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

Jornadas de Automática

Integración de cámaras RGB y un LiDAR 3D para la detección de participantes del tráfico en CARLA y ROS2

Montenegro, Jorge[a,*], Morales, Jesús[a], Martínez, Jorge Luis[a]

[a]*Dpto. de Ingeniería de Sistemas y Automática, Universidad de Málaga, Arquitecto Francisco Peñalosa, nº 6, 29071, Málaga, España*

To cite this article: Montenegro, J., Morales, J., Martínez, J.L. 2025. Integration of RGB cameras and a 3D LiDAR for traffic participant detection in CARLA and ROS2. Jornadas de Automática, 46. https://doi.org/10.17979/ja-cea.2025.46.12062

Resumen

Este artículo mejora la detección de participantes del tráfico en vehículos autónomos mediante la fusión de datos obtenidos con cámaras RGB y un LiDAR 3D a bordo. Para ello, se usa el modelo BEVFusion, el cual integra las características de estos sensores en un espacio compartido por ambos, denominado vista de pájaro. Este modelo se entrena y evalúa sobre un conjunto de datos sintéticos generado en el simulador CARLA (*Car Learning to Act*) que incluye diferentes tipos de participantes del tráfico como coches, peatones, camiones, autobuses y motocicletas. De forma cuantitativa, se obtiene la precisión media y los errores medios de traslación, escala, orientación, velocidad. Además, se realiza un análisis de la influencia en la precisión media de la distancia entre los participantes del tráfico y el vehículo sensorizado. De forma cualitativa, se evalúa el modelo sobre datos en línea procedentes de CARLA y procesados mediante un contenedor de BEVFusion en el entorno de desarrollo de aplicaciones robóticas ROS2 (*Robot Operating System*).

Palabras clave: Vehículos autónomos, sensores de automoción, redes neuronales, fusión sensorial, percepción y medida.

Integration of RGB cameras and a 3D LiDAR for traffic participant detection in CARLA and ROS2

Abstract

This paper improves the detection of traffic participants for autonomous vehicles by fusing data obtained from onboard RGB cameras and a 3D LiDAR. To this end, the BEVFusion model is employed, which integrates the features of these sensors in a space shared by both, called bird's eye view. This model is trained and evaluated on a synthetic dataset generated in the CARLA (Car Learning to Act) simulator, which includes different types of traffic participants such as cars, pedestrians, trucks, buses and motorbikes. Metrics such as average accuracy and average errors of translation, scale, orientation, speed are analysed. In addition, the influence on the average accuracy of the distance between the traffic participants and the sensorized vehicle are analysed quantitatively. The model is evaluated qualitatively with online data from CARLA and processed using a BEVFusion wrapper in the environment for developing robotic applications ROS2 (Robot Operating System).

Keywords: Autonomous vehicles, automotive sensors, neural networks, sensor data fusion, perception and sensing.

1. Introducción

La percepción del entorno es uno de los pilares fundamentales para el desarrollo de sistemas inteligentes, como los coches autónomos. Para que estos vehículos puedan operar de manera segura en entornos urbanos y dinámicos, requieren una representación densa y fiel del espacio que los rodea (Yang et al., 2025). En los vehículos autónomos se utilizan sensores como cámaras RGB (Moreau and Ibanez-Guzman, 2023), RADAR (Srivastav and Mandal, 2023) o LiDAR 3D (Urmila. and Megalingam, 2020), principalmente.

En este contexto, ningún sensor puede, por sí solo, ofre-

*Autor para correspondencia: jorgemn@uma.es

cer una percepción robusta ante las distintas condiciones del entorno (Huang et al., 2022). Por un lado, las cámaras RGB, proporcionan información sobre la forma, textura y color de los objetos, aunque carecen de información espacial y pueden fallar en condiciones de poca iluminación. Por otro lado, los sensores LiDAR 3D ofrecen información sobre la profundidad de los objetos, pero carecen de información semántica y presentan limitaciones en tareas de detección de peatones mediante redes neuronales (Montenegro et al., 2024).

Ante estas limitaciones, se presenta la fusión multi-modal de sensores, que es una técnica basada en combinar datos de diferentes fuentes para obtener una visión más completa del entorno superando las limitaciones que cada uno tiene de forma individual.

