Data in Brief 63 (2025) 112060

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

PY-CrackDB: A pavement crack dataset from paraguayan roads for context-aware computer vision models

Fredy Gabriel Ramírez-Villanueva[a,b,*], José Luis Vázquez Noguera[a], Horacio Legal-Ayala[a], Julio César Mello-Román[a], Pastor Enmanuel Pérez-Estigarribia[a]

[a] *Facultad Politécnica, Universidad Nacional de Asunción, San Lorenzo, Paraguay*
[b] *Facultad de Ciencias y Tecnologías, Universidad Nacional de Caaguazú (FCyT-UNCA), Coronel Oviedo, Paraguay*

A R T I C L E   I N F O

Article history:
Received 6 August 2025
Revised 4 September 2025
Accepted 8 September 2025
Available online 12 September 2025

Dataset link: PY-CrackDB: A Pavement Crack Dataset from Paraguayan Roads for Context-Aware Computer Vision Models (Original data)

Keywords:
Asphalt pavement
Crack detection
Deep learning
Image segmentation
Object detection
Paraguayan infrastructure
Road maintenance

A B S T R A C T

PY-CrackDB, a novel dataset of asphalt pavement images designed for developing context-aware artificial intelligence systems. The dataset contains 569 images (351 × 500 pixels), collected from national routes near Coronel Oviedo, Paraguay, and divided into 369 images with cracks and 200 without. A primary contribution of this work is its specific focus on fine fissures (< 3 mm wide), a category critical for early-stage maintenance according to Paraguayan road engineering standards. Data collection was performed under standardized conditions, and all annotations were created by civil engineering professionals and subsequently verified through a rigorous cross-review protocol to ensure accuracy. This methodological rigor resulted in a dataset that is particularly suitable for training and validating models for semantic segmentation and early defect detection, ultimately sup-

* Corresponding author.
E-mail addresses: framirez@fctunca.edu.py (F.G. Ramírez-Villanueva), jlvazquez@pol.una.py (J.L. Vázquez Noguera), hlegal@pol.una.py (H. Legal-Ayala), juliomello@pol.una.py (J.C. Mello-Román), peperez.estigarribia@pol.una.py (P.E. Pérez-Estigarribia).
Social media: @fgrv (F.G. Ramírez-Villanueva)

https://doi.org/10.1016/j.dib.2025.112060

porting the development of preventative road maintenance strategies.

Specifications Table

Subject	Computer Sciences Engineering & Materials science
Specific subject area	Computer Vision; Deep Learning; Pavement Health Monitoring; Automated Infrastructure Inspection.
Type of data	Raw images (2D-RGB) (.jpg) Ground truth images (2D-Binary images) (.png)
Data collection	Images were captured on national routes near Coronel Oviedo, Paraguay, using Motorola moto g52 and Tecno Spark 10 pro smartphones. A ring light and ruler were used to standardize lighting and scale at a consistent height of 1.10–1.20 m. Images were manually annotated for classification (“with crack” / “without crack”) and segmentation (pixel-wise binary masks) using Fiji software, a Wacom tablet, and custom Python scripts. A rigorous cross-review process was implemented to ensure annotation quality
Data source location	National routes near Coronel Oviedo, Caaguazú Department, Paraguay. The data were stored in Facultad de Ciencias y Tecnologías, Universidad Nacional de Caaguazú.
Data accessibility	Repository name: Zenodo Data identification number: 10.5281/zenodo.16749554 Direct URL to data: https://doi.org/10.5281/zenodo.16749554
Related research article	None

1. Value of the Data

- **High relevance for local contexts**: This dataset was developed as one of the first publicly available resources specifically capturing the characteristics of Paraguayan asphalt pavements, making it invaluable for training locally adapted AI models.
- **Focus on early-stage defect detection**: The dataset was designed to emphasizes fine fissures (< 3 mm wide), a critical category defined by Paraguay’s National Road Manual [1]. This is a key differentiator from many existing datasets (e.g., Crack500 [2]) that often contain more pronounced cracks.
- **Versatility for multiple computer vision tasks**: The dataset has been annotated for both image classification and semantic segmentation, which allows for direct, comparative benchmarking of different methodologies.
- **Standardized acquisition and rigorous quality control**: The standardized data collection process and a rigorous annotation quality check enhanced the reliability of the dataset for training robust models.

