ISSN 1870-4069

Detección de deterioros superficiales en pavimentos flexibles basada en segmentación semántica y redes transformer

Mario Alberto Roman-Garay[1], Luis Alberto Morales-Rosales[2], Héctor Rodríguez-Rangel[1], Sofía Isabel Fernández-Gregorio[3]

[1] Tecnólogico Nacional de México Campus Culiacán,
División de Posgrado,
México

[2] Universidad Michoacana de San Nicolás de Hidalgo,
Facultad de Ingeniería Civil,
México

[3] Universidad Veracruzana,
Facultad de Estadística e Informática,
Doctorado en Ciencias de la Computación,
México

{mario.rg, hector.rr}@culiacan.tecnm.mx,
lamorales@conacyt.mx, zS21000291@estudiantes.uv.mx

Resumen. La detección temprana y precisa de deterioros en carreteras es esencial para garantizar la seguridad vial y prevenir accidentes de tráfico. Además, la identificación temprana de agrietamientos y baches ayuda a reducir los costos de mantenimiento de las carreteras a largo plazo, ya que permite realizar reparaciones menores en lugar de costosas renovaciones de carreteras completas. El presente artículo se enfoca en la detección y segmentación automática de agrietamientos y baches en pavimentos por medio de redes Transformer en situaciones no controladas. Para abordar este problema, se utiliza una arquitectura de red llamada Segformer, que se encarga de la extracción de características de las imágenes. Además, se crea un conjunto de datos compuesto por 245 imágenes, en el cual se aplicaron técnicas de procesamiento digital de imágenes y aumento de datos para mejorar la precisión del modelo, por lo que el conjunto de imágenes se extendió a 2052. En el conjunto de datos se contemplaron distintos ambientes obtenidos en puntos situados en la ciudad de Culiacán, México, considerando cambios de iluminación y sombras, lo que permitió evaluar la robustez del modelo obtenido en condiciones similares a las de la vida real donde el ruido ambiental está presente. Nuestro modelo obtuvo métricas de *Precision* de 82.35 %, *F1 Score* de 88.89 % y *Recall* de 96.55 % en la detección de agrietamientos y baches en diferentes condiciones ambientales.

Palabras clave: Segmentación semántica, agrietamientos, baches, pavimento flexible, redes transformer.

Mario Alberto Roman-Garay, Luis Alberto Morales-Rosales, Héctor Rodríguez-Rangel, et al.

Detection of Surface Deterioration in Flexible Pavements based on Semantic Segmentation and Transformer Networks

Abstract. Early and accurate detection of road deterioration is essential to ensure road safety and prevent traffic accidents. In addition, early identification of cracks and potholes helps to reduce long-term road maintenance costs by allowing minor repairs instead of costly complete road renovations. This paper focuses on automatically detecting and segmenting pavement cracks and potholes by Transformer networks in uncontrolled situations. To address this problem, a network architecture called Segformer is used, responsible for feature extraction from the images. In addition, a dataset composed of 245 images was created, extending this dataset to 2052 images with digital image processing and data augmentation techniques to improve the model's accuracy. In the dataset, different environments obtained in points located in the city of Culiacan, Mexico, were contemplated, considering changes in illumination and shadows, allowing us to evaluate the robustness of the model obtained in conditions similar to those of real-life where environmental noise is present. Our model obtained a *Precision* of 82.35%, an *F1 Score* of 88.89%, and a *Recall* of 96.55% in detecting cracks and potholes in different environmental conditions.

Keywords: Semantic segmentation, cracking, potholes, flexible pavement, transformer networks

1. Introducción

Las carreteras son una infraestructura fundamental para el desarrollo económico y social de un país, así como para la movilidad terrestre. Sin embargo, muchas veces sufren deterioros que afectan su funcionamiento y seguridad. Entre los principales deterioros se encuentran los baches, agrietamientos o desprendimientos que incrementan el riesgo de accidentes y daños a vehículos. Para evitar estos daños, cada año se asignan recursos económicos por parte del gobierno para el programa de conservación de caminos. Una actividad importante dentro de la conservación es el monitoreo y evaluación del estado de las carreteras, denominada auscultamiento.

El auscultamiento o evaluación del estado de las carreteras es un conjunto de actividades que tienen como finalidad conservar los caminos en condiciones óptimas y seguras para los usuarios. Estas actividades consisten en obtener información de la calidad superficial mediante elementos tecnológicos como sensores, escáneres o cámaras digitales, instalados en vehículos o dispositivos que circulan por las carreteras.

Asimismo, se efectúan inspecciones visuales por parte de personal calificado para esta tarea. Con la información recopilada se diseñan planes de mantenimiento preventivo y correctivo según el grado de los deterioros identificados. Entre los deterioros que comprometen la seguridad y confort de los usuarios destacan los baches, agrietamientos o desprendimientos.