Dentro de las distintas técnicas de fusión multi-modal destacan Transfusion (Bai et al., 2022) y BEVFusion (Liu et al., 2023), que permiten unificar características procedentes de cámaras RGB y un LiDAR 3D en el espacio BEV (*Bird's Eye View*). Esta vista de pájaro es una representación del entorno mostrada desde una perspectiva aérea que permite conservar tanto la estructura geométrica como la densidad semántica. Sin embargo, estas técnicas tienen la desventaja de no estar diseñadas de forma nativa para su uso con datos en línea.

En este artículo se presenta un entorno de fusión multi-modal basada en BEVFusion, capaz de combinar en línea detecciones de datos procedentes de seis cámaras RGB y un LiDAR 3D integrada en ROS2 (*Robot Operating System*). Para ello se ha generado un repositorio (*dataset*) sintético en CARLA (*Car Learning to Act*) (Dosovitskiy et al., 2017) para entrenar y evaluar cuantitativa y cualitativamente la arquitectura.

El resto del artículo está organizado de la siguiente manera. El apartado 2 trata sobre la arquitectura de fusión BEVFusion y su implementación en ROS2. Tras esto, en la sección 3 se describe la generación del *dataset* sintético `PremoveSim` en CARLA. Posteriormente, en el apartado 4 se muestran los resultados cuantitativos y cualitativos sobre el simulador CARLA. Por último, el artículo finaliza con las conclusiones en la sección 5.

2. Arquitectura para la fusión multi-modal

En vehículos autónomos ha destacado el uso de redes neuronales para identificar participantes del tráfico con una elevada precisión. Entre éstas destacan las redes YOLO (*You Only Look Once*) para imágenes RGB (Redmon et al., 2016) o las PV-RCNN (*PointVoxel Regional based Convolutional Neural Network*) para nubes de puntos 3D (Shi et al., 2020).

Sin embargo, el uso de estos sensores por separado, tiene una serie de inconvenientes, como la dificultad de detección de peatones usando el LiDAR 3D o la pérdida de información espacial usando cámaras RGB (Montenegro et al., 2024). Para solucionarlo se presenta la fusión multi-modal, una técnica que combina datos de diferentes sensores para superar las limitaciones de cada uno por separado. Concretamente, al combinar la información visual de las cámaras RGB con la profundidad del LiDAR 3D, se pueden obtener detecciones más robustas y con mayor precisión media.

En cuanto a las técnicas de fusión multi-modal existen tres métodos: proyección de cámaras a LiDAR 3D, de LiDAR 3D a cámaras y a vista de pájaro (Wang et al., 2023). Por un lado, *Point Painting* (Vora et al., 2020) destaca en la proyección de cámaras a LiDAR, pero conlleva una pérdida de información semántica. Por otro lado, *Point Fusion* (Xu et al., 2018) destaca en la proyección de LiDAR a cámaras, pero distorsiona geométricamente la información. Finalmente, en la proyección a vista de pájaro destaca el modelo Transfusion (Bai et al., 2022), que soluciona las limitaciones anteriores pero implica un gran tiempo de inferencia, lo cual soluciona el modelo BEVFusion (Liu et al., 2023), haciéndolo especialmente adecuado para su uso en tiempo real en conducción autónoma.

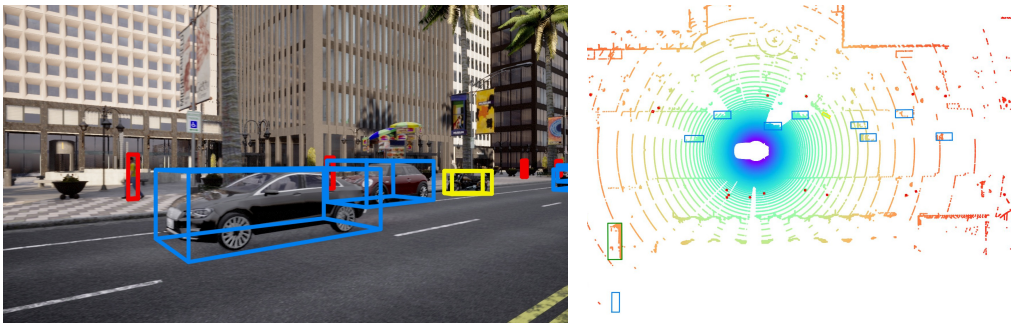
Introducido en 2022, el modelo BEVFusion, desarrollado por Mit-Han-Lab se basa en extraer las características de las cámaras y de las nubes de puntos por separado para, posteriormente, proyectarlas al espacio BEV (Liu et al., 2023). En el caso de las nubes de puntos tridimensionales (3D), únicamente es necesario proyectarlos sobre un plano perpendicular al eje z. Sin embargo, para el caso de las cámaras, se estima la profundidad de cada píxel empleando la técnica LSS (*Lift-Splat-Shoot*) (Philion and Fidler, 2020), con lo cual se genera una nube de puntos 3D, que nuevamente se proyecta sobre un plano perpendicular al eje z. Una vez proyectadas las características al espacio BEV, se fusionan ambas aplicando, en primer lugar, la concatenación y, posteriormente, un proceso de codificación basado en convoluciones de cara a compensar posibles desalineaciones entre las características del LiDAR y las cámaras. Finalmente, sobre estas características ya es posible realizar detección 3D de participantes del tráfico.