2. Background

The efficient and precise inspection of asphalt pavements is critical for maintaining safe road infrastructure. Traditional visual inspection methods are often subjective, costly, and can pose safety risks to personnel [3]. Automated systems that leverage computer vision and deep learning offer a powerful solution for objective and early damage detection. The development of such systems, however, is fundamentally dependent on high-quality, accurately labeled datasets.

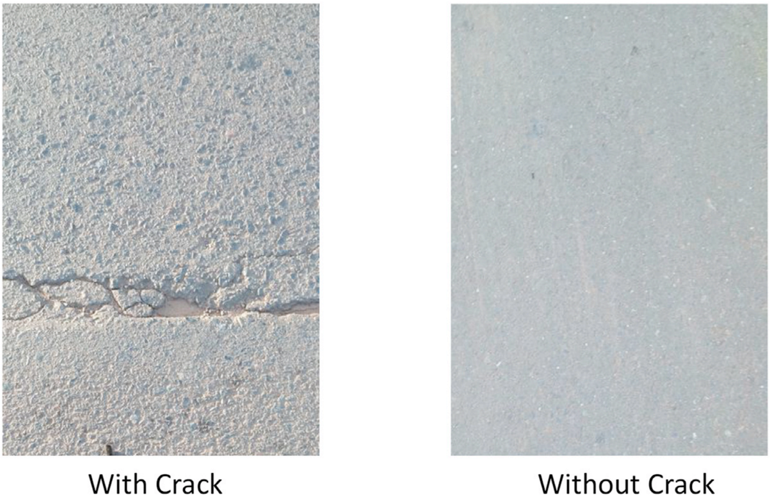

Fig. 1. Examples of images with and without crack.

While several public datasets exist, a preliminary review revealed significant limitations for our specific context. Many lack a sufficient representation of fine fissures (< 3 mm), which are of primary interest for early detection as defined by the Pavement Manual of Paraguay [1]. Furthermore, known labeling errors in some widely used datasets [2] compromise their reliability for training robust models. These shortcomings were the primary motivation for creating PY-CrackDB within the framework of the research project "Asphaltic Pavement Crack Segmentation using Computer Vision and Artificial Intelligence" to provide a specialized, high-quality resource tailored to Paraguayan roads.

3. Data Description

The PY-CrackDB dataset consists of **569** images standardized to a resolution of 351 × 500 pixels[4]. It is categorized into images containing cracks and those without, with a distribution of 369 "With crack" images (64.8 %) and 200 "Without crack" images (35.2 %), as illustrated in Fig. 1. This composition was designed to be representative of real-world inspection scenarios and as shown in the comparative analysis in Table 1, provides a suitable basis for training supervised learning models. To ensure model robustness, the dataset deliberately includes images with real-world complexities, such as the varied lighting conditions and pavement markings showcased in Fig. 3.

Table 1
Datasets comparisons.

Dataset	Number of images	Resolution	Environmental effect	Geotag	Pixel-wise annotations	Other annotations
CFD [5]	118	480 × 320	YES	NO	YES	NO
Aigle-RN [6]	38	991 × 462 / 311 × 462	NO	NO	YES	NO
Crack500 [7]	500	~2560 × 1440	NO	Partially	YES	NO
Cracktree200 [8]	206	800 × 600	YES	NO	YES	NO
Sut-crack [9]	130	3024 × 4032	YES	YES	YES	YES
PY-CrackDB	569	351 × 500	YES	NO*	YES	YES**

* Dataset sourced exclusively from Paraguayan roads to ensure local context and relevance.
** Classification.

