CEA Comité Español de Automática
XLV JORNADAS AUTOMÁTICA MÁLAGA 2024
UNIVERSIDADE DA CORUÑA

Jornadas de Automática

Detección 3D en infraestructuras inteligentes basada en imagen monocular

Borau Bernad, J.[a,*], Ramajo Ballester, Á.[a], Armingol Moreno, J.M.[a]

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

To cite this article: Borau Bernad, J., Ramajo Ballester, Á., Armingol Moreno, J.M. 2024. 3D object detection in smart infrastructures based on monocular image.
Jornadas de Automática, 45. https://doi.org/10.17979/ja-cea.2024.45.10737

Resumen

En los últimos años, los avances en Deep Learning y Visión por Computador han impulsado el desarrollo de algoritmos de detección monocular aplicados a la gestión y seguridad del tráfico urbano, con el objetivo de optimizar la recolección de datos en entornos urbanos para las ciudades inteligentes del futuro. Sin embargo, estos esfuerzos han estado predominantemente enfocados en la extracción de datos desde la perspectiva del vehículo, pasando por alto las ventajas que ofrece el uso de cámaras instaladas en la infraestructura. Este artículo se centra en el estudio de la obtención de datos tridimensionales del tráfico desde esta perspectiva alternativa, aprovechando un punto de vista superior para evitar oclusiones y obtener información más precisa sobre el tamaño y la posición de los vehículos. Así, esta investigación propone un nuevo enfoque metodológico para la integración de sistemas de visión por computador basados en infraestructuras, aplicados a los Sistemas Inteligentes de Transporte.

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

3D object detection in smart infrastructures based on monocular image

Abstract

In recent years, advances in Deep Learning and Computer Vision have driven the development of monocular detection algorithms applied to the management and safety of urban traffic, with the goal of optimizing data collection in urban environments for future smart cities. However, these efforts have predominantly focused on data extraction from the vehicle's perspective, overlooking the advantages offered by the use of cameras installed in infrastructure. This article focuses on the study of obtaining three-dimensional traffic data from this alternative perspective, leveraging a superior viewpoint to avoid occlusions and obtain more accurate information about the size and position of vehicles. Thus, this research proposes a new methodological approach for the integration of computer vision systems based on infrastructure, applied to Intelligent Transportation Systems.

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

1. Introducción

Durante la última década, la integración continua de las tecnologías digitales en los entornos urbanos ha hecho posible la aparición de conceptos innovadores como las Ciudades Inteligentes o *Smart Cities*. Estos conceptos representan la próxima generación de ciudades, buscando mejorar la eficiencia y seguridad de las urbes actuales. La visión por computador y la inteligencia artificial emergen como dos de las principales herramientas para el desarrollo de las *Smart Cities*, responsables de la obtención precisa de datos del entorno que se utilizan posteriormente para decisiones en tiempo real, como en la navegación de vehículos autónomos y en la gestión inteligente del tráfico.

Los sistemas de detección monocular basados en Deep

*Autor para correspondencia: jborau@pa.uc3m.es
Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

Learning se han consolidado como un grupo de tecnologías fiables para la obtención de datos a partir de cámaras convencionales, destacando por su bajo coste y sencillez comparados con sistemas basados en LiDAR o cámaras binoculares. Sin embargo, la mayoría de aplicaciones actuales se centran en la perspectiva del vehículo, lo que presenta desventajas como oclusiones y dificultad para extraer información precisa de los vehículos desde vistas laterales o frontales.

Para abordar estas limitaciones, surge la posibilidad de realizar detección utilizando cámaras situadas en la infraestructura, como semáforos, báculos y pórticos de control de tráfico, obteniendo así un punto de vista más elevado y una perspectiva que permite capturar información más precisa de los vehículos. Además, estos sistemas presentan una significativa ventaja económica al permitir monitorear todos los vehículos que atraviesan una ubicación específica utilizando una sola cámara, lo que no solo reduce el coste de implementación sino que también simplifica la gestión y el mantenimiento de la infraestructura de monitoreo.