Con el objetivo de mejorar los tiempos de inferencia e implementar BEVFusion con datos en línea, es posible utilizar la solución que ofrece NVIDIA-AI-IOT (NVIDIA, 2023). Dicha solución hace uso de TensorRT, una herramienta que mejora los tiempos de inferencia de modelos de redes neuronales fusionando capas y reduciendo la precisión. Con ello, se consigue una velocidad de inferencia de 25 FPS (*Frames Per Second*) (NVIDIA, 2023), frente a los 8.4 FPS (Liu et al., 2023) que ofrece el modelo BEVFusion de Mit-Han-Lab.

Además, la solución de NVIDIA de BEVFusion permite integrarse en entornos como ROS2, empleando para ello el contenedor BEVFusion-ROS-TensorRT (linClubs, 2023), que recibe y sincroniza imágenes y nubes de puntos en línea, los procesa y realiza la inferencia con BEVFusion devolviendo, finalmente, las detecciones.

3. Generación de repositorios sintéticos en CARLA

En el contexto de los vehículos autónomos, los *datasets* están compuestos por datos de diferentes sensores, como cámaras RGB, LiDAR 3D, RADAR o GPS (Caesar et al., 2020). Sin embargo, la obtención de datos etiquetados del mundo real es un proceso lento y costoso. Es por ello que surgen los *datasets* sintéticos generados mediante simuladores como Gazebo o CARLA, este último enfocado en conducción autónoma sobre entornos urbanos realistas y dinámicos con diferentes condiciones de tráfico y clima. Además, estos simuladores permiten el etiquetado automático de los datos durante el proceso de simulación, lo que elimina la necesidad de la intervención humana en gran parte del proceso de etiquetado (Song et al., 2023).



(a) Imagen RGB (b) Vista aérea de la nube de puntos LiDAR 3D

Figura 1: Muestra de datos etiquetados obtenidos con SimBEV.

Una herramienta empleada junto con CARLA para crear repositorios sintéticos es SimBEV (Mehr and Eskandarian, 2025), la cual permite definir el tipo y cantidad de sensores a bordo, las condiciones meteorológicas, el mapa o la cantidad de participantes del tráfico. Una vez definido lo cual, el vehículo recorre el mapa siguiendo una trayectoria definida y recopilando la información de los sensores junto con las cajas de detección (*bounding boxes*) de los participantes del tráfico en el mismo formato de nuScenes (Caesar et al., 2020), que es un repositorio estándar en conducción autónoma.

Empleando dicha herramienta, se ha creado un *dataset* sintético, denominado `PremoveSim`, con muestras de peatones, coches, camiones, autobuses y motocicletas en condiciones diurnas despejadas. Además, cuenta con un total de veinte escenas, de las cuales un 60 % son para entrenamiento (*train*), un 20 % para validación (*val*) y un 20 % para pruebas (*test*) en un entorno diurno y despejado cuyas condiciones adicionales se muestran resumidas en la Tabla 1 y 2.

Distribución	**Mapa**	**Escenas**	**# Muestras**
train	Town10 Town15	12	3828
val	Town10 Town15	4	1276
test	Town10 Town15	4	1276
Total	-	20	6380

Tabla 1: Distribución de escenas en el repositorio `PremoveSim`.

