (Zhou and Tuzel, 2017), SECOND voxeliza las nubes de puntos del LiDAR y predice las posiciones, dimensiones y orientación de los objetos con mayor fiabilidad que mediante la detección monocular. Este modo es especialmente adecuado en situaciones donde se requiere una representación tridimensional detallada, como intersecciones complejas, zonas con alta peligrosidad o escenarios con baja iluminación.

3.4. Modo Multimodal

Cuando la infraestructura dispone simultáneamente de dispositivos LiDAR y cámaras RGB, el sistema activa el modo multimodal, que utiliza el modelo MVX-Net. Este modelo integra la información de ambos sensores usando la técnica *PointFusion* que concatena los puntos 3D con su pixel de la imagen para su procesamiento mediante *Voxel Feature Encoding* (VFE). Esta estrategia permite incorporar información visual directamente en la representación tridimensional del entorno para enriquecer la descripción espacial con contenido semántico, disminuyendo la probabilidad de falsos positivos o confusiones de clases.

El modo multimodal es principalmente útil en entornos donde se requiere una alta precisión y fiabilidad. Por ejemplo, en situaciones de tráfico denso, objetos parcialmente ocluidos o intersecciones donde la seguridad sea crítica. Además, la instalación de sistemas multimodales permite la implementación de aplicaciones que requieren información visual, como el reconocimiento de matrículas, la clasificación semántica de vehículos o vigilancia de seguridad, tareas que no pueden realizarse únicamente con sensores LiDAR.

4. Experimentos y Resultados

4.1. Conjunto de Datos

Para el entrenamiento y evaluación del sistema propuesto se ha utilizado el conjunto de datos DAIR-V2X (Yu et al., 2022), el cual ofrece escenarios reales de tráfico que incluyen imágenes RGB, nubes de puntos LiDAR, así como sus anotaciones y datos de calibración correspondientes. En concreto, se ha empleado el subconjunto DAIR-V2I, que contiene los datos capturados desde la infraestructura, como pórticos de monitorización o postes elevados. Este conjunto incluye más de 10.000 imágenes con sus correspondientes nubes de puntos y anotaciones tridimensionales. El conjunto de datos ha sido dividido en un 80 % para entrenamiento, 10 % para validación y el 10 % restante para el test final.

4.2. Detalles de Implementación

Los tres modelos pertenecientes al sistema fueron entrenados en un mismo servidor equipado con una GPU Nvidia 3090Ti, sin embargo, cada modelo siguió unos detalles de implementación diferentes. La evaluación del rendimiento de los algoritmos se realizó utilizando la métrica Average Precision con 40 puntos de recall (AP40) y un umbral de IoU de 0.7, comúnmente utilizada en los benchmarks actuales de detección 3D (Geiger et al., 2012).

En primer lugar, el modelo monocular MonoLSS fue implementado en la librería RoadVision3D (Borau Bernad,

2024), que utiliza Pytorch, y entrenado durante 150 épocas, evaluando cada 10. Las imágenes son cargadas con una resolución de 1024x576 y un batch size de 14. Con el fin de mejorar la capacidad de generalización del modelo se aplicaron las técnicas de rotación y recorte de imágenes aleatorias para el aumento de datos durante el entrenamiento. El optimizador utilizado fue Adam, con una tasa de aprendizaje inicial de 0.001 que disminuye en las épocas 90 y 120 multiplicándose por un factor de 0.1.

En segundo lugar, el modelo SECOND fue entrenado utilizando la librería MMDetection3D (MMDetection3D Contributors, 2020), basada también en PyTorch, durante 40 épocas, con validación cada 8 épocas. Se aplicaron técnicas de aumento sobre los puntos LiDAR, como *RandomFlip3D* y escalado. La voxelización se configuró con un tamaño de vóxel de [0.05, 0.05, 0.1] y un rango espacial de [0, -40, -3, 70.4, 40, 1]. El optimizador utilizado fue AdamW, con una tasa de aprendizaje inicial de 0.0018, modificándose durante el entrenamiento con un programador de tipo coseno.

Por último, el modelo MVX-Net fue entrenado utilizando la librería MMDetection3D durante 40 épocas y evaluando cada 5 épocas. La voxelización se realizó con un tamaño de [0.05, 0.05, 0.1] y un rango espacial de [0, -40, -3, 70.4, 40, 1], de forma similar al algoritmo anterior. Las imágenes se cargaron con una resolución de 960x540 y además se aplicaron técnicas de aumento de datos tanto sobre la imagen como sobre el LiDAR, incluyendo escalado, rotación y *RandomFlip3D*. El modelo fue optimizado usando AdamW con una tasa de aprendizaje inicial de 0.003, ajustada durante el entrenamiento mediante *cosine annealing*.

