

Cano-Ortiz, S., Sainz-Ortiz, E., Lloret Iglesias, L., Martínez Ruiz del Árbol, P., Castro-Fresno, Daniel (2024). *Sistema de visión artificial basado en aprendizaje profundo para la detección automática de defectos superficiales en pavimentos* (IX ELAGEC2024).

SISTEMA DE VISIÓN ARTIFICIAL BASADO EN APRENDIZAJE PROFUNDO PARA LA DETECCIÓN AUTOMÁTICA DE DEFECTOS SUPERFICIALES EN PAVIMENTOS

Saúl Cano-Ortiz ¹ – saul.cano@unican.es

Eugenio Sainz-Ortiz ¹ – eugenio.sainz@unican.es

Lara Lloret Iglesias ² – lara.cern@gmail.com

Pablo Martínez Ruiz del Árbol ² – pablo.martinez@unican.es

Daniel Castro-Fresno* ¹ – daniel.castro@unican.es

¹Grupo de Investigación de Tecnología de la Construcción (GITECO), Universidad de Cantabria, España

²Instituto de Física de Cantabria, España.

RESUMEN

La monitorización y evaluación de defectos en pavimentos es esencial para garantizar la seguridad y longevidad de las infraestructuras viales, requiriendo metodologías y herramientas de gestión que sean eficientes, confiables, automáticas y de bajo costo. Las soluciones actuales, fundamentadas en la inspección visual por expertos y el uso de vehículos especializados, no satisfacen estas exigencias. En respuesta a esta problemática, se ha desarrollado un sistema de visión artificial basado en aprendizaje profundo, capaz de detectar y localizar múltiples defectos en carreteras. Este sistema ha sido optimizado mediante la incorporación de un modelo generativo, lo que ha permitido la creación de un índice de condición del pavimento basado en las detecciones obtenidas. Se ha implementado un software de gestión que integra imágenes georreferenciadas, resultados de detección, el índice calculado y datos de tráfico abiertos, ofreciendo una herramienta integral para la toma de decisiones en el mantenimiento de carreteras. Finalmente, se ha evaluado el potencial de un modelo de lenguaje visual para mejorar aún más este sistema. En definitiva, se ha creado una solución innovadora, automática, robusta y rentable, diseñada para digitalizar y optimizar la planificación estratégica del mantenimiento de la red vial.

PALABRAS CLAVE

Mantenimiento inteligente de carreteras; Digitalización; Visión Artificial; Aprendizaje Profundo; Inteligencia Artificial Generativa.

INTRODUCCIÓN

Las carreteras se deterioran debido al flujo continuo de tráfico pesado, condiciones climáticas, envejecimiento de materiales, construcción deficiente, etc. (Cano-Ortiz et al., 2022). Su deterioro tiene un impacto negativo económico, medioambiental, y de seguridad (Ai et al., 2023). Las soluciones actuales se dividen entre campañas de inspección visual humana, que son costosas, ineficientes y subjetivas, y vehículos especializados, que, aunque son semiautomáticos, presentan un coste muy elevado. Por lo tanto, es fundamental desarrollar una metodología que permita monitorizar y auscultar el pavimento de manera frecuente, eficiente, económica y automática (Koch et al., 2015). La frecuencia en la monitorización es esencial para fomentar el mantenimiento preventivo, realizar micro-rehabilitaciones y prevenir accidentes. La eficiencia es necesaria dado el gran tamaño de la red de carreteras, mientras que el bajo costo es crucial para ajustarse a los presupuestos limitados de las administraciones encargadas de la conservación. Además, la automatización del proceso es indispensable para reducir los tiempos de análisis y acelerar la toma de decisiones en el diseño del plan de mantenimiento estratégico.

El alto rendimiento de las soluciones basadas en Inteligencia Artificial (IA) en múltiples campos de Ingeniería Civil ha generado interés en la detección automática de defectos en el pavimento (Alzubaidi et al., 2021). En particular, la implementación de sistemas de visión artificial, que emulan la comprensión de tareas de visión humana, se basa en modelos de aprendizaje profundo, un subcampo de la inteligencia artificial. Estos sistemas son capaces de reconocer y clasificar los deterioros superficiales a partir de imágenes capturadas con cámaras de bajo coste (Zhang et al., 2021).