Por lo tanto, la investigación que se expone en este artículo se centra en la adaptación de algoritmos de visión monocular existentes para dotarles de la capacidad de realizar la localización de vehículos desde el punto de vista de la infraestructura, evaluando los resultados obtenidos y comparando su precisión con la obtenida originalmente. Este artículo demuestra la eficacia de la detección 3D de objetos desde infraestructuras urbanas, proporcionando una alternativa más precisa y económica que los métodos basados en vehículos. Los resultados obtenidos evidencian mejoras significativas en la precisión de detección, proponiendo un enfoque novedoso para la integración de sistemas de visión por computador en el ámbito de las *Smart Cities*.

2. Estado del arte actual

Para el desarrollo de algoritmos de detección tridimensional de vehículos basados en imágenes monoculares y Deep Learning, es necesario tanto el desarrollo de la arquitectura del modelo a implementar, como el conjunto de datos, o dataset, que se utilizará para entrenar el algoritmo.

De hecho, gran parte de la capacidad de detección del modelo depende de la calidad y nivel de variedad de los datos proporcionados durante el entrenamiento. Además, la similitud de los datos con los que entrena el algoritmo repercute directamente en su precisión, por lo que es importante la selección adecuada del dataset teniendo en cuenta cual va a ser la aplicación real del sistema de visión. En la Tabla 1, se presentan algunos de los datasets más importantes hoy en día, junto con una breve descripción de los datos que incluyen y el punto de vista en el que están enfocados.

Entre los datasets de la Tabla 1, destaca la popularidad en la comunidad científica de KITTI dataset, que se presenta como el principal conjunto de datos de referencia a la hora de comparar las capacidades de detección de los algoritmos desarrollados. Esto es debido principalmente a su marco de referencia centrado en las métricas AP11 y AP40, que permiten evaluar los modelos entrenados de una manera estandarizada y sencilla (Geiger et al., 2012).

Tabla 1: Selección de datasets disponibles públicamente para la detección 3D de vehículos, incluyendo el número de imágenes (Img.) y objetos (Obj.) en cada uno, así como la perspectiva (Persp.) desde la cual se capturan los datos. Las perspectivas son Vehículos (V), Infraestructura (I), y una combinación de ambos, cooperativos (C). (Ramajo-Ballester et al., 2023).

Dataset	Img.	Obj.	Persp.
KITTI (Geiger et al., 2013)	15k	200k	V
H3D (Patil et al., 2019)	83k	1.1M	V
nuScenes (Caesar et al., 2020)	1.4M	1.4M	V
WaymoOpen (Sun et al., 2020)	1M	12M	V
PandaSet (Xiao et al., 2021)	48k	1.3M	V
ONCE (Mao et al., 2021)	7M	417k	V
KITTI-360 (Liao et al., 2023)	300k	68k	V
Rope3D (Ye et al., 2022)	50k	1.5M	I
A9Dataset (Creß et al., 2022)	5.4k	215k	I
DAIR-V2X (Yu et al., 2022)	71k	1.2M	C
Argoverse2 (Wilson et al., 2023)	~1M	150k	V

A su vez, durante los últimos años se han desarrollado cada vez más precisos algoritmos de detección que se benefician de arquitecturas más complejas para la obtención de información más precisa del entorno. No obstante, este incremento en la complejidad de los modelos supone un aumento de la potencia de cálculo necesaria, tanto para el proceso de entrenamiento como la inferencia de resultados.

En la Tabla 2, se muestran algunos de los algoritmos con mejor puntuación en las métricas de KITTI dataset, ilustrando como conforme avanza la investigación en este ámbito se obtienen resultados cada vez más fiables y precisos.

Tabla 2: Comparativa del rendimiento de varios algoritmos de Detección 3D de Objetos a partir de Imágenes Monoculares, evaluados según las métricas AP40 en las categorías Fácil, Moderado y Difícil en KITTI Dataset. La puntuación de cada algoritmo está extraída de sus respectivos artículos científicos.