4.3. Resultados

Una vez finalizados los entrenamientos de los modelos explicados anteriormente, se ha realizado la etapa de test utilizando datos del conjunto DAIR-V2X. La Tabla 2 muestra los resultados obtenidos por cada algoritmo implementado en el estudio: MonoLSS, SECOND y MVX-Net. Estos resultados permiten comparar la precisión de cada modelo en términos matemáticos utilizando la métrica AP40 con un umbral de *Intersection over Union* (IoU) de 0.7 para los objetos de tipo coche.

Tabla 2: Resultados obtenidos en la etapa de test utilizando la métrica AP40 con umbral de IoU = 0.7 para los objetos de tipo coche.

	AP_{40} IoU 0.7		
Modelo	**Fácil**	**Moderado**	**Difícil**
MonoLSS	59.34	49.61	49.58
SECOND	69.07	56.68	56.71
MVX-Net	69.25	56.86	56.89

Como se puede observar en la tabla comparativa, el modelo MonoLSS demuestra unos resultados competitivos pese al uso de cámaras monoculares y su bajo coste. Aunque su rendimiento es más limitado que el de los algoritmos que utilizan LiDAR, el modelo mantiene una precisión más que aceptable incluso en la modalidad de detección difícil, donde se incluyen vehículos parcialmente ocluidos o más lejanos. Por lo tanto, el modo monocular se presenta como una solución práctica y suficientemente precisa para instalaciones que cuenten con un tráfico más sencillo o dispongan de un presupuesto menor.

Por otro lado, los modelos SECOND y MVX-Net alcanzan unos resultados superiores gracias a la mayor información espacial percibida del entorno mediante la integración del sensor LiDAR. Ambos modelos obtienen una puntuación más alta en las tres condiciones evaluadas, demostrando la ventaja que supone contar con datos tridimensionales reales provenientes de mediciones láser. A pesar de que MVX-Net incorpora información visual, su rendimiento cuantitativo es prácticamente idéntico al de SECOND en los experimentos realizados. No obstante, la integración de cámaras beneficia al sistema al aportar información semántica adicional y permitir la implementación de aplicaciones posteriores como el reconocimiento de matrículas o la clasificación de vehículos.

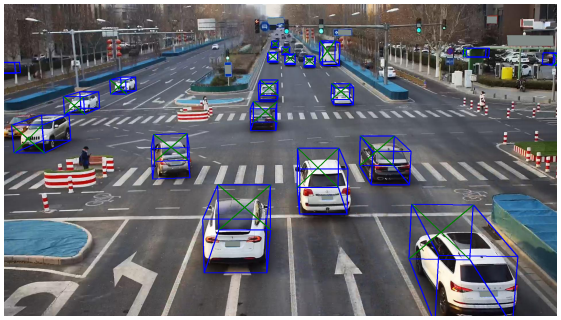

Figura 1: Resultado de inferencia del sistema en modo monocular utilizando el modelo MonoLSS. Se observa la proyección tridimensional de los vehículos detectados con precisión, demostrando su capacidad para identificar objetos lejanos, a pesar de la ausencia de datos de profundidad.

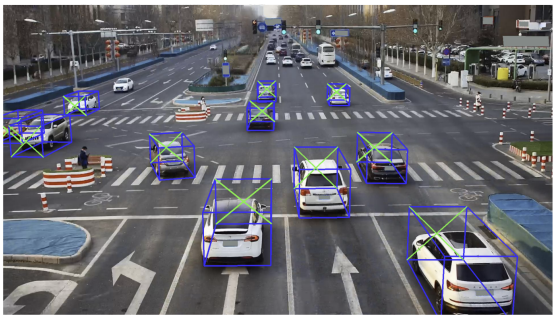

Figura 2: Resultado de inferencia del sistema en modo LiDAR utilizando el modelo SECOND. La figura muestra las detecciones proyectadas sobre la imagen RGB para facilitar la interpretación de los resultados. Se identifican con precisión todos los vehículos presentes en la escena dentro del rango del sensor, evidenciando la fiabilidad del sistema.