Aunado a ello se obtienen parámetros como la rugosidad para determinar la calidad del pavimento que permita garantizar la movilidad sin comprometer la seguridad vial. Es por ello, que la tarea de auscultamiento implica una gran inversión de tiempo y recursos humanos.

En los últimos años, se ha avanzado en la automatización de diversas tareas en el área de la ingeniería civil mediante el uso de la inteligencia artificial, y el auscultamiento no es una excepción. Existen numerosos trabajos que intentan realizar esta tarea mediante diferentes técnicas, destacando el empleo de la visión por computadora y el aprendizaje profundo [1].

La mayoría los trabajos se han enfocado en la detección y clasificación de deterioros superficiales como baches y agrietamientos utilizando cámaras digitales para obtener muestras [2]. No obstante, aún persiste un problema frecuente: el ruido ambiental que aparece en las imágenes, como sombras, diversidad de materiales en el pavimento o manchas sobre él.

Esto supone un inconveniente debido a que los algoritmos de aprendizaje profundo aprenden características como bordes, cambios de tonalidades en los píxeles o la textura de las superficies. Dado que las cámaras digitales captan la intensidad de la luz que reflejan las superficies, los cambios de tonalidad de elementos como sombras o manchas son semejantes a los cambios de tonalidad de la superficie cuando se presenta un deterioro como grieta o bache. Esto ocasiona que se presenten falsos positivos en los modelos de detección y clasificación al identificar deterioros cuando no existen.

Para evitar los problemas mencionados anteriormente, se han explorado diversos métodos basados en algoritmos de redes neuronales, tales como las redes convolucionales [3], las redes adversarias generativas (GAN) [4] y en los últimos años las redes transformer [5].

Aunado a la exploración de nuevas técnicas de detección y clasificación de deterioros, un desafío importante en la detección y clasificación de deterioros en carreteras es la falta de un conjunto de datos representativo que incluya distintos escenarios y condiciones de carreteras con y sin deterioros superficiales. Un conjunto de datos robusto hará que los modelos aprendan a reconocer patrones y características en una amplia variedad de escenarios.

El objetivo de un conjunto de datos robusto es contribuir a la mejora de la precisión y la capacidad de los modelos en la detección y clasificación de deterioros ante situaciones no previstas que se encuentran en campo. Además, se requiere distinguir entre el ruido ambiental y los deterioros. Por lo tanto, en este artículo se presenta una propuesta para la detección de grietas y baches en pavimentos flexibles por medio de segmentación semántica y la implementación de una arquitectura de redes transformer.

Se contribuye con un nuevo conjunto de datos que abarca 245 imágenes de deterioros. Este conjunto está compuesto por 130 imágenes de agrietamientos y 115 imágenes de baches donde se presentan distintos escenarios con presencia de sombras, cambios de iluminación, presencia de hojas u objetos varios.

Además, se genera un aumento de datos, obteniendo un total de 2052 imágenes, las cuales fueron preprocesadas mediante los filtros *Contrast Stretching* y conversión a escala de grises. La detección se efectúa empleando la red transformer SegFormer propuesta en [6].

Los resultados obtenidos muestran métricas de *precision* de 82.35 %, *F1 Score* de 88.89 % y *Recall* de 96.55 %. Estos resultados permitirán realizar de manera eficiente la detección de grietas y baches en pavimentos flexibles en ambientes no controlados, lo que demuestra la robustez del conjunto de datos de entrenamiento.

2. Estado del arte

Se han utilizado diferentes técnicas computacionales para abordar el problema de la detección automática de deterioros en pavimentos, como el procesamiento de imágenes por computadora y el aprendizaje automático. A continuación se presentan algunos trabajos que pertenecen a estas dos categorías y que ofrecen soluciones a diversas problemáticas en la detección automática de daños en carretera.

2.1. Procesamiento de imágenes

El procesamiento digital de imágenes ha sido pionero en la detección automática de deterioros en pavimentos. Uno de los primeros enfoques era inferir que los agrietamientos y baches presentaban una tonalidad más oscura con respecto al área sana del pavimento. Tomando esto en cuenta se utilizaron métodos de umbralización como Otsu [7] para segmentar las zonas más oscuras del pavimento y detectar agrietamientos o baches. Sin embargo, este enfoque presentaba detección de falsos positivos cuando existían cambios de tonalidad en el pavimento provocados por manchas, sombras u otros elementos ajenos al pavimento.

Otro enfoque aplicado a la detección de deterioros en pavimentos es el uso de métodos de detección de bordes como Canny o Sobel [8]. Con estos métodos se infiere que los bordes detectados pertenecen a las orillas de las grietas y los baches. Sin embargo, se presenta el mismo problema de falsos positivos en la presencia de entidades externas a la carretera.