	Resultado AP40		
Modelo	**Fácil**	**Mod.**	**Difícil**
MONOPAIR (Chen et al., 2020)	16.28	12.30	10.42
SMOKE (Liu et al., 2020)	14.03	9.76	7.84
MONORCNN (Shi et al., 2021)	18.36	12.65	10.03
PGD (Wang et al., 2021b)	19.05	11.73	9.39
MONODLE (Ma et al., 2021)	17.23	12.26	10.29
MONOCON (Liu et al., 2021)	22.50	16.46	13.95
DEVIANT (Kumar et al., 2022)	21.88	14.46	11.89
MONODDE (Li et al., 2022)	24.93	17.14	15.10
MONOLSS (Li et al., 2023)	26.11	19.15	16.94

3. Detección monocular desde la infraestructura

Durante esta sección se describirá el proceso de adaptación de los algoritmos de visión para la detección de vehículos desde el punto de vista de la infraestructura. Para ello, se debe seleccionar un dataset que contenga datos con esta perspectiva visual, para posteriormente entrenar los modelos seleccionados aportándoles, así, una mejor capacidad de detección que los modelos originales entrenados con datos capturados desde cámaras instaladas en el vehículo.

3.1. Dataset DAIR-V2X

En visión por computador, es importante que los modelos sean entrenados con datos similares a su aplicación real, más

importante aún en tareas tan complejas como la inferencia de posición tridimensional a partir de imágenes. En la Figura 1, se ilustra la comparativa de los resultados obtenidos al realizar la inferencia en una imagen desde la infraestructura con un modelo entrenado en KITTI dataset, y ese mismo modelo entrenado con DAIR-V2X, reflejando la importancia de realizar el entrenamiento de los algoritmos con la perspectiva similar a su aplicación final.

Figura 1: Inferencia con el algoritmo SMOKE entrenado con KITTI Dataset (arriba) y entrenado con DAIR-V2X Dataset (abajo) sobre una imagen capturada desde la infraestructura.

Para esta investigación, se ha seleccionado el dataset DAIR-V2X-I, la sección del dataset DAIR-V2X (Yu et al., 2022) que contiene los datos obtenidos desde la infraestructura. Este conjunto de datos incluye 10.000 imágenes capturadas desde báculos y pórticos de control de tráfico, junto con más de 500.000 objetos etiquetados. A pesar de no ser utilizados en este estudio, el dataset también incluye los mapas de puntos LiDAR capturados de forma simultánea a cada imagen, lo cual podría ser de gran utilidad en posibles investigaciones futuras basadas en detección mediante LiDAR.

El dataset ha sido convertido a un formato similar al usado en KITTI dataset, adaptando su estructura con herramientas proporcionadas por sus desarrolladores, permitiendo utilizar todas las herramientas de evaluación disponibles en KITTI, así como otras herramientas de desarrollo disponibles en la biblioteca de MMDetection3D (MMDetection3D Contributors, 2020).

3.2. *Algoritmos de detección*

Una vez escogido un dataset apropiado que sea capaz de reflejar correctamente el punto de vista desde la infraestructura, se puede continuar con el proceso de selección de los modelos a entrenar para la detección monocular de vehículos. Para ello, se han seleccionado 3 de los algoritmos más prometedores diseñados para KITTI dataset, aportando así una visión general de como se comportan los algoritmos actuales si son adaptados a la detección desde la infraestructura. La elección de los modelos se basa tanto en su rendimiento comprobado en el KITTI dataset como en su capacidad de adaptarse a la perspectiva desde la infraestructura.