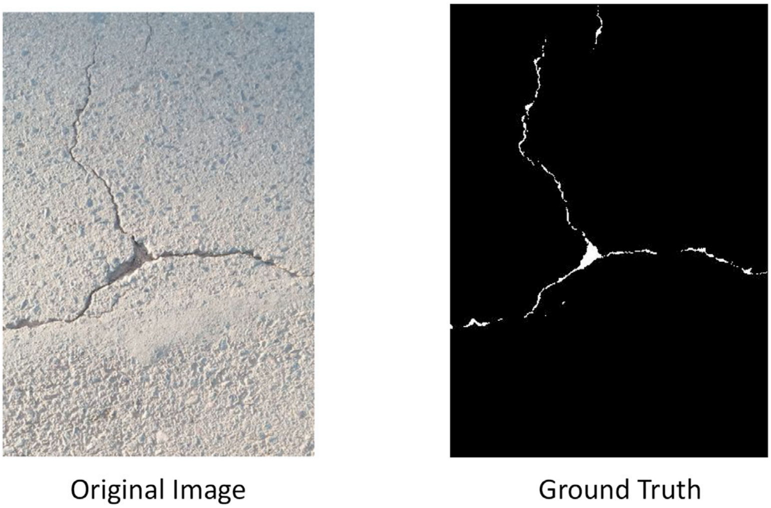

Fig. 2. Example of images for segmentation.

The dataset is organized into two main directories corresponding to the supported computer vision tasks:

1. Segmentation/: Contains data for semantic segmentation tasks. An example of an original image and its corresponding ground truth mask is shown in Fig. 2. The directory is divided into two sub-folders:
   - Original_image/: Contains 369 original RGB images of asphalt pavement with cracks.
   - Ground_truth/: Contains 369 corresponding binary mask images, where white pixels represent cracks and black pixels represent the background.
2. Classification/: This directory is structured for image classification tasks and contains two sub-folders:
   - With_crack/: Contains all 369 images that show pavement cracks.
   - Without_crack/: Contains 200 images of pavement sections with no visible cracks.

4. Experimental Design, Materials and Methods

4.1. Image collection and location

The 569 images in this dataset were captured between March and May 2024 near Coronel Oviedo, Paraguay (approximate UTM coordinates: 7185,311 N, 556,195 E, Zone 21 J). This location was strategically chosen because it hosts the intersection of Paraguay's most critical logistics corridors: Route PY02 (the primary east-west artery) and Route PY08 (a key north-south axis). This confluence provides a representative mix of pavement conditions under significant operational stress. Pavement sections were selected to capture a diversity of fissure types, textures, and lighting scenarios, with all images standardized to a 351 $\times$ 500 pixel resolution to balance feature detail with computational efficiency.

Pavements in this region endure compounding stress from environmental, operational, and historical factors. The area's Humid Subtropical (Cfa) climate imposes severe environmental loads, including high temperatures (25 °C to 33 °C) and substantial year-round precipitation exceeding 1600 mm [10]. Operationally, both PY02 and PY08 are Primary National Routes engineered for a maximum gross vehicle weight of 45 tons on six-axle vehicles. While historical traffic data for Route PY02 recorded an AADT of 2318 vehicles [11], recent estimates approach 20,000 users daily [12]. Although specific data for PY08 is unavailable, both corridors carry substantial heavy commercial traffic. This is compounded by the routes' histories: while PY02 has

Fig. 3. Asphalt pavement shown under varied conditions of sunlight and shadows, with some images also including pavement markings.

been undergoing modernization since 2019 [12], many images were captured on the deteriorating original alignment pending separate rehabilitation [13]. In contrast, Route PY08 has a non-uniform structure from being paved in stages over many years [14] and is maintained via medium-term CREMA contracts [15].

4.2. Capture devices and standardization tools

Two smartphone cameras were used for data acquisition: Motorola moto g52 and a Tecno Mobile Tecno Spark 10 pro.

To ensure consistency, two auxiliary tools were employed: (1) LUO ring light with an integrated phone stand for uniform illumination, and (2) a 55 cm multifunctional numeric ruler as a scale reference. The capture height was kept constant between 1.10 m and 1.20 m.