En [9, 10] utilizan preprocesamiento de imágenes para mejorar la detección de grietas en pavimentos. El primero utiliza un método basado en la fusión de imágenes multiescala para reducir el ruido y mejorar la calidad de las imágenes antes de aplicar el algoritmo de detección de grietas. En el segundo se emplea un procedimiento de preprocesamiento para eliminar el ruido espurio y rectificar los datos originales del pavimento. Ambos realizan la segmentación semántica de grietas.

2.2. Aprendizaje profundo

Redes convolucionales

Las redes convolucionales han sido las más utilizadas para la detección automática de grietas en imágenes digitales. Estas han sido utilizadas para identificar y clasificar los daños en distintas categorías.

En [3, 11] los autores presentan su propio conjunto de datos con imágenes de distintos tipos de deterioros, como grietas transversales, longitudinales, piel de cocodrilo o baches. Estos conjuntos de datos son utilizados para entrenar redes convolucionales como *SqueezeNet* y *U-net*.

Se han combinado técnicas como el procesamiento de imágenes y las redes convolucionales. En [18] utiliza preprocesamiento para eliminar reflejos de los vidrios y extraer regiones del pavimento debido a la posición de la cámara. Además, utiliza capas de convoluciones para la extracción de características, utilizando un total de 527 imágenes para entrenamiento.

Redes transformer

En los últimos años, las redes transformer han experimentado un gran auge debido al excelente rendimiento que han demostrado en el procesamiento del lenguaje natural. Estas redes han sido adaptadas para el procesamiento de imágenes, así como para tareas de reconocimiento y clasificación de objetos. Arquitecturas como ViT (Vision Transformer) [12], DeiT (Data-efficient Image Transformer) [13] y DETR (Detection TRansformer) han logrado resultados competitivos o incluso superiores a los obtenidos por las redes convolucionales en tareas tanto de reconocimiento como en la clasificación de objetos.

Desde hace muchos años la detección automática de deterioros en pavimentos es un tema que se ha desarrollado y se ha adaptado a las nuevas tecnologías. Se ha visto beneficiada tanto con dispositivos nuevos para la recolección de información como con algoritmos computacionales que entrenan modelos de inteligencia artificial y aprendizaje profundo que identifican automáticamente grietas. Sin embargo, aún existen algunas problemáticas abiertas que deben ser solucionadas, tales como la detección de falsos positivos ante la presencia de ruido ambiental o la evaluación de distintos tipos de deterioros.

3. Materiales y métodos

En esta sección se describe el diseño de la toma de muestras de imágenes, de grietas y baches en pavimentos flexibles, y la implementación de una red Transformer (Segformer) entrenada para la segmentación semántica de grietas y baches. Para el proceso de adquisición de imágenes se empleó una cámara de acción compacta GoPro 8 Black. La elección de la GoPro 8 Black se debe a la alta calidad de imagen, así como a su capacidad para grabar en alta definición, lo que la hace ideal para capturar imágenes de superficies de pavimento de forma rápida y sencilla.

A continuación, se describe la instalación y el montaje de la cámara para la captura de imágenes. Además, se detalla la configuración del hardware utilizado, la especificación técnica de la cámara y la computadora empleada para el proceso de entrenamiento, así como el software utilizado para el procesamiento de las imágenes.

3.1. Hardware

El uso de distintos dispositivos es indispensable para llevar a cabo las tareas de auscultación con precisión. Para la toma de muestras se utilizó una cámara de acción GoPro 8 black con una resolución de 4000x3000 píxeles, la cual se colocó a 1.20 metros de altura con apoyo de un tripie. Además, se utilizó un equipo de cómputo con las características que se muestran en la Tabla 1.

Tabla 1. Características del servidor.

Nombre	Características
Sistema Operativo	Ubuntu 20.04
RAM	16 GB
Almacenamiento	500 GB SSD
Procesador	Ryzen 7 3700
Tarjeta Gráfica	RTX 3060 Ti

3.2. Metodología

La metodología que se ha utilizado en este trabajo se basa en los pasos genéricos para la obtención de modelos de aprendizaje profundo presentada por [14] y [2]. Sin embargo, se ha añadido una etapa adicional de pre-preprocesamiento para mejorar la calidad de los datos de entrada, así como una etapa de pruebas de los resultados.

La Figura 1 muestra el proceso de la metodología utilizada durante este trabajo. Comienza con la recolección de datos, donde se recopilan imágenes relevantes que muestren el objeto en diferentes contextos.

Luego sigue el preprocesamiento, donde se considera si se aplican o no distintos filtros para resaltar características como bordes, texturas y tonalidades de colores. También se busca reducir el ruido presente en los datos. Una vez definidas las imágenes, se construye el conjunto de datos necesario para entrenar el modelo clasificador.

Dependiendo de la arquitectura de redes neuronales utilizada se determina si es o no necesaria una etapa de etiquetado de imágenes. Tras el preprocesamiento y la construcción del conjunto de datos, se procede al entrenamiento de las redes neuronales y una validación de los modelos resultantes. Esta validación proporciona métricas para evaluar el rendimiento del modelo final, las cuales son *Mean IOU, Presicion, Recall y F1 Score*.