Sin embargo, el estado del arte presenta varias limitaciones significativas. En primer lugar, existe una escasez de bancos de imágenes de acceso abierto con anotaciones de alto volumen y diversidad, lo que dificulta el desarrollo de soluciones avanzadas. De hecho, solo el 22% de los conjuntos de datos en la investigación están disponibles públicamente (Cano-Ortiz et al., 2022). Por otro lado, gran parte de los estudios se concentran exclusivamente en defectos de tipo fisura, dejando de lado otros tipos de deterioros que también requieren atención. El 69,6% de los estudios se centran en el procesamiento de fisuras basado en imágenes (Alzubaidi et al., 2021).

Otro desafío importante es el elevado coste del etiquetado de imágenes, un proceso esencial para la implementación de soluciones robustas de aprendizaje supervisado. Por otra parte, se observa una falta de diseño de un índice de condición del pavimento, herramienta fundamental para evaluar el estado de las superficies. Adicionalmente, aún no se han implementado softwares de gestión de pavimentos potenciados por inteligencia artificial, lo que limita las capacidades de automatización y análisis. Finalmente, es necesario diseñar herramientas explicativas que permitan determinar factores críticos como el tipo de mantenimiento necesario y los costos asociados, entre otros aspectos clave.

Este trabajo presenta un banco de imágenes anotadas y un modelo generativo que crea imágenes sintéticas de defectos, lo que reduce el costo y soluciona el desbalanceo en el etiquetado manual. Con estas imágenes, se entrenó una arquitectura capaz de detectar, clasificar y localizar deterioros en pavimentos. Se desarrolló un índice de condición del

pavimento, que se integró en un software de gestión junto con imágenes georreferenciadas, detecciones y datos de tráfico. También se demostró el potencial de un modelo de lenguaje visual para explicar los resultados. En conjunto, se diseñó un sistema automatizado, eficiente y de bajo costo para un mantenimiento de carreteras inteligente.

METODOLOGÍA

Se ha recogido un banco de imágenes (Cano-Ortiz et al., 2024) con 7099 de tamaño 640x640 con presencia de 13 tipos de defectos: malla en bloque, piel de cocodrilo, fisura diagonal, fisura transversal, fisura irregular, fisura transversal, fisura longitudinal de centro de eje, parche, bache, alcantarilla, sumidero, fisura sellada, y desprendimiento de partículas. Las imágenes han sido anotadas por expertos en pavimentos para una tarea de visión artificial de visión de objetos. Es decir, para cada defecto por imagen, se han guardado las coordenadas del polígono que encierra al defecto junto con el tipo de deterioro. De esta manera, la red de detección de objetos se podrá entrenar de manera supervisada, aprendiendo a mapear la relación entre las imágenes de entrada y las salidas (coordenadas y clase de defecto). El sistema de adquisición de imágenes es una cámara anclada al vehículo con una estructura realizada con impresora 3D y un sistema de ventosas. El sistema recoge vídeos, y de manera programática, los videos se dividen en imágenes. También, a partir de los metadatos de la imagen, se han obtenido las coordenadas espaciales.

La arquitectura seleccionada para la detección de defectos es *You Only Look at Once* versión 5 (YOLOv5) (Reis et al., 2023). Es una red basada en capas convolucionales, especialmente diseñada para tener un tiempo de inferencia reducido y métricas de detección altas. Está dividida en tres partes secuenciales: *backbone*, *neck* y *head*. El *backbone* está pre-entrenado en el conjunto de imágenes ImageNet y utiliza la arquitectura *CSPDarknet53* (Wang et al., 2019). Esta arquitectura se compone de módulos *CBS* (capa convolucional, normalización por lotes, función de activación *SiLU*) y *C3* (Reis et al., 2023). Finalmente, se añade un módulo *SPPF* (He et al., 2014). Los módulos *CBS* y *C3* se encargan de extraer características de las imágenes, mientras que el *SPPF* mejora la expresión de estas características y reduce el coste computacional. Debido a que el *backbone* está pre-entrenado, es decir, se utiliza transferencia de aprendizaje, y se reduce el tiempo de entrenamiento de YOLOv5. El *backbone* aprende las características principales de las imágenes (por ejemplo, bordes). El *neck* emplea dos tipos de módulos: *FPN* (Tsung-Yi Lin, 2016) y *PAN* (Zheng Ma, 2019). *PAN* es un mecanismo de atención que mejora la representación de características, mientras que *FPN* genera múltiples mapas de características para detectar objetos a diferentes escalas. El *head*, basado en capas convolucionales, realiza la predicción generando como salida las coordenadas de las cajas delimitadoras, la clase del defecto y la confianza de la detección. Cuando YOLOv5 produce sus detecciones, solo se mantienen aquellas que superan un umbral de confianza y son filtradas mediante la técnica de post-procesado de supresión no máxima (Hosang et al., 2017).