4.3. Annotation and quality control

The manual annotation process was executed as follows:

- **Segmentation:** Pixel-wise segmentation was performed by a Civil Engineering professional using Fiji (v1.54f) software and a Wacom Intuos BT S drawing tablet for precise manual tracing of cracks.
- **Classification:** Each processed image was manually inspected to determine the presence or absence of pavement fissures and sorted into "With Crack" or "Without Crack" directories.
- **Quality Control:** A strict quality control protocol was implemented to ensure annotation consistency and accuracy. The initial segmentation masks were independently verified by the first author, who holds a master's degree in civil engineering sciences, through a cross-review process. The accuracy of the final masks was confirmed by visually superimposing them onto the original images, as illustrated in Fig. 4.

Limitations

Geographic Specificity: The dataset's focus on a single Paraguayan region, while a contextual strength, means models trained on this data may not directly generalize to areas with different asphalt compositions or climatic wear patterns without further validation.

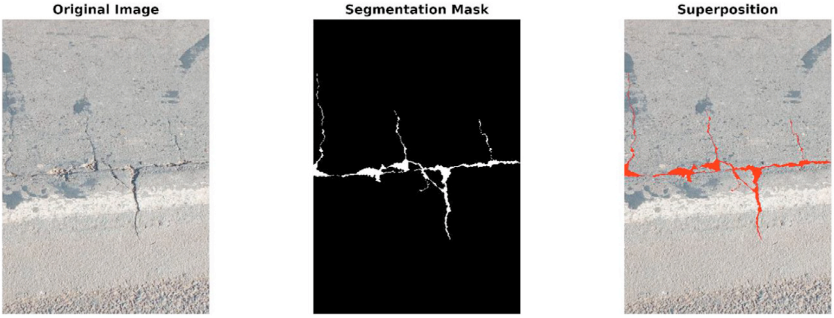

Fig. 4. Superposition for quality control.

Sensor Variability: The dataset was collected using two distinct smartphone models to balance data consistency with a baseline of sensor diversity. Although this provides more variability than is common in single-sensor benchmarks, models intended for broad deployment would benefit from training on a wider range of capture devices. Future work will aim to expand the dataset with additional sensors to enhance model generalization.

Pavement History and Technical Data: While this study provides context on the general age and maintenance strategies for Routes PY02 and PY08, a limitation is the unavailability of specific technical data. Detailed records, including the original asphalt mix designs, granular maintenance history, and precise construction dates for older pavement sections, are not publicly accessible, which precludes a more granular analysis of how these specific factors influence fissure formation.

Ethics Statement

The authors have read and follow the ethical requirements for publication in Data in Brief and confirm that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

Data Availability

PY-CrackDB: A Pavement Crack Dataset from Paraguayan Roads for Context-Aware Computer Vision Models (Original data) (Zenodo)

CRediT Author Statement

Fredy Gabriel Ramírez-Villanueva: Conceptualization, Investigation, Writing – original draft, Funding acquisition, Data curation; **José Luis Vázquez Noguera:** Conceptualization, Supervision, Writing – review & editing; **Horacio Legal-Ayala:** Methodology, Validation, Writing – review & editing; **Julio César Mello-Román:** Visualization, Formal analysis, Writing – review & editing; **Pastor Enmanuel Pérez-Estigarribia:** Visualization, Data curation.

Acknowledgements

We would like to extend our sincere gratitude to Engineer Luz Rashel Soto Olguín for her valuable collaboration and expertise in the manual segmentation process of this work.

This research was funded by the CONACYT, Paraguay, under grant INIC01-302, through the PROCIENCIA Program with resources from the Fund for Excellence in Education and Research (FEEI).

