Red neuronal semántica para la detección de carriles peatonales

Semantic Neural Network for pedestrian lane detection

[1]Juan Manuel Aldana Porrasx, [2]John Fredy Montes Mora

[1,2]Universidad Nacional Abierta y a Distancia, Colombia

Recibido: 20/10/2023 Aprobado 20/11/2023

Resumen

El presente trabajo forma parte del proyecto PG2402ECBTI2022 aprobado por la Universidad Nacional Abierta y a Distancia. Su objetivo es desarrollar un sistema de percepción espacial con técnicas de inteligencia artificial (IA) para mejorar la orientación y movilidad de personas con discapacidad visual en la ciudad de Ibagué. El sistema propuesto consta de varios módulos, y este texto se enfoca en los resultados del módulo de procesamiento de imágenes orientado a la segmentación de carriles peatonales mediante el uso de redes neuronales semánticas. La red propuesta para esta tarea parte de una arquitectura Deeplab y se entrenó a partir de un dataset compuesto por 5.443 imágenes. Los resultados muestran que la arquitectura obtuvo un nivel de exactitud de 0.95 y que su capacidad para generalizar nuevas muestras es acorde a la tarea exigida para el sistema en general.

Palabras clave: discapacidad visual, red neuronal, deeplab, carriles peatonales.

Abstract

This work is part of the PG2402ECBTI2022 project approved by the Universidad Nacional Abierta y a Distancia. Its objective is to develop a spatial perception system with artificial intelligence (AI) techniques to improve the orientation and mobility of visually impaired people in Ibagué. The proposed system consists of several modules, and this text focuses on the results of the image processing module oriented to the segmentation of pedestrian lanes through the use of semantic neural networks. The network proposed for this task is based on a Deeplab architecture and was trained on a dataset composed of 5443 images. The results show that the

Citación: Aldana Porras, J. M. ., & Montes Mora, J. F. . (2023). Red Neuronal semántica para la detección de carriles peatonales. Publicaciones E Investigación, 17(4). https://doi.org/10.22490/25394088.7476

[1]juan.aldana@unad.edu.co / https://orcid.org/0000-0002-6969-1044
[2]john.montes@unad.edu.co / https://orcid.org/0000-0003-0466-343

https://doi.org/10.22490/25394088.7476

architecture obtained an accuracy level of 0.95 and that its capacity to generalize new samples is in accordance with the task required for the system in general.

Keywords: Visually impaired, neural semantic network, deeplab.

1. Introducción

Las personas en situación de discapacidad en Colombia representan un porcentaje significativo de la población total, con un 6.4 % según el Departamento Administrativo Nacional de Estadística (DANE, 2018). Esto se traduce en aproximadamente 3.134.036 personas afectadas en el país, de las cuales un considerable 62.17 % se relaciona con limitaciones visuales. En la ciudad de Ibagué, que alberga a 529.635 habitantes, un 4.4 % de la población sufre de alguna discapacidad, entre los cuales 6.291 individuos enfrentan dificultades visuales.

A pesar de esto, los entornos urbanos distan bastante de las necesidades reales que tienen estos grupos poblacionales para su integración en la vida en sociedad, entendiendo que dichos entornos son esenciales para la vida en comunidad y que reflejan complejas interacciones humanas (Pérgolis & Moreno, 2009). Estos entornos, aunque cruciales para el desarrollo integral, a menudo limitan la movilidad y autonomía de las personas con discapacidad visual, revelando la desigualdad espacial y arquitectónica que obstaculiza su participación plena (Cabrera, 2019).

En este sentido, se destaca la necesidad de una profunda reflexión sobre las condiciones que permitan la autonomía de estas personas, subrayando la importancia de su inclusión en los procesos urbanos y arquitectónicos (Vega, 2015). Las limitaciones físicas y sociales en el entorno afectan significativamente la autonomía y la integración social de estas personas, lo que enfatiza la importancia de estudiar la interacción entre el espacio, la orientación y la movilidad para su empoderamiento y reconocimiento en la sociedad (Regalado, 2019).