3.3. Recolección de datos

Con el propósito de obtener imágenes de deterioros superficiales en pavimentos, se realizaron recorridos por diversas vías de la ciudad de Culiacán, Sinaloa. Se enfatizó en el registro de pavimentos flexibles.

Para la captura de las imágenes se empleó una cámara de acción GoPro 8 black y un tripié con una inclinación de 20 grados respecto a la superficie del pavimento y a una altura de 1.20 metros, de tal manera que la imagen capturada mostrara el ancho total del carril, el cual tiene una anchura de tres metros en promedio.

3.4. Anotación de imágenes

El etiquetado o anotación de imágenes para el entrenamiento de redes neuronales para segmentación, es un proceso donde se indica el área que representa un objeto en una imagen [15]. Esta anotación se lleva a cabo a nivel de píxel, y existen herramientas con las cuales esto es posible.

En este trabajo se llevó a cabo la anotación con Roboflow [16], esta herramienta permite a los usuarios realizar anotaciones de distintos tipos, aplicar técnicas de

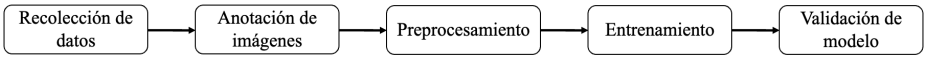

Fig. 1. Metodología implementada.

pre-procesamiento y hacer aumento de datos para crear distintas versiones del conjunto de datos.

La anotación utilizada fue la segmentación semántica, en la cual se debe señalar el contorno de los objetos de interés. En el caso de las grietas y baches se indicaron el contorno en todas las imágenes y se indicaba que el área pertenecía a alguno de los dos deterioros. De esta forma, al momento de descargar el conjunto de datos, la herramienta Roboflow nos proporciona la imagen original y una máscara, la cual es una imagen que contiene el área señalada durante la anotación.

En la segmentación semántica, la forma en que se diferencia un objeto de otro, está dado por el valor de los píxeles del área que se señala durante las anotaciones. Estos valores van desde 0 hasta 255, que son los posibles valores de un píxel en una imagen en escala de grises.

En la imagen mascara, los píxeles que representan el área de una grieta tienen un valor de uno y los que representan el área de un bache tienen el valor dos. Debido a que estos valores son muy cercanos a cero, el cual es la representación de un píxel totalmente negro, por lo que la mascara se aprecia como una imagen completamente oscura.

3.5. Pre-procesamiento

Actualmente, los modelos de aprendizaje profundo tienen la capacidad de ser entrenados sin necesidad de tener un preprocesamiento de datos; sin embargo, añadir esta etapa antes del entrenamiento mejora la calidad de los datos y aumenta la precisión del modelo [17].

Las imágenes crudas llegan a contener ruido como sombras o iluminación inconsistente que interfieren con la capacidad del modelo para identificar y segmentar los objetos de interés. Al aplicar técnicas de preprocesamiento como el estiramiento por contraste o el cambio a escala de grises, se mejora la calidad de las imágenes y facilita la identificación de los objetos de interés.

Estiramiento por contraste

El *Contrast Stretching* (estiramiento de contraste) es una técnica de procesamiento de imágenes que se utiliza para mejorar la calidad visual de las imágenes. Consiste en expandir el rango de valores de píxeles de una imagen de manera que los valores más bajos se mapeen a un valor mínimo y los valores más altos se mapeen a un valor máximo.

De esta manera, se amplía el rango de valores de la imagen para obtener una imagen con mayor contraste y detalle. El *Contrast Stretching* es una técnica común utilizada en el preprocesamiento de imágenes antes de aplicar algoritmos de segmentación o clasificación en tareas de procesamiento de imágenes. Con esta técnica se busca reducir la afección de las sombras y los cambios de iluminación en las imágenes.

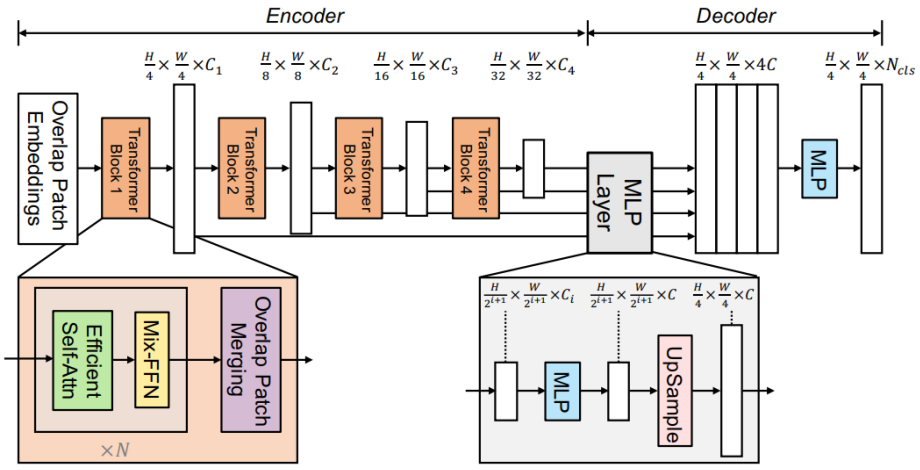

Fig. 2. Arquitectura de red Segformer [6].