YOLOv5 es un excelente candidato para implementar un modelo de detección de defectos de carretera en tiempo real debido a su alta precisión y velocidad de procesamiento, lo que permite identificar y localizar defectos con rapidez y eficiencia a partir de imágenes, incluso en entornos de alta variabilidad y condiciones adversas.

Se observó que las métricas de detección relativa de los defectos minoritarios, es decir, los defectos que aparecen menos relativamente en el banco de imágenes eran menores. Por ello, se entrenó un modelo generativo para crear imágenes sintéticas de dichos defectos. Además, dicho modelo era condicional, por lo que permitía guiar la creación de imágenes sintéticas, aliviando el problema de tener que etiquetarlas para añadirlas a las imágenes originales. El modelo generativo es un *Denoising Diffusion Probabilistic Model* condicional (Ho et al., 2020), tomando como red de denoising, *UNet* (Ronneberger et al., 2015). Esta estrategia se conoce como *data augmentation*, pero a diferencia de las técnicas tradicionales basadas en cambios geométricos, se añaden imágenes sintéticas con una distribución de probabilidad similar a la de las imágenes reales. YOLOv5 fue reentrenado con imágenes reales y sintéticas, mostrando un incremento notorio en la detección de dichos deterioros minoritarios. Un modelo generativo de difusión es superior a otros modelos generativos para la creación de imágenes sintéticas, ya que ofrece un control más adecuado de las características deseadas y produce imágenes de mayor calidad y realismo. A diferencia de modelos como *Generative Adversarial Networks* (Zhang et al., 2021), que pueden sufrir de inestabilidad y colapso de modo, los modelos de difusión garantizan una mayor estabilidad y diversidad.

A partir de los resultados del modelo, que incluyen las coordenadas de la caja delimitadora, la clase de defecto y la confianza de detección, se confeccionó un índice de condición del pavimento basado en áreas. Este índice se define como la suma ponderada de áreas calculadas a partir de las coordenadas, donde la ponderación se realiza según el tipo de defecto, lo que refleja la severidad. En este caso, la severidad se ha calculado como el nivel de importancia de mantenimiento obtenido a partir de una encuesta a empresas y administraciones de conservación de carreteras. Su valor varía de 0-100%, siendo 100% una sección de carretera con extrema urgencia de mantenimiento, y 0%, una carretera ideal.

Todos los resultados, que incluyen imágenes georreferenciadas, resultados de YOLOv5 y el índice de condición del pavimento, se han integrado en una herramienta que actúa como un sistema de gestión de pavimentos enfocado en defectos superficiales. Además, para aportar un valor añadido en la toma de decisiones, se han incorporado datos abiertos de estaciones de aforo, como el porcentaje de vehículos pesados, la intensidad media diaria, entre otros, proporcionados por el Ministerio de Transportes, Movilidad y Agenda Urbana de España. Esta herramienta, sumada a la metodología de recogida de imágenes, pretende facilitar, agilizar, y automatizar, el diseño de planes de mantenimiento estratégico de carreteras para promover actuaciones rápidas a un coste reducido.

Por último, se ha querido mostrar una funcionalidad muy interesante que podría incorporar el sistema de gestión de pavimentos, analizando sus ventajas y desventajas. Se trata de un modelo de lenguaje visual (LVM). Un LVM (J. Zhang et al., 2023) integra procesamiento de lenguaje natural (texto) y Visión Artificial (imágenes) para interpretar y generar descripciones de imágenes, realizar tareas de reconocimiento visual y responder preguntas sobre contenido visual. Concretamente, se podrían utilizar para realizar preguntas sobre las imágenes para saber el tipo de mantenimiento, el coste aproximado, el tiempo de ejecución para la rehabilitación, etc. En este caso, se ha utilizado el VLM pre-entrenado Kosmos-2 (Peng et al., 2023), en una tarea de *Visual Question-Answer* (VQA). Es decir, el modelo recibe una imagen y una pregunta, devolviendo de manera

textual, la respuesta. Este tipo de sistemas ayudarían a los resultados del software de gestión de pavimentos como una funcionalidad añadida, para una mayor interpretabilidad de los resultados.