Considerando la necesidad de apoyo para este grupo poblacional, la tecnología asistiva ha surgido como una herramienta invaluable. Según Bryant & Bryant (2003, p. 2), la tecnología asistiva abarca una amplia gama de instrumentos y dispositivos diseñados para facilitar actividades cotidianas y suplir limitaciones físicas o cognitivas (Smythe *et al.*, 2020). Este tipo de tecnología abarca desde facilitar tareas básicas hasta permitir una participación plena en diversas actividades, como aprender, trabajar y disfrutar de actividades recreativas (Bodine, 2013).

En consonancia con el impulso hacia soluciones innovadoras y la aplicación de tecnologías asistivas, el trabajo se inscribe en el marco del proyecto PG2402ECBTI2022, aprobado en la convocatoria 010 de la Universidad Nacional Abierta y a Distancia. El enfoque de este proyecto es la implementación de un sistema de percepción espacial mediado por técnicas de inteligencia artificial (IA) con el objetivo de fortalecer la orientación y la movilidad de la población con discapacidad visual en Ibagué. En tal sentido, se presentan los resultados concernientes a uno de los módulos de procesamiento que hace parte del sistema propuesto y que se ilustra en la Figura 1.

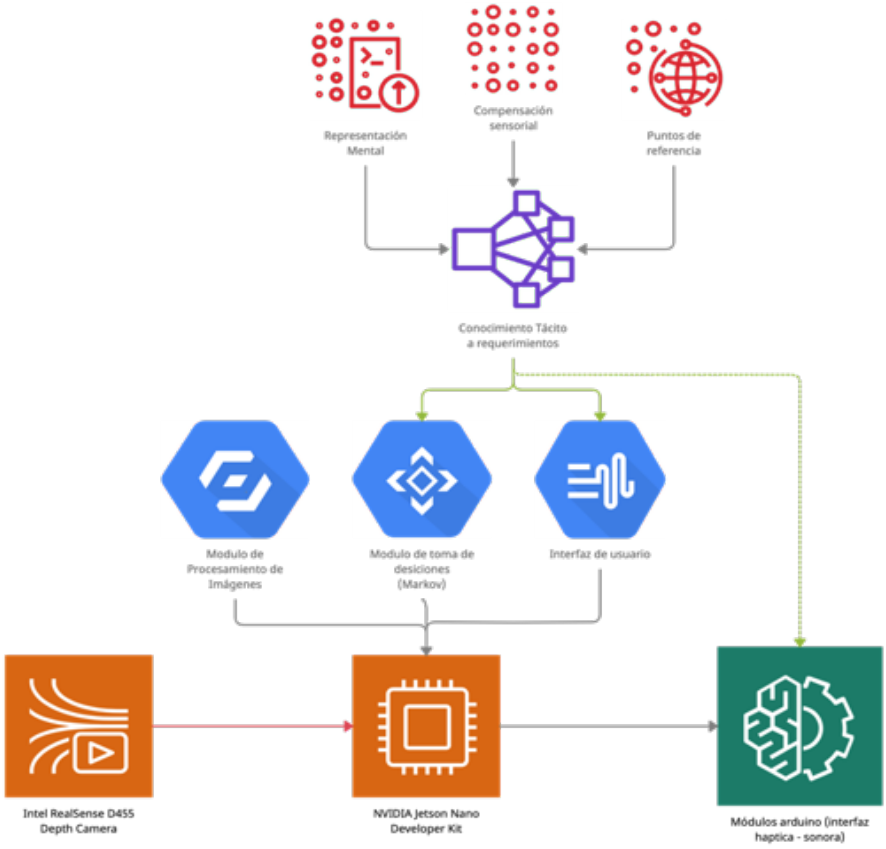

Figura 1. Sistema propuesto.

Así pues, dentro del módulo de procesamiento de imágenes es posible diferenciar dos grandes componentes, el primero dedicado al procesamiento tridimensional de imágenes y el segundo enfocado a la segmentación de carriles peatonales (Le *et al*., 2014) mediante la aplicación de redes neuronales semánticas y del cual parte la presente ponencia. En cuanto a la problemática en cuestión, la detección automática de carriles peatonales es de gran importancia para la navegación asistida (Thanh Le *et al*., 2022), a pesar de esto, garantizar la seguridad de los usuarios con problemas de visión mientras se movilizan por entornos desconocidos se limita en gran medida a la facilidad para gestionar con eficacia las distintas superficies de los carriles y las condiciones de iluminación, lo que lo hace un problema difícil de abordar (Ang *et al*., 2021).