Escala de grises

Es importante convertir las imágenes a escala de grises para mejorar la eficiencia del procesamiento de la información y reducir la cantidad de datos que se deben procesar en la red neuronal.

Al eliminar la información de color, se reduce el tamaño de la imagen y se elimina la redundancia de datos, lo que acelera el entrenamiento de la red y reduce la complejidad del modelo.

Redimensión de imágenes

El proceso de redimensión de imágenes cambia el tamaño y la resolución de las muestras para adaptarlas a un formato que permita ser introducidas a las redes neuronales. El tamaño original de las imágenes era de 4000x3000 píxeles, pero para introducirlas en las redes se redimensionaron a 512x512 píxeles.

Esta reducción se debe a que la red transformer (SegFormer) empleada tiene este requerimiento para el tamaño de imagen como dato de entrada, ya que cada píxel es introducido a una neurona como parte de la red.

Aumento de datos

Para aumentar la variedad y el número de muestras disponibles para el entrenamiento del modelo clasificador, se aplicaron técnicas de aumento de datos conocidas como *Tiling* y rotación.

– **Rotación:** Esta consiste en girar las imágenes originales un cierto ángulo alrededor de su centro. En este caso, se rotaron las imágenes 90 grados en sentido horario y se repitió el proceso hasta completar un giro completo de 270 grados. De esta manera, se obtuvieron tres imágenes rotadas por cada imagen original.

Tabla 2. Cantidad de imágenes obtenidas durante la toma de muestras.

Clase	Total
Agrietamiento	130
Baches	115
Total	**245**

– ***Tiling:*** consiste en dividir una imagen en varias partes más pequeñas y utilizar cada parte como una imagen independiente en el conjunto de datos. Esto se hace para aumentar la cantidad de datos de entrenamiento y evitar problemas de memoria cuando las imágenes son demasiado grandes para ser procesadas por la red neuronal en su totalidad.

3.6. Entrenamiento

En esta etapa, se utilizó la red SegFormer [6], la cual se basa en la arquitectura Transformer.

Esta red tiene dos características principales: el uso de un codificador para extraer características finas y gruesas de las imágenes, y un decodificador de perceptrones multicapa para fusionar estas características y predecir la máscara de segmentación. En la Figura 2 se muestra la arquitectura de la Red Segformer donde se observa gráficamente como están estructurados cada uno de los módulos que se utilizan, así como el codificador y el decodificador.

3.7. Validación de modelo

A continuación se muestran las métricas que se utilizan para llevar a cabo la evaluación del modelo de segmentación.

Mean Intersection Over Union

El mIoU (Intersección sobre la Unión Promedio) es una medida de la superposición entre la máscara de segmentación (también conocida como máscara de verdad terrenal) y la máscara de predicción producida por el modelo de segmentación. Se calcula como la relación entre el área de la intersección entre estas dos máscaras y el área de la unión entre ellas, para cada clase de objeto. Luego, se calcula el promedio de las mIoU de todas las clases para obtener una medida general del rendimiento del modelo. Para el cálculo de mIoU se consideran la ecuaciones 1 y 2 descritas a continuación:

$$IoU_c = \frac{TP_c}{TP_c + FP_c + FN_c}, \tag{1}$$

$$meanIoU = \frac{1}{c}\sum_{c} IoU_c. \tag{2}$$

TP= *Verdaderos positivos*, FP=*Falsos positivos*, FN= *Falsos negativos*

Precision, Recall, F1 Score

Las métricas *Precision, Recall y F1 Score* se obtuvieron al comparar la información real de las anotaciones realizadas con las predicciones generadas por el modelo entrenado.