En primer lugar, se ha seleccionado el modelo SMOKE (Liu et al., 2020), Single-Stage Monocular 3D Object Detection via Keypoint Estimation, qué se caracteriza por ser un algoritmo de una etapa, capaz de obtener unos resultados precisos sin ser un modelo de elevada complejidad. La arquitectura de SMOKE se compone de un primer grupo de redes neuronales, o backbone, que generan un mapa de características directamente de la imagen. Concretamente, SMOKE utiliza el backbone DLA-34 (Yu et al., 2017) para esta tarea. A partir del mapa de características obtenido, SMOKE utiliza dos redes neuronales, también llamadas cabezas detectoras, para la obtención de los datos tridimensionales de los objetos. Por un lado, la cabeza de detección de puntos clave infiere el centro de los vehículos en la imagen, mientras que la cabeza de regresión tridimensional calcula los parámetros tridimensionales de ese píxel en la imagen, como son la profundidad, dimensiones del vehículo y su rotación. Finalmente, se obtiene la posición tridimensional utilizando la matriz de parámetros intrínsecos de la cámara y los datos obtenidos en las cabezas detectoras.

En segundo lugar, se ha escogido el modelo PGD (Wang et al., 2021b), Probabilistic and Geometric Depth, que utiliza la arquitectura del modelo FCOS3D (Wang et al., 2021a). El modelo comienza generando el mapa de características mediante un backbone ResNET (He et al., 2015) y un conjunto de cabezas detectoras extraen la información tridimensional de los objetos de forma similar al algoritmo SMOKE. La principal diferencia del modelo PGD, con respecto a FCOS3D, radica en su novedosa forma de obtener la profundidad de los vehículos, la cual es calculada mediante un método probabilístico y otro geométrico, siendo la profundidad final una media ponderada de ambas, obteniendo una precisión adicional en la posición inferida de los vehículos.

Por último, se ha optado por el modelo MONOCON (Liu et al., 2021), Monocular Context, un modelo con mayor complejidad que los anteriores, que introduce la metodología de aportar información auxiliar durante el entrenamiento para mejorar el rendimiento del algoritmo. MONOCON utiliza el backbone DLA-34, al igual que SMOKE, para generar el mapa de características a partir de la imagen. A continuación, presenta dos grupos de cabezas detectoras: las cabezas de regresión tridimensional y las cabezas de contexto auxiliar. Las cabezas de regresión tridimensional se usan durante el entrenamiento y la inferencia y son las encargadas de realizar la detección 3D de los objetos, localizando el centro bidimensional de los vehículos, su profundidad, y posteriormente sus dimensiones y orientación. Por otro lado, las cabezas de contexto auxiliar, son únicamente utilizadas por el algoritmo durante la fase de entrenamiento, y sirven para aportar información adicional de los objetos, como las esquinas de las Bounding Boxes o su dimensión bidimensional, permitiendo al algoritmo obtener una capacidad de generalización mejorada.

4. Experimentos realizados

4.1. Detalles de la implementación

Los detalles de implementación de los algoritmos explicados en el apartado anterior varían ligeramente, sin embargo, el entorno computacional utilizado ha sido el mismo en todos los casos. Se ha contado con un ordenador equipado con una tarjeta gráfica Nvidia GeForce 3090 TI, con 24 GB de VRAM y se ha instalado la librería CUDA 12.2, que permite utilizar la aceleración de GPU para el entrenamiento de los algoritmos. La implementación de los modelos ha sido realizada con la librería MMDetection3D 1.2, basada en PyTorch 1.13, que aglutina los datasets y modelos más significativos del estado del arte bajo una arquitectura software común.

A la hora de realizar el entrenamiento de los modelos, el dataset se ha dividido en splits de entrenamiento, validación y test. El 80 % del dataset se ha asignado al entrenamiento, el 10 % a la validación y el 10 % restante a las pruebas de evaluación final. Esta distribución asegura que cada modelo sea evaluado sobre conjuntos de datos que no han sido vistos anteriormente, garantizando así resultados que reflejen adecuadamente la precisión y la capacidad de generalización de los algoritmos.

El dataset original contiene 10.000 imágenes y los algoritmos han sido configurados para solamente detectar los objetos de tipo coche, con un máximo de 30 por imagen. No obstante, para mejorar la capacidad de generalización de los algoritmos, se ha implementado una etapa inicial de preprocesamiento de datos. Al introducir las imágenes a la red neuronal, se establece una probabilidad del 50 % de ser rotadas sobre el eje horizontal y un 30 % de ser escaladas con un factor entre 0.2 y 0.4, utilizando las herramientas de preprocesamiento de MMDetection3D.