2. Materiales y métodos

En coherencia con lo planteado, para la creación del modelo se emplea la metodología CRISP-ML (CRoss-Industry Standard Process model for the development of Machine Learning applications) propuesto por Studer *et al*. (2021), como marco de referencia para llevar a cabo proyectos de machine learning donde se estipulan las actividades que deben realizarse para asegurar la calidad y replicabilidad de los modelos implementados. Su ejecución se lleva a cabo en seis etapas (Figura 2):

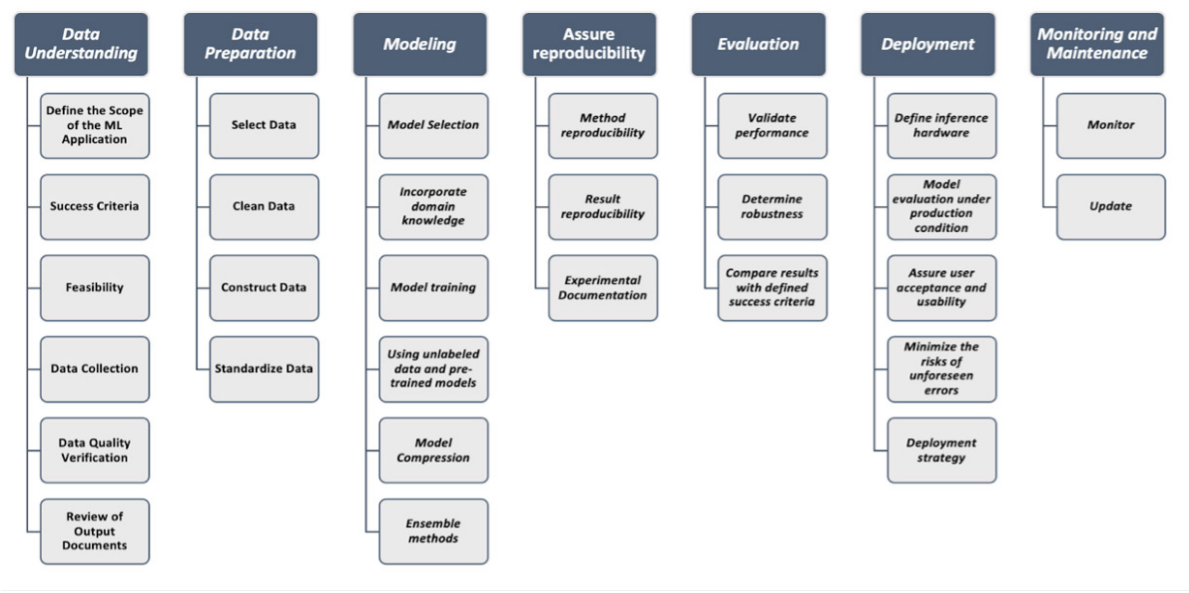

Figura 2. Modelo CRISP-ML.

Por otra parte, para estructurar la red neuronal se dispuso de la creación de un dataset compuesto por imágenes de andenes y aceras presentes en la ciudad de Ibagué, para posteriormente estructurar el modelo propuesto a partir de una arquitectura DeepLab (Chen *et al*., 2018). El dataset en cuestión cuenta con un total de 5.443 imágenes tomadas con un sensor Sony IMX 586 (12 MP, f/2.2, 1/2.9", 1.25µm, PDAF) en distintas partes de la ciudad tratando de emular la mayor cantidad de escenarios posibles. Para el proceso de etiquetado se utilizó un mapa de bitmaps generado manualmente para cada una de las imágenes del dataset (Figura 3). No se utilizaron técnicas de aumentación de datos para la creación del mismo.