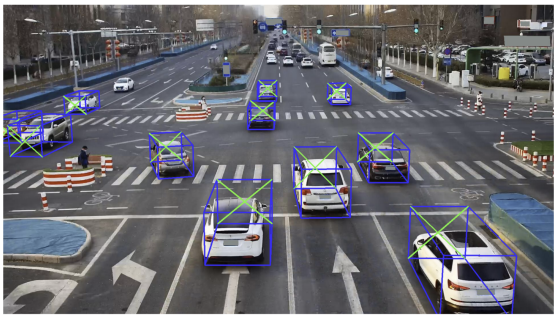

Figura 3: Resultado de inferencia del sistema en modo multimodal utilizando el modelo MVX-Net. La figura muestra las detecciones proyectadas sobre la imagen RGB donde se localizan todos los vehículos en la intersección de forma precisa demostrando la robustez del sistema multimodal.

Figura 4: Inferencias realizadas con MonoLSS (izquierda), SECOND (centro) y MVX-Net (derecha) proyectando los resultados sobre la imagen. La figura muestra cómo en situaciones de tráfico denso con oclusiones entre vehículos el sistema monocular no es capaz de detectar algunos objetos, mientras que el LiDAR y el multimodal son capaces de detectarlos incluso en condiciones de visibilidad parcial. Del mismo modo, el modo monocular no es capaz de detectar los vehículos en situaciones de baja luminosidad, mientras que sí lo son los sistemas que integran LiDAR. Sin embargo, el modo multimodal se beneficia de la integración de imágenes detectando el vehículo más lejano.

Las Figuras 1, 2 y 3 muestran ejemplos visuales de inferencia obtenidos con los modelos MonoLSS, SECOND y MVX-Net, respectivamente. En estas figuras, se demuestra la capacidad de detección del sistema en sus tres modos: monocular, LiDAR y multimodal, en escenas de tráfico poco concurrido, donde la buena visibilidad y la baja presencia de oclusiones permiten al sistema proporcionar detecciones precisas de cada vehículo. Sin embargo, en situaciones de tráfico complejo y baja luminosidad, como se observa en la Figura 4, el modelo monocular pierde eficacia, dejando sin detectar algunos vehículos parcialmente ocultos o los no iluminados, mientras que los modos LiDAR y multimodal mantienen un rendimiento sólido, identificando correctamente la mayoría de los objetos presentes en la escena. Esto refuerza la ventaja que supone integrar sensores LiDAR en infraestructuras donde se requiera una mayor robustez del sistema, especialmente en entornos con alta densidad de tráfico, visibilidad reducida o despliegues que requieran un funcionamiento óptimo durante la noche.

5. Conclusiones y trabajos futuros

En este artículo se ha presentado un sistema escalable para la detección tridimensional de vehículos orientado a facilitar el despliegue de este tipo de sistemas en las infraestructuras urbanas. El sistema integra tres modos de detección: monocular, LiDAR y multimodal, seleccionando automáticamente en base a los sensores instalados. Esta flexibilidad y facilidad de instalación permite adaptar el sistema a distintos niveles de equipamiento, optimizando el equilibrio entre coste, precisión y escalabilidad en función de las necesidades de cada implementación.

El modo monocular, basado en el modelo MonoLSS, proporciona una solución eficiente y de bajo coste para entornos menos exigentes o con presupuesto limitado. Este modelo ha demostrado un rendimiento competitivo en las métricas AP40, alcanzando valores aceptables incluso en escenarios difíciles con objetos lejanos. Sin embargo, el modelo se ve afectado por situaciones de baja luminosidad o alta densidad de tráfico donde las oclusiones reducen significativamente la capacidad del algoritmo para detectar vehículos con fiabilidad. En cambio, el modo de funcionamiento LiDAR, basado en el modelo SECOND, se presenta como una solución más fiable para instalaciones donde se requieran capacidades avanzadas de detección ante condiciones desfavorables, como en zonas de tráfico denso, intersecciones peligrosas o espacios con visibilidad limitada. Los resultados muestran una mejora significativa respecto al método monocular, especialmente en escenarios moderados y difíciles, destacando su idoneidad en entornos donde la seguridad es crítica. Además, en los experimentos de inferencia realizados, el modelo obtiene un rendimiento superior en las situaciones de tráfico denso o luminosidad baja, demostrando los beneficios que supone de integrar sistemas LiDAR en infraestructuras inteligentes.