A nivel de desarrollo, la herramienta de etiquetado es *Roboflow Annotate*. El lenguaje de programación utilizado es *Python* con la librería de Aprendizaje Automático, *PyTorch*, y las de procesamiento de imagen, *Pillow* y *OpenCV*. La aplicación se ha desarrollado en *Java*, y la demostración del potencial del VLM con *Gradio*.

RESULTADOS Y DISCUSIÓN

En primer lugar, en la Tabla 1, se muestran los resultados de una métrica de detección comúnmente utilizada en detección de objetos: *mean Average Precision* (mAP). Cuanto más se aproxima su valor a la unidad, mayor es la precisión de detección y localización del modelo. En la Tabla 1 se presentan los resultados para los distintos tipos de defectos minoritarios. Una fisura sellada es una grieta reparada para prevenir daños mayores. Una fisura diagonal atraviesa el pavimento en un ángulo oblicuo. Una fisura de geometría irregular presenta un patrón desordenado y no lineal. Un bache es un hoyo causado por el desgaste y la erosión. Detectar estos defectos es crucial para mantener la seguridad vial y prolongar la vida útil de la carretera, evitando deterioros graves y costosos.

Como se puede comprobar, el hecho de entrenar YOLOv5, antes y después de incorporar imágenes sintéticas de dichos defectos minoritarios, implica unos mejores resultados en términos de mAP. Principalmente, cabe destacar la gran mejoría de las fisuras irregulares, que no tienen una geometría definida, y los mejores resultados de los baches, que son defectos que preocupan en mayor medida a las administraciones de carreteras. En términos generales, la arquitectura entrenada da resultados robustos, con un mAP elevado de 0.62. Además, cabe destacar que no ha sido necesario etiquetar las imágenes sintéticas, ya que el modelo es condicional, es decir, recibe como entrada la etiqueta del tipo de defecto a generar. Este mecanismo mejora el detector de objetos, y además, alivia el problema de la costosa anotación.

Tabla 1. Métricas de detección de YOLOv5 en la partición de *test* del banco de imágenes Mosquitonet.

	Diagonal	Irregular	Sellada	Bache	Total
Antes	0.43	0.06	0.76	0.59	0.60
Después	0.50	0.21	0.80	0.62	0.62

Se ha elegido un tramo de carretera como nuestro caso de estudio. Para hacer esta selección, el Sistema de Gestión de Pavimentos utiliza un mapa de calor cuya intensidad varía según el índice de condición del pavimento. Este mapa resalta claramente las áreas con peor estado de la carretera, mostrando un color más intenso en aquellas zonas con un índice más reducido, indicando así una condición deficiente. Una vez identificadas estas áreas problemáticas, es posible hacer *zoom* para examinar más detalladamente los deterioros. Al hacerlo, se muestran marcadores de diferentes colores que funcionan como señales de advertencia. Estas señales indican la presencia de problemas detectados, y el color varía según el índice: verde para un rango de 0-0.2, azul para 0.21-0.40, amarillo para 0.41-0.60, naranja para 0.61-0.80 y rojo para 0.81-1.00. Cada señal proporciona

imágenes de áreas problemáticas con detecciones superpuestas e información detallada sobre el índice por imagen. Además, las estaciones de aforo se muestran como círculos naranjas, cuyo tamaño depende del Tráfico Anual Promedio Diario (AADT) de vehículos pesados.