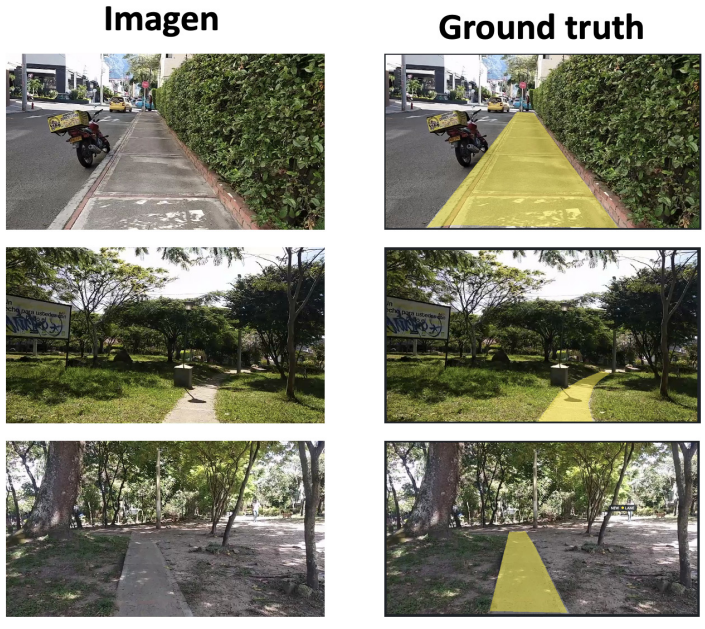

Figura 3. Imágenes de ejemplo del dataset.

Para el conjunto de datos de entrenamiento y validación se utilizó una proporción 80 -20. La Figura 4 muestra el gráfico de dispersión resultante para cada uno de los conjuntos de datos.

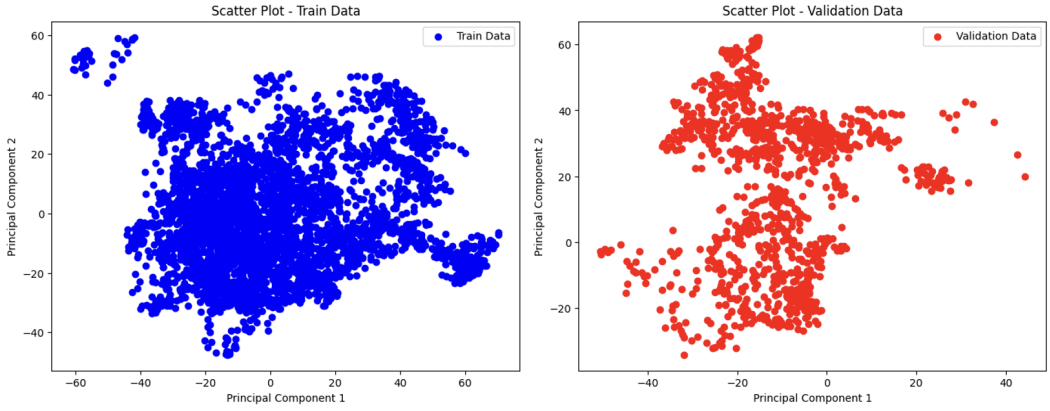

Figura 4. Gráficos de dispersión para los conjuntos de datos.

La Figura 5 muestra el promedio de los valores de los píxeles a lo largo de todas las imágenes para los dos conjuntos de datos (entrenamiento y validación).

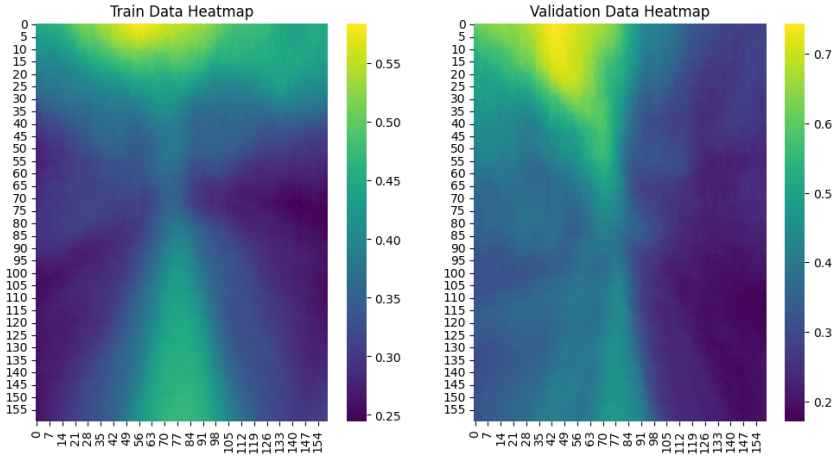

Figura 5. Heatmap: valores de los pixeles del dataset.