La métrica *presicion* o precisión, en español, se define como la proporción de verdaderos positivos (TP, por sus siglas en inglés), es decir, los casos positivos correctamente clasificados, sobre el total de casos clasificados como positivos. En otras palabras, la precisión mide la capacidad del modelo para identificar correctamente los casos positivos. La Ecuación 3 muestra la fórmula de la precisión:

$$Precision = \frac{TP}{TP + FP}. \tag{3}$$

La recuperación, también conocida como *recall*, se define como la proporción de verdaderos positivos sobre el total de casos positivos (TP + Falsos negativos). En otras palabras, la recuperación mide la capacidad del modelo para identificar todos los casos positivos. La Ecuación 4 muestra la fórmula de la Recuperación (Recall):

$$Recall = \frac{TP}{TP + FN}. \tag{4}$$

La medida *F1 Score* es una combinación de precisión y recuperación, y se utiliza como una medida general de rendimiento del modelo. Se calcula como el promedio armónico de precisión y recuperación. La Ecuación 5 muestra la fórmula de la medida *F1 Score*:

$$F1Score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{5}$$

4. Resultados

Durante la etapa de recolección de datos se llevaron a cabo distintos recorridos para obtener muestras de deterioros superficiales en pavimentos. Se recorrieron las calles Cerro Monte Largo, y el bulevar Luciernilla, además del circuito interior del Tecnológico Nacional de México Campus Culiacán, todo esto en la ciudad de Culiacán, Sinaloa. A continuación se muestra el recorrido total que se realizó en las calles mencionadas:

- Calle Juan de Dios Batiz con un total de 832 metros,
- Calle Cerro Monte Largo con un total de 903 metros,
- Calle Luciernilla con un total de 1650 metros.

En la Tabla 2 se muestran los resultados de las imágenes que se obtuvieron durante los recorridos realizados. En ella se observa que en total se lograron recolectar 245 imágenes.

Posteriormente, se llevó a cabo el proceso de anotación de imágenes. La Figura 3 muestra un conjunto de imágenes normales de pavimentos junto a sus respectivas máscaras de segmentación. En las máscaras se han delimitado con precisión las áreas

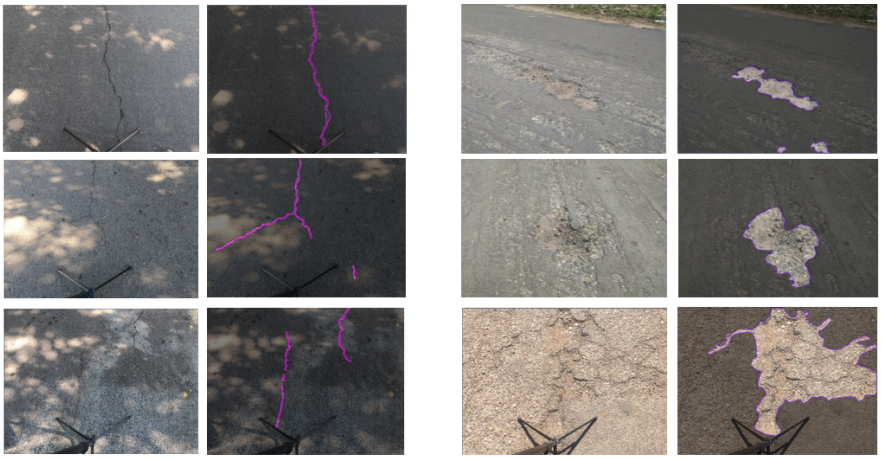

Fig. 3. a) Imagen normal. b) Imagen con anotación a mano.

que presentan agrietamientos y baches, utilizando técnicas manuales de segmentación. Estas máscaras de segmentación sirven de referencia durante el entrenamiento para indicar el área de interés que se requiere analizar.

Los ejemplos presentados son solo una muestra de las 245 imágenes que se etiquetaron manualmente durante el proceso de anotación. Es importante mencionar que este proceso es el que consume más tiempo. La anotación de cada imagen toma alrededor de 10 a 15 minutos, dependiendo de la complejidad de los deterioros. Por lo tanto, el tiempo total necesario para completar estas anotaciones fue de aproximadamente 60 horas.

Resultados de preprocesamiento

Después de realizar las anotaciones en las imágenes, se aplican técnicas de preprocesamiento para resaltar ciertas características, como los bordes, o para reducir el impacto de la iluminación o las sombras en las imágenes.

En este caso, además de las técnicas mencionadas anteriormente, se aplicaron técnicas como *Contrast Stretching* y conversión a escala de grises. También se realizaron cambios en las dimensiones de las imágenes.

Una vez aplicadas las técnicas de Tiling y rotación de imágenes, se produjo un aumento significativo en la cantidad de imágenes. De las 245 imágenes originales, se obtuvo un conjunto de datos de, 2940 imágenes. Sin embargo, debido a una restricción en la versión gratuita de Roboflow en cuanto a la cantidad de imágenes que se pueden descargar, el conjunto de datos final resultó en, 2052 imágenes.

Se realizó un recorte en la cantidad de imágenes totales en los conjuntos de validación y pruebas para lograr esto. La distribución de imágenes en los conjuntos de entrenamiento, validación y pruebas fue realizada de la siguiente manera 1780, 184 y 88 respectivamente.