En primer lugar, el entrenamiento de SMOKE se ha programado para durar 100 épocas, con etapas de validación cada 5 y un batch size de 4. Las imágenes después del preprocesamiento son reducidas con un factor de 1/4 desde su resolución original, 1920x1080. El optimizador escogido es el Adam, con una tasa de aprendizaje inicial de $2{,}5 \times 10^{-4}$, la cual se reducirá al comienzo de la época 51 a un valor de $2{,}5 \times 10^{-5}$, para realizar un ajuste de parámetros más suave al final del entrenamiento.

En segundo lugar, el entrenamiento del modelo PGD se diseñó para durar 48 épocas, con intervalos de validación cada 12 y un batch size de 1, debido a las restricciones en la memoria de la GPU. Sin embargo, en este modelo las imágenes no son reescaladas y son directamente cargadas en su resolución original de 1920x1080 píxeles. El optimizador utilizado es el SGD, con un momentum de 0.9. La tasa de aprendizaje inicial es de 1×10^{-3}, reduciéndose con un factor de 10 en la época 32 y 44, finalizando el entrenamiento con una tasa de aprendizaje de 1×10^{-5}.

Por último, el modelo MONOCON es entrenado durante 100 épocas, con etapas de validación cada 25 épocas, un batch size de 4 y, al igual que SMOKE, se utiliza el optimizador Adam. La tasa de aprendizaje inicial es de $2{,}5 \times 10^{-5}$, siendo incrementada linealmente hasta $2{,}5 \times 10^{-4}$ en la época 5 y disminuida, también linealmente, hasta un valor de $2{,}5 \times 10^{-9}$ en el final del entrenamiento.

4.2. Métricas de evaluación

La principal ventaja de transformar el DAIR-V2X dataset al formato utilizado por KITTI dataset es la posibilidad de usar las herramientas de implementación, incluyendo las métricas de evaluación AP40, ampliamente utilizadas por la comunidad científica. Estas métricas clasifican los objetos en 3 categorías: fácil, moderado y difícil, en función del tamaño de los objetos o si están parcialmente ocluidos. Posteriormente, se calcula la Intersección sobre Unión (IoU) de la predicción de los objetos con su valor teórico, filtrando las predicciones que superen un determinado valor, por ejemplo 0.50 o 0.70. Finalmente, los objetos que superen este filtro se utilizan para calcular la precisión media en 40 puntos, siguiendo la siguiente ecuación:

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

Siendo $R = [r_0, r_0 + \frac{r_1 - r_0}{N-1}, r_0 + \frac{2(r_1 - r_0)}{N-1}, ..., r_1]$ y $P(r) = \max_{r':r' \geq r} P(r')$. Las métricas AP40 proporcionan un reflejo más detallado del comportamiento del algoritmo que las métricas AP11. Sin embargo, las métricas AP11 siguen siendo ampliamente reconocidas en el mundo de la detección tridimensional de objetos.

4.3. Resultados

Una vez finalizados los entrenamientos de los algoritmos explicados anteriormente se ha procedido a realizar una etapa de test con cada uno de los modelos elaborados sobre imágenes del dataset DAIR-V2X. Los resultados obtenidos se muestran en la Tabla 3. Estos resultados sirven para comparar el rendimiento de los modelos SMOKE, PGD y MONOCON según las métricas AP40, en el nivel de IoU igual a 0.7, aportando una visión más amplia de las capacidades de detección de cada uno de los algoritmos con imágenes tomadas desde la infraestructura. Los datos revelan que tanto el algoritmo PGD como SMOKE se comportan de forma muy similar en las tres condiciones evaluadas. Sin embargo, el modelo MONOCON es ligeramente menos eficaz en la detección de vehículos, probablemente debido a su enfoque en contextos auxiliares que no resultan tan útiles como los de los otros algoritmos al realizar la detección desde un punto de vista más elevado.