3. Resultados y discusión

Para el modelo propuesto se parte de una arquitectura de red DeepLab (Figura 5), orientada a la segmentación semántica (Buduma *et al.*, 2022) buscando comprender las imágenes a nivel de pixel, cuyo objetivo es asignar a cada pixel de la imagen una clase de objeto específica o una etiqueta semántica (Dey, 2018). La arquitectura propuesta parte de una red U-Net modificada (Antiga *et al.*, 2020) con capas de convolución dilatadas y se compone de tres partes principales: el codificador (downsampler), el bloque central y el decodificador (upsampler).

El codificador está diseñado para reducir la resolución espacial y extraer características de alto nivel de la imagen de entrada, consta de dos capas convolucionales seguidas de una capa de activación ReLU y una capa de normalización por lotes (Batch Normalization) (Chollet, 2021). La primera capa convolucional

utiliza 112 filtros y un tamaño de kernel de (3, 3) con un desplazamiento (stride) de 2, mientras que la segunda capa convolucional utiliza 32 filtros con las mismas configuraciones.

Para el caso del bloque central, utiliza las convoluciones dilatadas para aumentar el campo receptivo y permitir que el modelo capture características contextuales de la imagen. Consta de dos capas convolucionales dilatadas (Gulli *et al.*, 2019), ambas con un factor de dilatación (dilation rate) de 2 y el mismo número de filtros (128) que se establece en el hiperparámetro middle_filters. Al igual que en el codificador, cada capa convolucional está seguida por una capa de activación ReLU y una capa de normalización por lotes.

Por último, el decodificador está diseñado para aumentar la resolución espacial y producir una máscara segmentada con el mismo tamaño que la imagen de entrada original. Este cuenta con dos capas de convolución transpuesta, seguidas de una capa de activación ReLU y una capa de normalización por lotes. La primera capa de convolución transpuesta utiliza 160 filtros y un tamaño de kernel de (3, 3) con un desplazamiento (stride) de 2, mientras que la segunda capa de convolución transpuesta utiliza un número de filtros igual al número de clases objetivo (2) y también tiene un tamaño de kernel de (3, 3) con un desplazamiento de 2. La última capa de convolución transpuesta utiliza la función de activación "softmax" para generar la máscara segmentada final. La Figura 6 ilustra la arquitectura de la red.

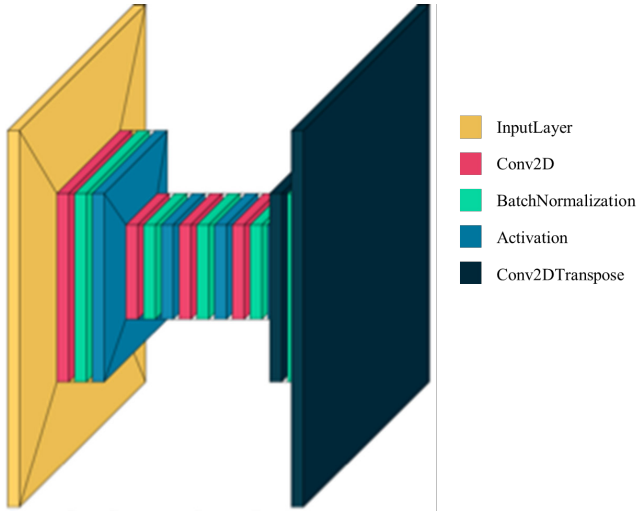

Figura 6. Representación gráfica de la arquitectura de la red.

En relación a la optimización del modelo se definió el espacio de búsqueda de hiperparámetros (Kneusel, 2021) y se utilizó un algoritmo de optimización que realiza una búsqueda adaptativa basada en un proceso de eliminación temprana (Ekman, 2021), el cual permite realizar múltiples experimentos con diferentes combinaciones de hiperparámetros en paralelo y descarta las peores combinaciones. El resultado final de esta búsqueda se resume en la Tabla 1.