A continuación, se muestran dos casos de estudio, donde el Caso I (Figura 1) representa un tramo con urgente necesidad de mantenimiento, mientras que el Caso II (Figura 2) se trata de un tramo sin necesidad de mantenimiento urgente. En el Caso I, se trata de un tramo de carretera con presencia de un defecto con un alto grado de severidad de tipo fisura en malla. Como se puede comprobar, el índice es muy elevado, lo que sugeriría una rápida actuación de conservación. Probablemente, dado el estado del firme, sería necesario aplicar técnicas de mantenimiento correctivo. Este resultado tiene bastante sentido ya que se trata del polígono de Guarnizo (Cantabria, España) donde el flujo de tráfico pesado es elevado y continuo. Una posible explicación es que este tipo de pavimento debería soportar un alto número de ejes de vehículos pesados, además, una extensión del parque industrial fue construida años más tarde, lo que resultó en un aumento en el volumen de tráfico pesado predominantemente. En consecuencia, el pavimento ha alcanzado el número calculado de ejes para su vida útil antes de lo esperado. De hecho, se observa que en esta área industrial la mayoría de las imágenes muestran un índice elevado. Esto resalta la necesidad urgente de mantenimiento vial en este escenario.



Figura 1. (Caso I) Software de gestión de pavimentos para la detección automática de defectos superficiales a partir de imágenes.

En el Caso II (Figura 2), se trata de un escenario donde el índice es reducido, y por tanto, no sugiere una actividad de mantenimiento urgente. Concretamente, se detectan microfisuras de tipo longitudinal. Por otra parte, se trata de una autovía (S-10, Cantabria, España), con un gran porcentaje de vehículos pesados (5.7%). Entonces, se trata de un escenario donde si la administración de conservación decidiese actuar porque encuentra

múltiples imágenes consecutivas (como en este caso), con un índice adecuado reducido, podría actuar utilizando técnicas de mantenimiento preventivo. Este tipo de actuación promovería el mantenimiento preventivo, con unos costes e impacto medioambiental, más adecuados en comparación con el mantenimiento correctivo. Además, este tipo de iniciativas no solo garantizaría una mayor seguridad de la vía, sino una extensión del ciclo de vida del pavimento.

Por tanto, se valida visualmente que se trata de un sistema que proporciona un valor añadido a la detección de pavimentos cumpliendo con las necesidades de las administraciones y empresas de conservación. Por otra parte, es importante destacar que más allá de detectar donde se encuentran los defectos, y cuantificar la necesidad de conservación estableciendo una prioridad, hay otros factores que ayudarían a las entidades mencionadas. Se trata de determinar cuáles son los costes, tiempos de ejecución, descripción visual de los pavimentos para una mayor interpretabilidad, determinación del tipo de actividad de mantenimiento (ej., sellado), etc.