Tabla 3: Resultados obtenidos en las pruebas de test de los experimentos explicados anteriormente, evaluados según las métricas AP40 en una Intersección sobre Unión (IoU) de 0.7.

	AP_{40} IoU 0.7		
Modelo	**Fácil**	**Moderado**	**Difícil**
SMOKE	59.82	51.14	50.93
MONOCON	51.35	43.41	43.19
PGD	59.99	50.87	49.54

Además de la tabla comparativa anterior, es interesante valorar la mejora de rendimiento que tienen los algoritmos entrenados con imágenes capturadas desde la infraestructura, comparado con aquellos entrenados con imágenes desde el punto de vista del vehículo. Para analizar este aumento de rendimiento, se muestra en la Tabla 4 una comparación de la puntuación obtenida por cada uno de los modelos entrenados

Tabla 4: Comparativa de los resultados obtenidos en las métricas AP40 y IoU de 0.70 de los modelos entrenados en DAIR-V2X con los resultados teóricos establecidos por los modelos en KITTI.

	DAIR-V2X AP_{40} IoU 0.7			**KITTI** AP_{40} IoU 0.7		
Modelo	**Fácil**	**Moderado**	**Difícil**	**Fácil**	**Moderado**	**Difícil**
SMOKE	59.82	51.14	50.93	14.03	9.76	7.84
MONOCON	51.35	43.41	43.19	22.50	16.46	13.95
PGD	59.99	50.87	49.54	19.05	11.73	9.39

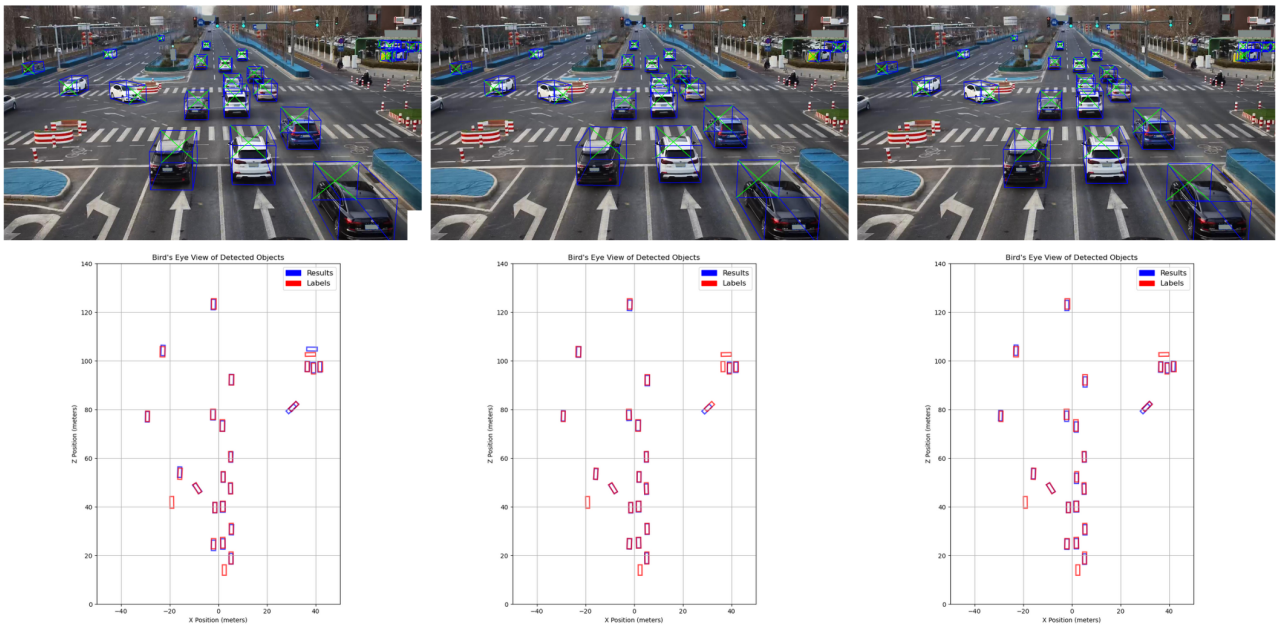

Figura 2: Inferencias realizadas con SMOKE (izquierda), PGD (centro) y MONOCON (derecha) sobre una imagen de DAIR-V2X dataset, ilustrando las Bounding Boxes (BB) proyectadas sobre la imagen y su punto de vista de pájaro (bird's-eye view). Esta representación destaca la gran precisión alcanzada por los modelos desarrollados, identificando los objetos correctamente en diferentes distancias y perspectivas.