Tabla 1. Hiperparámetros de la red

Value	Best Value So Far	Hyperparameter
112	64	conv1_filters
32	224	conv2_filters
128	128	middle_filters
160	32	Transpl_filters
0,0001	0,01	learning_rate
7	3	tuner/epochs

Finalmente, luego de realizar el proceso de entrenamiento correspondiente, el modelo se utiliza para predecir máscaras segmentadas a partir de las imágenes de prueba. Las figuras 7 y 8 muestran respectivamente la evolución del modelo tanto en precisión como en érdida en el transcurso de los 20 epoch de entrenamiento.

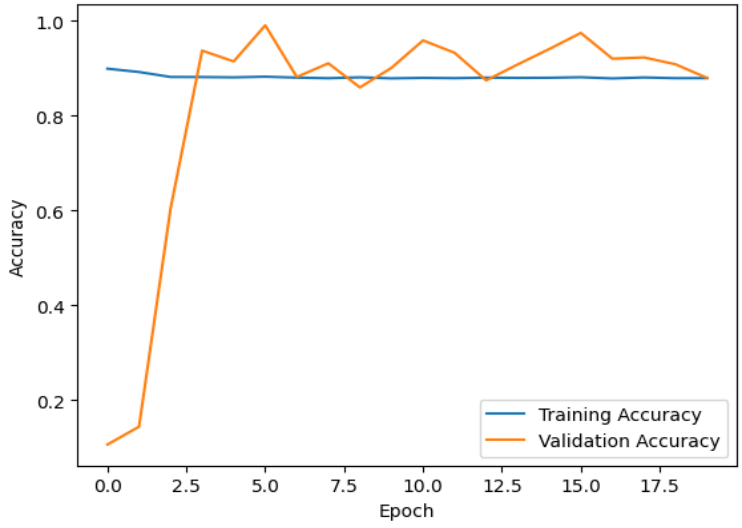

Figura 7. Validación de la precisión (accuracy) del modelo en el proceso de entrenamiento.

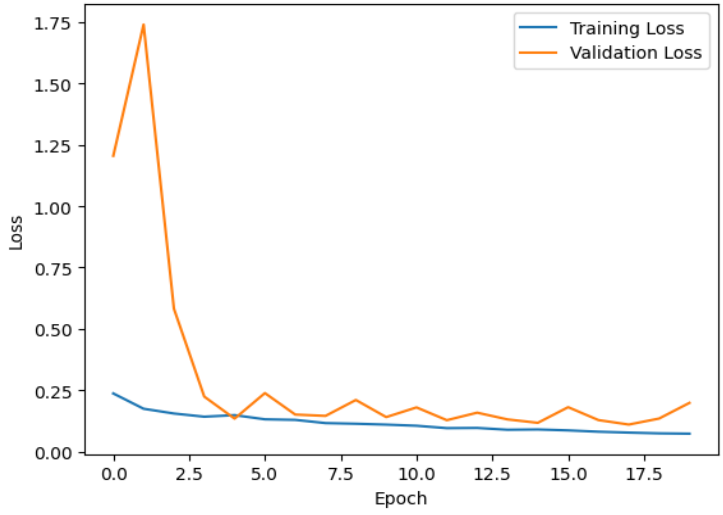

Figura 8. Validación de la pérdida (loss) del modelo en el proceso de entrenamiento.

Dentro del proceso de entrenamiento se denota una fluctuación entre los picos de precisión del modelo en cada uno de los epoch, esto dado a la dificultad propia del problema abordado en donde la diferenciación semántica de cada uno de los pixeles del modelo es compleja, y ciertas áreas de las muestras tienden a marcarse como falsos positivos. A pesar de eso, las métricas del modelo (Patterson & Gibson, 2017) muestran un Recall de 0.86 y un F1 score de 0.82. El modelo final obtiene un 0.95 de accuracy, que si bien muestra que existe dificultad para identificar algunos escenarios puntuales, en general no afecta el rendimiento general del modelo en su uso práctico. La Figura 9 ilustra las detecciones realizadas por el modelo.

Figura 9. Detecciones realizadas por el modelo.

4. Conclusiones

DeepLab es una arquitectura de red neuronal profunda altamente eficaz para la segmentación semántica de imágenes, que para el caso particular de este proyecto resulta ser muy beneficiosa, sobre todo en apartados relacionados a la captura de características contextuales mediante convoluciones dilatadas, y capacidad de generalización en diferentes contextos. En contraprestación tiende a presentar una mayor complejidad computacional que representan requerimientos de memoria y sensibilidad a la calidad de los datos de entrenamiento. A pesar de estos aspectos, la red propuesta se mostró robusta ante los desafíos presentes en el dataset, que variaba según el ángulo y las condiciones de iluminación.