Clases	***train***	***val***	***test***
Coches	18,96 %	26,38 %	21,69 %
Peatones	68,17 %	50,03 %	63,51 %
Camiones	8,70 %	13,04 %	7,76 %
Autobuses	2,09 %	2,24 %	1,45 %
Motocicletas	2,08 %	8,31 %	5,59 %
Total detecciones	139959	28545	31327

Tabla 2: Porcentaje de clases etiquetadas en las muestras.

Para obtener el *dataset*, se ha simulado un vehículo Toyota Prius, sobre el cual se ha replicado la configuración sensorial del vehículo empleado en el conjunto de datos nuScenes, compuesta por un LiDAR 3D sobre el techo del vehículo, en su centro, y seis cámaras RGB: tres en la parte frontal, orientadas hacia delante (centro, derecha e izquierda) y otras tres en la parte trasera, orientadas hacia atrás de forma análoga. Las características de estos sensores se recogen en la Tabla 3 para el LiDAR 3D y en la Tabla 4 para las cámaras RGB. En la Figura 1 se muestra como ejemplo una imagen RGB y una nube de puntos 3D etiquetadas y generadas mediante SimBEV con sus correspondientes *ground truth bounding boxes*.

Característica	**Valor**
Alcance máximo	≤ 100 m
Puntos por revolución	17500 puntos
Campo de visión vertical	[−30,7°, 10,7°]
Campo de visión horizontal	360°
Velocidad de rotación	10 Hz
Número de haces	32

Tabla 3: Especificaciones del sensor LiDAR 3D Velodyne HDL32E.

Característica	**Valor**
Frecuencia de captura	60 Hz
Resolución	1600x1200 px
Campo de visión	50º

Tabla 4: Especificaciones de las cámaras RGB Basler acA1600-60gc.

La Figura 2 muestra cómo SimBEV configura CARLA, recibe las muestras etiquetadas y genera el repositorio sintético. Las herramientas empleadas se muestran con un rectángulo, mientras que los datos con elipses.

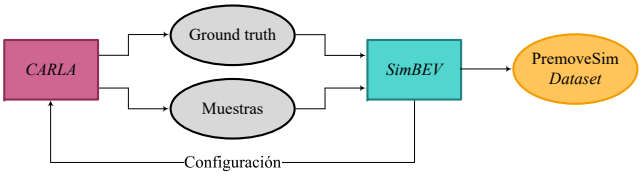


Figura 2: Esquema del entorno de generación de datos sintéticos.

4. Resultados sobre el simulador CARLA

El equipo informático empleado para entrenar el modelo BEVFusion y ejecutar el simulador junto a la red neuronal cuenta con un procesador Intel Core™ i7-13700F a 5,2 GHz con 16 núcleos, 24 hilos y 16 MB de caché L2, tarjeta gráfica GigaByte RTX 4070 TI GAMING OC con 12 GB de memoria y dos módulos de memoria RAM DDR5 de 16 GB a 5600 MHz. Cuantitativamente se han analizado las métricas del entrenamiento y se ha realizado un estudio de la influencia de la distancia en la precisión media de las detecciones, mientras que, el análisis cualitativo se ha realizado en línea mediante ROS2 con datos del simulador CARLA.

4.1. *Métricas del entrenamiento del modelo BEVFusion en CARLA*

Para entrenar el modelo BEVFusion usando el repositorio sintético `PremoveSim`, se ha empleado la configuración mostrada en la Tabla 5.

Parámetro	Valor
Red base para imágenes RGB	ResNet-50
Red base para nubes de puntos 3D	VoxelNet
Tasa de aprendizaje	1×10^{-5}
Número total de épocas	20

Tabla 5: Configuración para el entrenamiento del modelo BEVFusion en el *dataset* `PremoveSim`.

Los resultados del entrenamiento del modelo BEVFusion en `PremoveSim` se muestran en la Tabla 6 para diferentes clases de objetos, evaluados con la precisión media (AP), el error medio de traslación (ATE), el error medio de escala (ASE), el error de orientación (AOE) y el error de velocidad (AVE).

Clases	AP (0 – 1)	ATE (m)	ASE (m)	AOE (rad)	AVE (m/s)
Coches	0,505	0,197	0,129	0,292	0,440
Camiones	0,591	0,314	0,209	0,164	2,133
Autobuses	0,471	0,355	0,163	0,085	2,870
Peatones	0,491	0,121	0,075	0,215	0,496
Motocicletas	0,412	0,217	0,238	0,837	0,125
Media	0,494	0,241	0,163	0,319	1,213

Tabla 6: Métricas del modelo BEVFusion con el subconjunto de validación (`val`) del *dataset* `PremoveSim`.