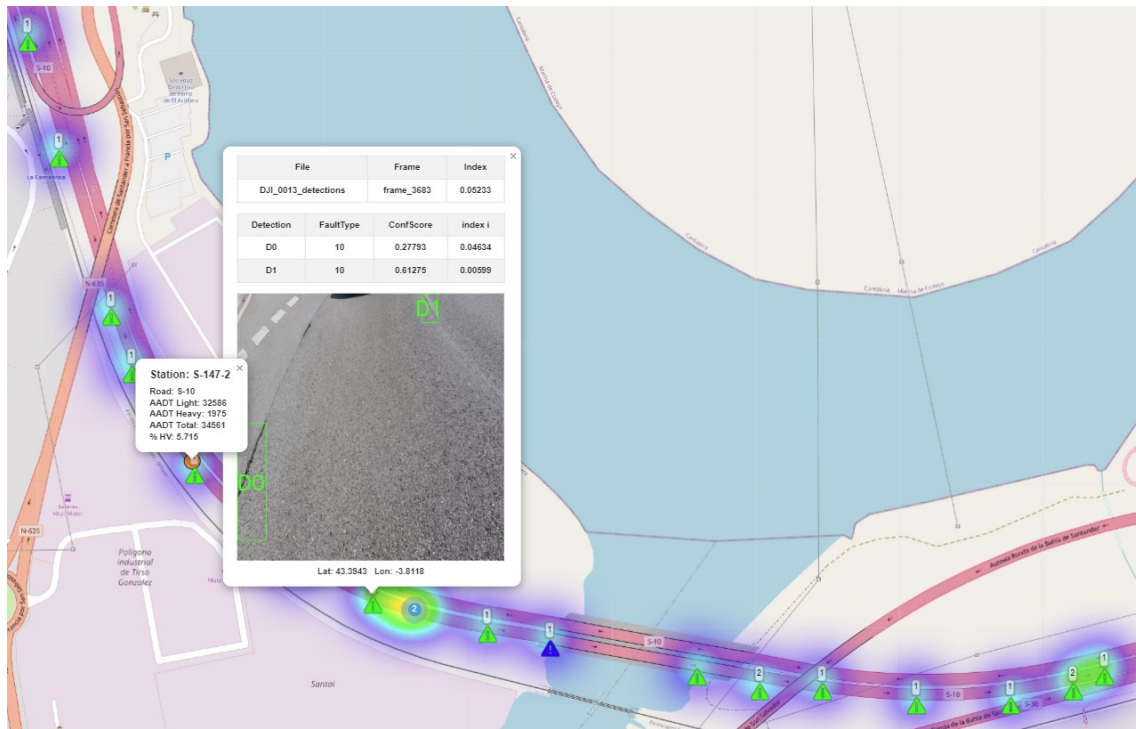


Figura 2. (Caso II) Software de gestión de pavimentos para la detección automática de defectos superficiales a partir de imágenes.

Para ello, en la Figura 3, se ha realizado un prototipo con dicha funcionalidad. Concretamente, se ha utilizado un VLM pre-entrenado en un *dataset* genérico en modo inferencia que recibe la imagen, además de una pregunta, y predice la respuesta. Como se puede observar las respuestas son muy genéricas porque realmente no está entrenado en un banco de imágenes en conjunto con múltiples proyectos de mantenimiento que alberguen técnicas de mantenimiento, costes usuales, etc. Por tanto, sería una buena estrategia para futuras investigaciones crear *datasets* con imágenes de defectos en conjunto con descripciones textuales con el histórico de mantenimiento, estructurado como un dataset de preguntas y respuestas. Es decir, que, para cada imagen, haya un conjunto de pares pregunta-respuesta, para que el modelo aprenda a realizar la tarea de

VQA. No obstante, es un reto complejo porque la elaboración de dichos bancos de datos multimodales es costosa de realizar, y además se necesitan muchos ejemplos, con lo que una buena opción puede ser pre-entrenar un VLM en un *dataset* existente genérico, y a posteriori, añadir ciertas capas basadas en *Transformers* para aplicar *fine-tuning*, entrenándolo en el *dataset* de menor volumen, pero más específico en proyectos de mantenimiento de carreteras.

Kosmos-2 Road Image Processor

Upload a road image and get: maintenance activity, expected cost, severity, and description

Upload Image

Description

An image of a road with cracks

Maintenance activity

What type of maintenance should be approached? To maintain the road's condition, it is essential to address the cracks and potholes on the road. Cracks and potholes can lead to uneven and potentially dangerous driving conditions, which can lead drivers to lose control of their vehicles. To address these issues, it's essential to inspect the road regularly for any signs of

Severity

What is the severity? The severity of the cracked road surface can vary depending on the depth and extent of the damage. In the image, the road appears to be relatively shallow, with cracks extending from the pavement to the ground. However, the severity of this damage may vary depending upon the depth of the cracks and the extent of damage. The

Expected cost

What is the approximate cost of maintaining this road? The approximate cost to maintain this road is approximately \$10,000 per year. This includes repairing and maintaining the cracked and damaged road surface, as well as maintaining the electrical and water lines. The cost of this maintenance is typically borne by the local municipality or the owner of the property.

Flag

Clear Submit

Figura 3. Prueba de concepto del potencial de los modelos de lenguaje visual para la interpretabilidad de los resultados en la toma de decisiones estratégica aplicada al mantenimiento de carreteras.

CONCLUSIONES

Se ha diseñado un sistema de gestión de pavimentos potenciado con Inteligencia Artificial capaz de reconocer y clasificar 13 tipos de defectos en pavimentos. Para lograrlo, se entrenó una red de detección de objetos YOLOv5, que se mejoró utilizando imágenes sintéticas generadas por un modelo de difusión, lo que incrementó el mAP en un 3%. Con los resultados de esta arquitectura mejorada, se creó y calculó un índice de condición del pavimento. Todas las imágenes georreferenciadas, las detecciones, el índice y los datos de tráfico se integraron en una herramienta de software. Este sistema permitió el análisis de dos tramos de carreteras, validando la utilidad del índice propuesto tanto para escenarios que requieren mantenimiento urgente como para aquellos con potencial para aplicar técnicas de mantenimiento preventivo. Además, se mostró el potencial de un modelo de lenguaje visual pre-entrenado, Kosmos-2, que podría añadir valor a la herramienta presentada al determinar el tipo de mantenimiento, costes, entre otros aspectos. En resumen, se ha propuesto un sistema de visión artificial basado en aprendizaje profundo y potenciado con inteligencia artificial generativa, destinado a ayudar a administraciones y empresas de conservación a optimizar la gestión de la red de carreteras.