Referencias

Ang, S. P., Phung, S. L., Bouzerdoum, A., Nguyen, T. N. A., Duong, S. T. M., & Schira, M. M. (2021). Real-time Pedestrian Lane Detection for Assistive Navigation using Neural Architecture Search. *2020 25th International Conference on Pattern Recognition (ICPR),* 8392-8399. https://doi.org/10.1109/ICPR48806.2021.9412741

Antiga, L. P. G., Stevens, E., & Viehmann, T. (2020). *Deep Learning with PyTorch*. Simon and Schuster.

Bodine, C. (2013). *Assistive Technology and Science.* Sage Publications. http://search.ebscohost.com.bibliotecavirtual.unad.edu.co/login.aspx?direct=true&db=nlebk&AN=525937&lang=es&site=eds-live&scope=site

Bryant, D. P., & Bryant, B. R. (2003). *Assistive technology for people with disabilities.* Pearson.

Buduma, N., Buduma, N., & Papa, J. (2022). *Fundamentals of Deep Learning*. O'Reilly Media, Inc.

Cabrera, F. (2019). *Movilidad Urbana, espacio público y ciudadanos sin autonomía: el caso de Lima.* (Tesis doctoral) Universidad Autónoma de Barcelona. https://ddd.uab.cat/pub/tesis/2019/hdl_10803_667392/icv1de1.pdf

Chen, L-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(4), 834-848. https://doi.org/10.1109/TPAMI.2017.2699184

Chollet, F. (2021). *Deep Learning with Python.* Second Edition. Simon and Schuster.

DANE. (2018). *Censo Nacional de Población y Vivienda 2018.* https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/censo-nacional-de-poblacion-y-vivenda-2018

Dey, S. (2018). *Hands-On Image Processing with Python: Expert techniques for advanced image analysis and effective interpretation of image data.* Packt Publishing Ltd.

Ekman, M. (2021). *Learning deep learning: Theory and practice of neural networks, computer vision, nlp, and transformers using tensorflow*. Addison-Wesley.

Gulli, A., Kapoor, A., & Pal, S. (2019). *Deep Learning with TensorFlow 2 and Keras: Regression, ConvNets, GANs, RNNs, NLP, and More with TensorFlow 2 and the Keras API*. Packt Publishing.

Kneusel, R. T. (2021). *Math for Deep Learning: What You Need to Know to Understand Neural Networks*. No Starch Press.

Patterson, J., & Gibson, A. (2017). *Deep learning: A practitioner's approach*. O'Reilly Media, Inc.

Pérgolis, j. C., & Moreno Hernández, D. (2009). La capacidad comunicante del espacio. *Revista de Arquitectura, 11*, 68-73.

Regalado, G. D. (2019). El capital de la movilidad urbana cotidiana: motilidad en la periferia de Lima Metropolitana. *Revista de Arquitectura, 22*(1). 67-81. https://doi.org/10.14718/RevArq.2020.3038

Smythe, K.., Adam, D. , da Silva, C. , & Okimoto, M. (2020). Including users in the evaluation process of assistive technology: Applied methods and techniques. *Advances in Intelligent Systems and Computing, 972*(1), 940–947. https://doi.org/10.1007/978-3-030-19135-1_92

Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Müller, K. R. (2021). Towards CRISP-ML (Q): a machine learning process model with quality assurance methodology. *Machine Learning and Knowledge Extraction, 3*(2), 392-413.https://arxiv.org/pdf/2003.05155.pdf

Thanh Le, H., Phung, S. L., & Bouzerdoum, A. (2022). Bayesian Gabor Network With Uncertainty Estimation for Pedestrian Lane Detection in Assistive Navigation. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(8), 5331-5345. https://doi.org/10.1109/TCSVT.2022.3144184

Vega, M. (2015). Ciudad, espacio y ceguera en ciudad Juárez México. *Revista Latinoamericana de Estudios sobre Cuerpos, Emociones y Sociedad, 7*(17),42-50. https://www.redalyc.org/articulo.oa?id=273238564005