During the preparation of this work, the authors used Google Gemini to improve readability and language. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Paraguayan Highway Association (APC)Road Manual of Paraguay, Asunción, Paraguay, 2019 https://mopc.gov.py/servicios/manual-de-carreteras/ accessed 5 August 2025.
[2] F. Yang, L. Zhang, S. Yu, D. Prokhorov, D. Zhang, Y.J. Zhu, Feature pyramid and hierarchical boosting network for pavement crack detection, IEEE Transac. Intell. Transport. Syst. 21 (4) (2020) 1525–1535, doi:10.1109/TITS.2019.2910597.
[3] Y. Hamishebahar, Z. Guan, T.D. Le, A. Manaf, A comprehensive review of deep learning-based crack detection approaches, Appl. Scienc. 12 (3) (2022) 1374, doi:10.3390/app12031374.
[4] F.G. Ramírez-Villanueva, J.L. Vázquez Noguera, H. Legal-Ayala, J.C. Mello-Román, P.E. Pérez-Estigarriba, PY-CrackDB: a pavement crack dataset from Paraguayan roads for context-aware computer vision models [Data set], Zenodo (2025), doi:10.5281/zenodo.16749554.
[5] L. Cui, Z. Qi, Z. Chen, F. Meng, Y. Shi, Pavement distress detection using random decision forests, in: *data mining*, 2015, pp. 95–102, doi: 10.1007/978-3-319-24474-7_14.
[6] R. Amhaz, S. Chambon, J. Idier, V. Baltazart, A new minimal path selection algorithm for automatic crack detection on pavement images, in: 2014 IEEE International Conference on Image Processing (ICIP), IEEE, 2014, pp. 788–792, doi:10.1109/ICIP.2014.7025158.
[7] L. Zhang, F. Yang, Y.D. Zhang, Y.J. Zhu, Road crack detection using deep convolutional neural network, in: 2016 IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp. 3708–3712, doi:10.1109/ICIP.2016.7533052.
[8] Q. Zou, Y. Cao, Q. Li, Q. Mao, S. Wang, CrackTree: automatic crack detection from pavement images, Pattern Recognit Lett. 33 (3) (2012) 227–238, doi:10.1016/j.patrec.2011.11.004.
[9] M. Sabouri, A. Sepidbar, SUT-Crack: a comprehensive dataset for pavement crack detection across all methods, Data Brief. 51 (2023) 109642, doi:10.1016/j.dib.2023.109642.
[10] The World Bank Group, "Climate Change Knowledge Portal: paraguay," *World Bank*. [Online]. Available: https://climateknowledgeportal.worldbank.org/country/paraguay/climate-data-historical. [Accessed: Sep. 4, 2025].
[11] Japan International Cooperation Agency, "Study on the Integrated Rural Development in the Southern Area of the Eastern Region in the Republic of Paraguay, final report, appendix," Rep. 11831740, 2005. [Online]. Available: https://openjicareport.jica.go.jp/pdf/11831740_02.pdf. [Accessed: Sep. 4, 2025].
[12] Sacyr, "Ruta PY02: uno de los proyectos viales más importantes de Paraguay," ⟨13:italic ⟩https://www.google.com/search?q=sacyr.com⟨/13:italic⟩. [Online]. Available: https://sacyr.com/-/ruta-py-02. [Accessed: Sep. 4, 2025].
[13] MOPC licita modernización del antiguo trazado de la PY02 en Coronel Oviedo, ABC Color. (Jun. 24, 2024) [Online]. Available: https://www.abc.com.py/nacionales/2024/06/24/mopc-licita-modernizacion-del-antiguo-trazado-de-la-py02-en-coronel-oviedo/. [Accessed: Sep. 4, 2025].
[14] Asfaltado de la ruta 8 avanza, pero preocupa calidad de obra„ Última Hora. (Nov. 29, 2010) [Online]. Available: https://www.ultimahora.com/asfaltado-la-ruta-8-avanza-pero-preocupa-calidad-obra-n292360. [Accessed: Sep. 4, 2025].
[15] MOPC recibirá ofertas para rehabilitar 150 km de rutas en tres departamentos del país, Agencia IP. (Oct. 11, 2020) [Online]. Available: https://www.ip.gov.py/ip/2020/10/11/mopc-recibira-ofertas-para-rehabilitar-150-km-de-rutas-en-tres-departamentos-del-pais/. [Accessed: Sep. 4, 2025].