con KITTI dataset y con DAIR-V2X dataset, donde se puede apreciar el amplio margen de mejora que supone la detección desde la infraestructura. Concretamente, todos los modelos experimentan mejoras significativas al ser evaluados desde la perspectiva de la infraestructura, destacando así las ventajas de realizar la detección de vehículos desde esta nueva perspectiva en comparación con el punto de vista del vehículo.

Por último, en la Figura 2 se muestran los resultados de inferencia con cada uno de los modelos entrenados, exponiendo la gran capacidad de detección y fiabilidad que resulta de los experimentos realizados. Además de la proyección de la Bounding Box de los vehículos sobre la imagen, se ilustra una representación de los vehículos vistos desde arriba según las etiquetas del dataset, junto con la correspondiente predicción, para demostrar de una forma más intuitiva la precisión alcanzada por el sistema de visión desarrollado.

5. Conclusiones y trabajos futuros

El objetivo del presente artículo ha sido explorar las posibles implementaciones de la detección de objetos 3D desde el punto de vista de la infraestructura y sus posibles ventajas en la capacidad de detección de los algoritmos. Durante esta investigación, se han implementado y entrenado modelos de detección 3D monocular, concretamente SMOKE, PGD y MONOCON con el dataset DAIR-V2X, cuya principal característica diferenciadora es el punto de vista de sus imágenes capturadas desde la infraestructura, en comparación con la mayoría de datasets disponibles públicamente.

Los resultados obtenidos demuestran las ventajas de aplicar algoritmos de detección 3D a los sistemas de detección de infraestructuras, obteniendo resultados que evidencian una mayor precisión en comparación con los mismos modelos entrenados en el conjunto de datos KITTI. Debido al mayor campo de visión que proporcionan las cámaras instaladas en las infraestructuras, los algoritmos rinden con mayor precisión y eficacia, beneficiándose de la información adicional que aporta esta perspectiva de vista. Por lo tanto, este rendimiento superior obtenido en los experimentos realizados plantea un nuevo camino para lineas de investigación futuras relativas a las Smart Cities y los Sistemas Inteligentes de Transporte.

Por un lado, futuras investigaciones se pueden enfocar en la implementación de nuevos algoritmos de visión, así como adaptar otros disponibles actualmente para su optimización en la detección desde la infraestructura, mejorando todavía más los resultados obtenidos en este estudio. Por otro lado, existe un amplio campo de investigación en cuanto a las posibles aplicaciones que surgen de la instalación de sistemas de visión 3D en la infraestructura viaria. Ejemplos de estas aplicaciones podrían ser el diseño de sistemas de gestión inteligente de tráfico basados en sensores situados en la infraestructura, que analicen la información del entorno para el control de semáforos en tiempo real o la priorización inteligente de vehículos de

emergencia.

En resumen, este estudio ha confirmado la eficacia de las tecnologías de detección de objetos 3D implementadas desde infraestructuras urbanas, demostrando mejoras significativas en la precisión y eficiencia de los sistemas de transporte. Estos resultados no solo validan la utilidad de los algoritmos en entornos reales, sino que también abren el camino para futuras investigaciones. Continuar explorando y expandiendo estas tecnologías será vital para maximizar sus beneficios en la gestión y seguridad de las ciudades inteligentes, contribuyendo así a los esfuerzos globales por alcanzar los objetivos de desarrollo sostenible.

Agradecimientos

Subvenciones PID2021-124335OB-C21, PID2022-140554OB-C32 y PDC2022-133684-C31 financiadas por MCIN/AEI/10.13039/501100011033.