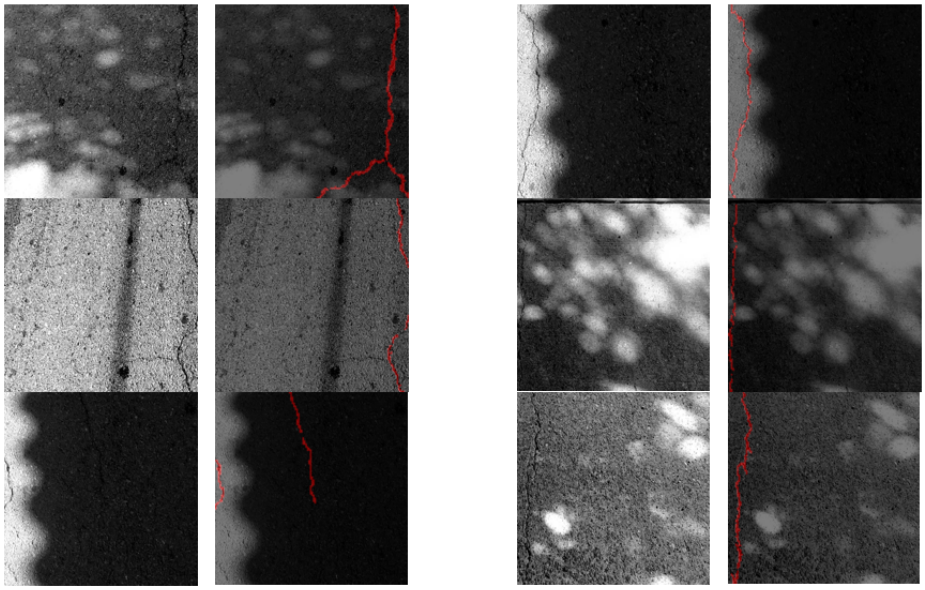

Fig. 4. a) Imagen normal. b) Mascara de predicción.

Resultados del entrenamiento del modelo de segmentación

En la Figura 4 se muestran algunos resultados obtenidos con el conjunto de pruebas una vez realizado el entrenamiento del modelo de predicción. Se observa que esté es capaz de identificar correctamente los agrietamientos que aparecen en el pavimento, incluso con la presencia de cambios de iluminación y sombras.

En la Tabla 3 se muestran los resultados de las métricas que se utilizaron para evaluar el entrenamiento del modelo de segmentación, así como una comparación con trabajos de otros autores.

Los resultados mostrados de nuestra solución en la Tabla 3 son las métricas obtenidas en el conjunto de pruebas. En la Tabla 3 los valores de nuestras métricas de evaluación son superiores al 80 %, excepto Mean IoU. Se obtuvo un *Recall* del 96.55 %, lo que indica que el modelo detecta correctamente la mayoría de las grietas y baches con una cantidad pequeña de falsos positivos, superando a los otros modelos presentados en la tabla comparativa.

En general, los resultados obtenidos son prometedores y sugieren que el modelo de detección de objetos es eficaz para la tarea que fue entrenado en presencia de ruido ambiental, cambios de iluminación y sombras, como se observa en la Figura 4.

No obstante, es importante tener en cuenta que siempre hay margen de mejora en cualquier modelo de aprendizaje automático y se pueden explorar técnicas adicionales para mejorar aún más la precisión y el rendimiento del modelo en futuros trabajos.

Tabla 3. Comparativa de las métricas de evaluación con el estado del arte.

Autor	Técnica	Imágenes	*Mean IOU*	*Presicion*	*Recall*	*F1 Score*
Li et al. [9]	Procesamiento de imágenes	No requiere Entrenamiento	No indica	88.38 %	93.15 %	90.68 %
Li et al. [10]	Procesamiento de imágenes	200 de prueba	No indica	89.90 %	89.47 %	88.04 %
Bang et al.[14]	Redes Convolucionales	527	No indica	93.57 %	84.90 %	89.03 %
Solución propuesta	Redes transformer	2940	67.00 %	82.35 %	96.55 %	88.89 %

5. Conclusiones

En este trabajo se ha presentado una propuesta para la detección y segmentación semántica de deterioros superficiales en pavimentos mediante el uso de técnicas de visión artificial y aprendizaje profundo. Además, se presentó la construcción de un conjunto de datos de deterioros superficiales de 245 imágenes, a las cuales se les aplicaron técnicas de aumento de datos de Tiling y rotación, obteniendo un total de 2052 imágenes.

A pesar de no contar con una gran cantidad de imágenes, se logró llevar a cabo la segmentación semántica que permite identificar la presencia de baches y grietas, incluso en presencia de ruido como sombras, cambios de iluminación, manchas en el pavimento y objetos ajenos al pavimento.

Esto demuestra que el modelo entrenado tiene la capacidad de detectar deterioros en ambientes no controlados y diferentes condiciones. Los factores que contribuyeron a ello fue la inclusión de técnicas de aumento de datos, preprocesamiento, la inclusión de datos con distinto ruido externo y la arquitectura de la red Segformer.