Las principales limitaciones del estudio radican en el cálculo del índice de condición del pavimento, dado que la severidad debería basarse en parámetros geométricos, como la

profundidad de los baches, y en la falta de experiencia en el procesamiento de lenguaje natural del VLM aplicado al mantenimiento de carreteras. Futuras investigaciones se centrarán en desarrollar técnicas de visión y ópticas para evaluar parámetros geométricos de los defectos, así como en la creación de un conjunto de datos y una arquitectura entrenada para abordar el mantenimiento de pavimentos.

AGRADECIMIENTOS

La investigación es parte del proyecto TED2021-129749B-I00, financiado por MCIN/AEI/10.13039/501100011033 (MAPSIA) y por la Unión Europea “NextGenerationEU”/PRTR, así como por el programa marco de investigación e innovación Horizonte Europa de la Unión Europea en virtud del acuerdo de subvención No. 101103698 (LIAISON).

REFERENCIAS

- Ai, D., Jiang, G., Lam, S.-K., He, P., & Li, C. (2023). Computer vision framework for crack detection of civil infrastructure—A review. *Engineering Applications of Artificial Intelligence*, 117, 105478. <https://doi.org/10.1016/j.engappai.2022.105478>
- Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1), 53. <https://doi.org/10.1186/s40537-021-00444-8>
- Cano-Ortiz, S., Lloret Iglesias, L., Martínez Ruiz del Árbol, P., Lastra-González, P., & Castro-Fresno, D. (2024). An end-to-end computer vision system based on deep learning for pavement distress detection and quantification. *Construction and Building Materials*, 416, 135036. <https://doi.org/10.1016/j.conbuildmat.2024.135036>
- Cano-Ortiz, S., Pascual-Muñoz, P., & Castro-Fresno, D. (2022). Machine learning algorithms for monitoring pavement performance. *Automation in Construction*, 139, 104309. <https://doi.org/10.1016/j.autcon.2022.104309>
- He, K., Zhang, X., Ren, S., & Sun, J. (2014). *Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition*. https://doi.org/10.1007/978-3-319-10578-9_23
- Ho, J., Jain, A., & Abbeel, P. (2020). *Denoising Diffusion Probabilistic Models*.
- Hosang, J., Benenson, R., & Schiele, B. (2017). *Learning non-maximum suppression*.
- Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., & Wei, F. (2023). *Kosmos-2: Grounding Multimodal Large Language Models to the World*.
- Reis, D., Kupec, J., Hong, J., & Daoudi, A. (2023). *Real-Time Flying Object Detection with YOLOv8*.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation*.
- Tsung-Yi Lin, P. D. R. G. K. H. B. H. S. B. (2016). Feature Pyramid Networks for Object Detection. *Computer Vision and Pattern Recognition*.
- Wang, C.-Y., Liao, H.-Y. M., Yeh, I.-H., Wu, Y.-H., Chen, P.-Y., & Hsieh, J.-W. (2019). *CSPNet: A New Backbone that can Enhance Learning Capability of CNN*.
- Zhang, J., Huang, J., Jin, S., & Lu, S. (2023). *Vision-Language Models for Vision Tasks: A Survey*.

- Zhang, K., Zhang, Y., & Cheng, H.-D. (2021). CrackGAN: Pavement Crack Detection Using Partially Accurate Ground Truths Based on Generative Adversarial Learning. *IEEE Transactions on Intelligent Transportation Systems*, 22(2), 1306–1319. <https://doi.org/10.1109/TITS.2020.2990703>
- Zheng Ma, M. L. Y. W. (2019). PAN: Path Integral Based Convolution for Deep Graph Neural Networks. *ICML 2019 Workshop on Learning and Reasoning with Graph-Structured Representations (Oral)*.