En cuanto a las métricas obtenidas, se observa una precisión elevada para camiones, coches y peatones, con valores de AP de 0,591, 0,505 y 0,491, respectivamente, lo que indica que el modelo ha aprendido a identificar estos objetos con una alta precisión. Sin embargo, la precisión de las motocicletas y autobuses es algo inferior ya que, en cada escena hay, comparativamente, menos muestras.

En relación a la localización espacial de los participantes del tráfico, evaluada mediante el ATE, el modelo muestra una mayor precisión en la localización de coches y peatones frente al resto de clases, ya que los camiones o autobuses pueden tener una mayor variación de forma. En relación al ASE, el modelo BEVFusion muestra una mejor capacidad para estimar el tamaño de objetos pequeños como peatones y coches frente a objetos grandes como autobuses y camiones. Esto se debe a que los objetos más grandes, al ocupar más píxeles, tienen más posibilidades de error al estimar su profundidad.

El AOE muestra una peor capacidad en la estimación de la orientación de las detecciones sobre peatones y motocicletas, lo cual se debe a que estas clases tienen una mayor movilidad y, en consecuencia, más orientaciones distintas, por lo que harían falta más muestras de dichas clases para que la red aprendiese mejor. Finalmente, analizando el AVE se observa una peor estimación de la velocidad en vehículos grandes, como camiones o autobuses, mientras que mejora conforme disminuye el tamaño del vehículo.

4.2. *Influencia de la distancia en la precisión media*

Una vez obtenidas las métricas del modelo BEVFusion sobre `PremoveSim`, se ha realizado un análisis de la influencia de la distancia de los participantes del tráficos respecto al vehículo sensorizado en la precisión media, dentro del rango inmediato de colisiones del vehículo, de cero a cuatro metros. Los resultados del análisis se muestran la Tabla 7, para distancias de 0,5 m, 1,0 m, 2,0 m, 4,0 m respecto al centro del vehículo. La información de la Tabla 7 se pueden representar de forma gráfica como se muestra en la Figura 3.

Se puede observar cómo la precisión en las detecciones aumenta conforme mayor es la distancia al vehículo sensorizado, lo cual puede deberse a que conforme más cerca está un objeto, mayor es la pérdida de información espacial sobre el mismo. Además, en algunas clases, como las motocicletas, la variación del AP con la distancia es menor, lo cual se puede comprobar numéricamente en la Tabla 7. La razón para ello es que, en las muestras del conjunto de datos, estos objetos únicamente han aparecido a distancias menores de 1 m, con lo que también pertenecen al resto de casos, como distancias menores a 2 m y en adelante.

Clases	AP			
	0,5 m	**1,0 m**	**2,0 m**	**4,0 m**
Coches	0,335	0,502	0,589	0,595
Camiones	0,369	0,625	0,683	0,687
Autobuses	0,346	0,500	0,514	0,526
Peatones	0,348	0,495	0,557	0,565
Motocicletas	0,391	0,414	0,422	0,422

Tabla 7: Métricas de AP a diferentes distancias en el modelo BEVFusion para varias clases de objetos con `PremoveSim`.

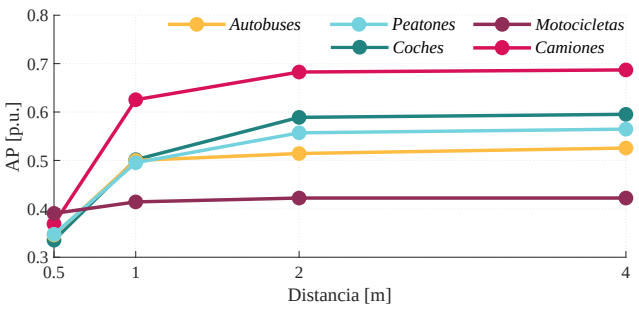


Figura 3: Evolución del AP frente a la distancia.