Es importante considerar la calidad de la segmentación semántica que se lleva a cabo manualmente al construir el conjunto de datos. Esta calidad es crucial para el desempeño del modelo, ya que las delimitaciones precisas del área de interés permiten que la red neuronal se enfoque en las características que se desean identificar.

En trabajos futuros, se aumentará la cantidad de imágenes anotadas para llevar a cabo una clasificación de los diferentes tipos de agrietamientos y baches. Además, se obtendrán medidas de longitud, anchura y profundidad para evaluar la severidad del daño que representa cada deterioro para la carretera, con el fin de desarrollar un algoritmo que permita la evaluación de los deterioros de manera automática.

Referencias

1. Du, Z., Yuan, J., Xiao, F., Hettiarachchi, C.: Application of image technology on pavement distress detection: A review. Measurement, vol. 184, no. 109900 (2021) doi: 10.1016/j.measurement.2021.109900
2. Hou, Y., Li, Q., Zhang, C., Lu, G., Ye, Z., Chen, Y., Wang, L., Cao, D.: The state-of-the-art review on applications of intrusive sensing, image processing techniques, and machine learning methods in pavement monitoring and analysis. Engineering, vol. 7, no. 6, pp. 845–856 (2021) doi: 10.1016/j.eng.2020.07.030

3. Ha, J., Kim, D., Kim, M.: Assessing severity of road cracks using deep learning-based segmentation and detection. The Journal of Supercomputing, vol. 78, pp. 17721–17735 (2022) doi: 10.1007/s11227-022-04560-x
4. Zhang, K., Zhang, Y., Cheng, H. D.: CrackGAN: Pavement crack detection using partially accurate ground truths based on generative adversarial learning. In: IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 2, pp. 1306–1319 (2021) doi: 10.1109/TITS.2020.2990703
5. Xiao, S., Shang, K., Lin, K., Wu, Q., Gu, H., Zhang, Z.: Pavement crack detection with hybrid-window attentive vision transformers. International Journal of Applied Earth Observation and Geoinformation, vol. 116 (2023) doi: 10.1016/j.jag.2022.103172
6. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P.: SegFormer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems, vol. 34, pp. 12077–12090 (2021)
7. Ha, J., Kim, D., Kim, M.: Assessing severity of road cracks using deep learning-based segmentation and detection. The Journal of Supercomputing, vol. 78, pp. 17721-–17735 (2022) doi: 10.1007/s11227-022-04560-x
8. Ahmed-Talab, A. M., Huang, Z., Xi, F., HaiMing, L.: Detection crack in image using Otsu method and multiple filtering in image processing techniques. Optik, vol. 127, no. 3, pp. 1030–1033 (2016) doi: 10.1016/j.ijleo.2015.09.147
9. Li, H., Song, D., Liu, Y., Li, B.: Automatic pavement crack detection by multi-scale image fusion. IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 6, pp. 2025–2036 (2019) doi: 10.1109/TITS.2018.2856928
10. Li, B., Wang, K. C., Zhang, A., Fei, Y., Sollazzo, G.: Automatic segmentation and enhancement of pavement cracks based on 3D pavement images. Journal of Advanced Transportation, vol. 2019 (2019) doi: 10.1155/2019/1813763
11. Eisenbach, M., Stricker, R., Seichter, D., Amende, K., Debes, K., Sesselmann, M., Ebersbach, D., Stoeckert, U., Gross, H. M.: How to get pavement distress detection ready for deep learning? A systematic approach. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 2039–2047 (2017) doi: 10.1109/IJCNN.2017.7966101
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021) doi: 10.48550/arXiv.2010.11929
13. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers and distillation through attention. In: International conference on machine learning, vol. 139, pp. 10347–10357 (2021)
14. Hamishebahar, Y., Guan, H., So, S., Jo, J.: A comprehensive review of deep learning-based crack detection approaches. Applied Sciences, vol. 12, no. 3 (2022) doi: 10.3390/app12031374
15. Guo, Y., Liu, Y., Georgiou, T., Lew, M. S.: A review of semantic segmentation using deep neural networks. International Journal of Multimedia Information Retrieval, vol. 7, pp. 87–93 (2018) doi: 10.1007/s13735-017-0141-z
16. Roboflow: Roboflow universe: Open source computer vision community (2023) https://universe.roboflow.com/
17. Ranganathan, G.: A study to find facts behind preprocessing on deep learning algorithms. Journal of Innovative Image Processing (JIIP), vol. 3, no. 1, pp. 66–74 (2021) doi: 10.36548/jiip.2021.1.006
18. Bang, S., Park, S., Kim, H., Kim, H.: Encoder–decoder network for pixel-level road crack detection in black-box images. Computer-Aided Civil and Infrastructure Engineering, vol. 34, no. 8, pp. 713–727 (2019) doi: 10.1111/mice.12440