Por último, el modo multimodal, que integra el modelo MVX-Net, combina las ventajas del LiDAR con la información visual de las cámaras RGB, aportando robustez adicional frente a oclusiones y permitiendo tareas complementarias como la clasificación semántica de vehículos o el reconocimiento de matrículas. Si bien sus métricas de precisión y su comportamiento en las inferencias realizadas son similares al modo LiDAR, su capacidad para ofrecer información contextual lo convierte en la opción más avanzada para entornos urbanos complejos o para integrar futuras aplicaciones de las smart cities.

Este sistema modular y flexible establece las bases para una implementación progresiva de soluciones inteligentes de monitorización del tráfico, capaces de adaptarse a las necesidades y presupuesto de cada implementación para facilitar el crecimiento y evolución tecnológica de las infraestructuras urbanas. Una vez diseñado este sistema, en futuros trabajos se plantea realizar pruebas de despliegue real en ciudades, evaluando su rendimiento y fiabilidad bajo condiciones dinámicas.

Agradecimientos

Este trabajo ha sido realizado parcialmente gracias al apoyo de las subvenciones PID2021-124335OB-C21 y PID2022-140554OB-C32 financiadas por MCIN/AEI/10.13039/501100011033, el proyecto TEC-2024/ECO-277 financiado por la Comunidad de Madrid y el Programa de Investigación Predoctoral en Formación en Inteligencia Artificial (PIPF-IA) de la Universidad Carlos III de Madrid.

Referencias

Arnold, E., Al-Jarrah, O. Y., Dianati, M., Fallah, S., Oxtoby, D., 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., Armingol Moreno, J. M., 2024. Three-dimensional vehicle detection and pose estimation in monocular images for smart infrastructures. Mathematics 12 (13), 2027.

Dong, Q., Zhou, Z., Qiu, X., Zhang, L., 2025. A survey on self-supervised monocular depth estimation based on deep neural networks. IEEE Transactions on Neural Networks and Learning SystemsAdvance online publication.
URL: https://doi.org/10.1109/TNNLS.2025.3552598
DOI: 10.1109/TNNLS.2025.3552598

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. 2012 IEEE Conference on Computer Vision and Pattern Recognition, 3354–3361.
URL: https://api.semanticscholar.org/CorpusID:6724907

Huang, K., Shi, B., Li, X., Li, X., Huang, S., Li, Y., 2024. Multi-modal sensor fusion for auto driving perception: A survey.
URL: https://arxiv.org/abs/2202.02703

Li, Z., Jia, J., Shi, Y., 2024. Monolss: Learnable sample selection for monocular 3d detection. In: 2024 International Conference on 3D Vision (3DV). pp. 1125–1135.
DOI: 10.1109/3DV62453.2024.00088

Liu, X., Xue, N., Wu, T., 2021. Learning auxiliary monocular contexts helps monocular 3d object detection.
URL: https://arxiv.org/abs/2112.04628

Liu, Z., Wu, Z., Tóth, R., 2020. Smoke: Single-stage monocular 3d object detection via keypoint estimation.
URL: https://arxiv.org/abs/2002.10111

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.
DOI: 10.1109/TITS.2023.3344533

Sindagi, V. A., Zhou, Y., Tuzel, O., 2019. Mvx-net: Multimodal voxelnet for 3d object detection. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 7276–7282.
DOI: 10.1109/ICRA.2019.8794195

Wang, Y., Mao, Q., Zhu, H., Deng, J., Zhang, Y., Ji, J., Li, H., 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., Li, B., 2018. Second: Sparsely embedded convolutional detection. Sensors 18 (10).
URL: https://www.mdpi.com/1424-8220/18/10/3337
DOI: 10.3390/s18103337

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

Yu, H., Luo, Y., Shu, M., Huo, Y., Yang, Z., Shi, Y., Guo, Z., Li, H., Hu, X., Yuan, J., Nie, Z., June 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). pp. 21361–21370.

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

Zhang, Y., Carballo, A., Yang, H., 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.
URL: https://www.sciencedirect.com/science/article/pii/S0924271622003367
DOI: 10.1016/j.isprsjprs.2022.12.021

Zhou, Y., Tuzel, O., 2017. Voxelnet: End-to-end learning for point cloud based 3d object detection.
URL: https://arxiv.org/abs/1711.06396

Zimmer, W., Ercelik, E., Zhou, X., Ortiz, X. J. D., Knoll, A., 2022. A survey of robust 3d object detection methods in point clouds.
URL: https://arxiv.org/abs/2204.00106