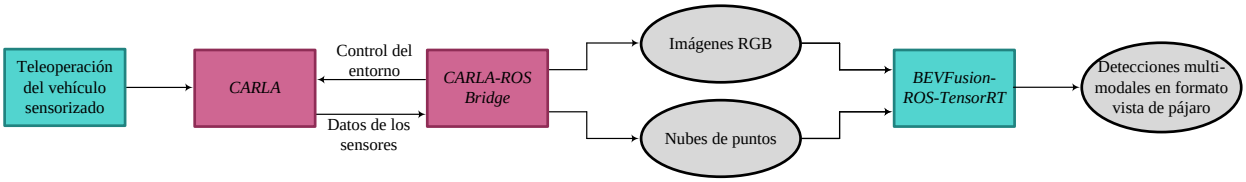


Figura 4: Esquema de la inferencia de BEVFusion en ROS2. Los nodos de ROS2 se representan con un rectángulo azul, mientras que los tópicos con elipses.

Figura 5: Detección de coches, camiones y peatones sobre imágenes RGB y una nube de puntos LiDAR 3D.

4.3. *Inferencia en línea de BEVFusion*

Una vez obtenidas las métricas del modelo BEVFusion en el conjunto de datos `PremoveSim`, se ha convertido el fichero de pesos `pytorch` a formato `onnx` y se ha aplicado post-cuantificación (*Post Training Quantization*) mediante TensorRT. Tras esto, se ha empleado el fichero de pesos con post-cuantificación en BEVFusion-ROS-TensorRT sobre datos en línea del simulador CARLA obtenidos mediante ROS2. En la Figura 4 se puede apreciar cómo el entorno ROS2 se comunica con CARLA, permitiendo controlar los parámetros del simulador y teleoperar el vehículo sensorizado. Además, CARLA publica las imágenes RGB y las nubes de puntos LiDAR 3D en tópicos de ROS2 que, posteriormente, emplea el contenedor BEVFusion-ROS-TensorRT para hacer la inferencia con la herramienta BEVFusion, devolviendo las detecciones multi-modales en formato de vista de pájaro.

En la Figura 5 se muestran algunos resultados de detección de participantes del tráfico sobre imágenes RGB y una nube de puntos LiDAR 3D, con su correspondiente etiqueta de identificación. El vehículo sensorizado se representa con una caja de color verde, los coches con una caja de color naranja, los peatones con una caja de color azul y los camiones con una caja de color rojo. Sobre estas cajas se muestra una etiqueta con la clase detectada y su precisión.

Analizando de forma cualitativa los resultados obtenidos, se pueden observar detecciones de coches con una precisión entre 0,4 y 0,6. Además aparecen detecciones de peatones con una precisión inferior a 0,3, lo cual puede deberse a que los peatones presentan una mayor variación en sus atributos, como la ropa, el movimiento o tamaño respecto a otras clases. En cuanto a los camiones, se puede observar en la Figura 5 la detección de uno en la cámara frontal central y en la cámara frontal izquierda con una precisión de 0,2, lo cual nuevamente indica que la red tiene margen para aprender sobre esta clase. Finalmente, se puede apreciar que no se han detectado aquellos participantes del tráfico más lejanos, como el autobús de la cámara frontal central o aquellos que se camuflan con el entorno, como el peatón de la cámara trasera izquierda que se encuentra entre el coche azul y el coche de policía. Para solucionarlo, sería necesario ampliar el repositorio con un mayor número de escenas en diferentes situaciones.

5. Conclusiones

En este trabajo se ha presentado una solución para aumentar la robustez de la detección de participantes del tráfico empleando la fusión multi-modal de datos procedentes de seis cámaras RGB y un LiDAR 3D en un entorno simulado con CARLA y ROS2. Para ello se ha empleado el modelo BEVFusion, el cual ha sido entrenado con un conjunto de datos sintético creado en CARLA.

De forma cuantitativa, las métricas muestran un buen desempeño de la red neuronal en términos de precisión media, la cuál es cercana a 0,5 en las diferentes clases. Además, el análisis de la influencia de la distancia en la precisión media ha mostrado que ésta aumenta conforme mayor es la distancia entre los participantes del tráfico y el vehículo sensorizado, dentro del rango de cero a cuatro metros. De forma cualitativa, con datos en línea del simulador CARLA obtenidos mediante ROS2, se ha observado que se detectan correctamente la gran mayoría de participantes del tráfico, aunque aquellos demasiado alejados no han sido detectados adecuadamente y algunas clases, como los peatones, han presentado una menor precisión.