Referencias

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O., June 2020. nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Chen, Y., Tai, L., Sun, K., Li, M., June 2020. Monopair: Monocular 3d object detection using pairwise spatial relationships. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Creß, C., Zimmer, W., Strand, L., Fortkord, M., Dai, S., Lakshminarasimhan, V., Knoll, A., 2022. A9-dataset: Multi-sensor infrastructure-based dataset for mobility research. In: 2022 IEEE Intelligent Vehicles Symposium (IV). pp. 965–970.
DOI: 10.1109/IV51971.2022.9827401

Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets robotics: The kitti dataset. The International Journal of Robotics Research 32, 1231 – 1237.
DOI: 10.1177/0278364913491297

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.
DOI: 10.1109/CVPR.2012.6248074

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition.

Kumar, A., Brazil, G., Corona, E., Parchami, A., Liu, X., 2022. Deviant: Depth equivariant network for monocular 3d object detection.

Li, Z., Jia, J., Shi, Y., 2023. Monolss: Learnable sample selection for monocular 3d detection.

Li, Z., Qu, Z., Zhou, Y., Liu, J., Wang, H., Jiang, L., June 2022. Diversity matters: Fully exploiting depth clues for reliable monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2791–2800.

Liao, Y., Xie, J., Geiger, A., 2023. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (3), 3292–3310.
DOI: 10.1109/TPAMI.2022.3179507

Liu, X., Xue, N., Wu, T., 2021. Learning auxiliary monocular contexts helps monocular 3d object detection.

Liu, Z., Wu, Z., T'oth, R., 2020. Smoke: Single-stage monocular 3d object detection via keypoint estimation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 4289–4298.
DOI: 10.1109/CVPRW50498.2020.00506

Ma, X., Zhang, Y., Xu, D., Zhou, D., Yi, S., Li, H., Ouyang, W., June 2021. Delving into localization errors for monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4721–4730.

Mao, J., Niu, M., Jiang, C., Liang, H., Chen, J., Liang, X., Li, Y., Ye, C., Zhang, W., Li, Z., Yu, J., Xu, H., Xu, C., 2021. One million scenes for autonomous driving: Once dataset.

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

Patil, A., Malla, S., Gang, H., Chen, Y.-T., 2019. The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 9552–9557.
DOI: 10.1109/ICRA.2019.8793925

Ramajo-Ballester, Á., de la Escalera Hueso, A., Armingol Moreno, J. M., 2023. 3D Object Detection for Autonomous Driving: A Practical Survey. In: 9th International Conference on Vehicle Technology and Intelligent Transport Systems. pp. 64–73.
DOI: 10.5220/0011748400003479

Shi, X., Ye, Q., Chen, X., Chen, C., Chen, Z., Kim, T.-K., October 2021. Geometry-based distance decomposition for monocular 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15172–15181.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D., June 2020. Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Wang, T., Zhu, X., Pang, J., Lin, D., 2021a. Fcos3d: Fully convolutional one-stage monocular 3d object detection.

Wang, T., Zhu, X., Pang, J., Lin, D., 2021b. Probabilistic and geometric depth: Detecting objects in perspective.

Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J. K., Ramanan, D., Carr, P., Hays, J., 2023. Argoverse 2: Next generation datasets for self-driving perception and forecasting.

Xiao, P., Shao, Z., Hao, S., Zhang, Z., Chai, X., Jiao, J., Li, Z., Wu, J., Sun, K., Jiang, K., Wang, Y., Yang, D., 2021. Pandaset: Advanced sensor suite dataset for autonomous driving. In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). pp. 3095–3101.
DOI: 10.1109/ITSC48978.2021.9565009

Ye, X., Shu, M., Li, H., Shi, Y., Li, Y., Wang, G., Tan, X., Ding, E., June 2022. Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection task. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21341–21350.

Yu, F., Wang, D., Darrell, T., 2017. Deep layer aggregation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2403–2412.
DOI: 10.1109/CVPR.2018.00255

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.