Como línea futura de trabajo podría plantearse extender el repositorio de datos con un mayor número de escenas en diferentes situaciones, probar diferentes configuraciones de entrenamiento y añadir cámaras térmicas al sistema sensorial para mejorar su robustez de noche o ante condiciones atmosféricas adversas.

Agradecimientos

Este trabajo ha sido realizado gracias al Proyecto de Investigación de Excelencia de la Junta de Andalucía PREMOVE (ProyExcel_00684).

Referencias

Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., Tai, C.-L., 2022. TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, New Orleans, LA, USA, 1090–1099.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O., 2020. nuScenes: A multimodal dataset for autonomous driving. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Seattle, WA, USA, 11621–11631.
DOI: 10.1109/CVPR42600.2020.01164

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V., 2017. CARLA: An Open Urban Driving Simulator. Proceedings of the 1st Annual Conference on Robot Learning, Mountain View, California, USA, 1–16.
DOI: 10.48550/arXiv.1711.03938

Huang, K., Shi, B., Li, X., Li, X., Huang, S., Li, Y., 2022. Multi-modal sensor fusion for auto driving perception: A survey.
DOI: 10.48550/arXiv.2202.02703

linClubs, 2023. BEVFusion-ROS-TensorRT. https://github.com/linClubs/BEVFusion-ROS-TensorRT.

Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D. L., Han, S., 2023. BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. IEEE International Conference on Robotics and Automation (ICRA), London, UK, 2774–2781.
DOI: 10.1109/ICRA48891.2023.10160968

Mehr, G., Eskandarian, A., 2025. SimBEV: A Synthetic Multi-Task Multi-Sensor Driving Data Generation Tool and Dataset.
DOI: 10.48550/arXiv.2502.01894

Montenegro, J., García-Guillén, A., Castro, F. M., Martínez, J. L., Morales, J., 2024. Detección de participantes del tráfico en entornos urbanos sobre imágenes RGB y nubes de puntos 3D. Jornadas de Automática 45, Málaga, Spain.
DOI: 10.17979/ja-cea.2024.45.10870

Moreau, J., Ibanez-Guzman, J., 2023. Emergent Visual Sensors for Autonomous Vehicles. IEEE Transactions on Intelligent Transportation Systems 24 (5), 4716–4737.
DOI: 10.1109/TITS.2023.3248483

NVIDIA, 2023. CUDA-BEVFusion. https://github.com/NVIDIA-AI-IOT/Lidar_AI_Solution/tree/master/CUDA-BEVFusion.

Philion, J., Fidler, S., 2020. Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D. Computer Vision–ECCV: 16th European Conference, Glasgow, UK, Proceedings, Part XIV 16, 194–210.
DOI: 10.1007/978-3-030-58568-6$_1$2

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You Only Look Once: Unified, Real-Time Object Detection. Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 779–788.
DOI: 10.1109/CVPR.2016.91

Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H., 2020. PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 10526–10535.
DOI: 10.1109/CVPR42600.2020.01054

Song, Z., He, Z., Li, X., Ma, Q., Ming, R., Mao, Z., Pei, H., Peng, L., Hu, J., Yao, D., et al., 2023. Synthetic datasets for autonomous driving: A survey. IEEE Transactions on Intelligent Vehicles 9 (1), 1847–1864.
DOI: 10.1109/TIV.2023.3331024

Srivastav, A., Mandal, S., 2023. Radars for autonomous driving: A review of deep learning methods and challenges. IEEE Access 11, 97147–97168.
DOI: 10.48550/arXiv.2306.09304

Urmila., O., Megalingam, R. K., 2020. Processing of LiDAR for Traffic Scene Perception of Autonomous Vehicles. International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 298–301.
DOI: 10.1109/ICCSP48568.2020.9182175

Vora, S., Lang, A. H., Helou, B., Beijbom, O., 2020. PointPainting: Sequential Fusion for 3D Object Detection. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Seattle, WA, USA, 4604–4612.
DOI: 10.1109/CVPR42600.2020.00466

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.
DOI: 10.1007/s11263-023-01784-z

Xu, D., Anguelov, D., Jain, A., 2018. PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 244–253.
DOI: 10.1109/CVPR.2018.00033

Yang, B., Li, J., Zeng, T., 2025. A Review of Environmental Perception Technology Based on Multi-Sensor Information Fusion in Autonomous Driving. World Electric Vehicle Journal 16 (1).
DOI: 10.3390/wevj16010020