Artículo Científico / Scientific Paper

UNIVERSIDAD POLITÉCNICA SALESIANA ECUADOR
INGENIUS
Revista de Ciencia y Tecnología
https://doi.org/10.17163/ings.n35.2026.09
pISSN: 1390-650X / eISSN: 1390-860X

Mejorando la segmentación semántica para la accesibilidad urbana mediante datos sintéticos de alta fidelidad

Enhancing semantic segmentation for urban accessibility using high-fidelity synthetic data

Santiago Felipe Luna Romero[1], Renato Gouveia[1,*], Mauren Abreu de Souza[1]

Recibido: 11-07-2025, Recibido tras revisión: 01-12-2025, Aceptado: 09-12-2025, Publicado: 01-01-2026

Resumen

La segmentación semántica de escenas urbanas es un componente clave para el desarrollo de ciudades inteligentes; sin embargo, su efectividad depende de grandes volúmenes de datos anotados a nivel de píxel, los cuales son costosos y especialmente escasos en clases críticas relacionadas con la accesibilidad y la movilidad asistida. Este trabajo tiene como objetivo mejorar la segmentación semántica para aplicaciones de accesibilidad urbana mediante el uso de datos sintéticos. La metodología propuesta integra la generación de datos sintéticos hiperrealistas utilizando Unreal Engine 5.1, el procesamiento automático de máscaras semánticas con etiquetas perfectas y el entrenamiento de modelos de segmentación de referencia. Se generaron 5036 imágenes anotadas en 22 clases, incluyendo aceras, sillas de ruedas y bastones. Se evaluaron dos arquitecturas de segmentación: una *U-Net* básica y DeepLabv3+ con módulos ASPP. El preentrenamiento con datos sintéticos incrementó el mIoU global de 0.0626 a 0.84, lo que representa una mejora de 13.4×, y produjo aumentos significativos en precisión, *recall* y *F1-score* (aproximadamente 6.8×, 9.3× y 10.4×, respectivamente). En clases críticas para la accesibilidad, se alcanzó un IoU de 0.94 para sillas de ruedas motorizadas y un *recall* de 0.98 para aceras. En total, las 22 clases superaron el umbral operativo de despliegue (IoU $\geq$ 0.75). Estos resultados demuestran que la incorporación de datos sintéticos, junto con estrategias de entrenamiento sensibles al desbalance de clases, constituye una solución efectiva y escalable para el desarrollo de sistemas robustos de segmentación semántica orientados a la accesibilidad urbana.

Palabras clave: accesibilidad, aprendizaje profundo, ciudades inteligentes, datos sintéticos, segmentación semántica, inteligencia artificial

Abstract

Semantic segmentation of urban scenes is essential for the development of smart cities; however, its effectiveness relies heavily on large, pixel-level annotated datasets, which are particularly scarce for mobility aids. This study aims to enhance semantic segmentation for urban accessibility applications by leveraging synthetic data. The proposed methodology integrates high-fidelity synthetic data generation using Unreal Engine 5.1, automated semantic mask processing, and the training of state-of-the-art segmentation models. A dataset of 5,036 images with pixel-perfect labels across 22 classes, including sidewalks, wheelchairs, and walking aids, was created to support this investigation. Two architectures were benchmarked: a baseline U-Net and DeepLabv3+ with ASPP. Pretraining with synthetic data increased global mIoU from 0.0626 to 0.84 (13.4×) and substantially improved precision, recall, and F1-score (by approximately 6.8×, 9.3×, and 10.4×, respectively). For accessibility-critical classes, motorized wheelchairs achieved an IoU of 0.94, and sidewalks attained a recall of 0.98. Overall, all 22 classes surpassed the deployment threshold ($\geq$ 0.75 IoU). These findings demonstrate that synthetic data, combined with imbalance-aware training strategies, provides a viable pathway toward robust semantic segmentation solutions for urban accessibility applications.

Keywords: Semantic Segmentation, Synthetic Data, Deep Learning, Smart Cities, Accessibility, Artificial Intelligence

[1,*]Pontifícia Universidade Católica do Paraná, Brasil.
Autor para correspondencia ✉: gouveia.renato@pucpr.edu.br.

Forma sugerida de citación: S. F. Luna Romero, R. Gouveia y M. Abreu de Souza, "Mejorando la segmentación semántica para la accesibilidad urbana mediante datos sintéticos de alta fidelidad," *Ingenius, Revista de Ciencia y Tecnología*, N.° 35, pp. 122-137, 2026. DOI: https://doi.org/10.17163/ings.n35.2026.09.

1. Introducción

La segmentación semántica, que asigna una etiqueta semántica a cada píxel de una imagen, constituye un componente fundamental para la comprensión de escenas urbanas complejas. Esta técnica respalda aplicaciones como la conducción autónoma, el monitoreo del tráfico, la realidad aumentada y la navegación asistida para peatones con discapacidades de movilidad o visuales [1, 2]. Al discriminar entre carreteras, aceras, edificios, vehículos, peatones y elementos asociados a la accesibilidad —como rampas de bordillo y ayudas para la movilidad—, los modelos de segmentación proporcionan la conciencia espacial necesaria para la planificación del transporte, el diseño urbano inclusivo y la detección de obstáculos en tiempo real en dispositivos de asistencia [1, 2].

No obstante, los modelos profundos de segmentación continúan dependiendo de grandes conjuntos de datos anotados a nivel de píxel. La generación de dichas etiquetas resulta costosa y requiere un considerable esfuerzo humano, particularmente para estructuras finas y clases poco frecuentes. Esta limitación es especialmente crítica en categorías relevantes para la seguridad pero subrepresentadas, como peatones que utilizan sillas de ruedas, andadores o bastones, cuya escasa presencia en los conjuntos de datos disponibles restringe la capacidad de generalización en entornos urbanos reales [3, 4].

Conjuntos de datos ampliamente utilizados para escenas urbanas, como Cityscapes y KITTI, contienen muy pocos o ningún ejemplo de usuarios con movilidad reducida, lo que introduce sesgos sistemáticos y provoca que los modelos omitan estas clases o las confundan con el fondo [3, 5].

La simulación de alta fidelidad ha surgido como una estrategia eficaz para mitigar la dependencia de grandes conjuntos de datos del mundo real. Motores de videojuegos modernos, como Unreal Engine, así como simuladores especializados como CARLA, permiten generar grandes volúmenes de imágenes fotorrealistas con máscaras semánticas renderizadas automáticamente y precisas a nivel de píxel. Este enfoque reduce de manera significativa los costos de anotación y facilita la experimentación controlada sobre clases poco frecuentes o críticas para la seguridad [1, 4, 6].

Estudios previos han demostrado que el preentrenamiento de modelos de segmentación con imágenes sintéticas, seguido de un ajuste fino con conjuntos de datos reales más pequeños, mejora el desempeño en escenas urbanas complejas [1, 7]. Asimismo, técnicas de aleatorización de dominio —que varían sistemáticamente la iluminación, las texturas, las condiciones climáticas, los puntos de vista de la cámara y la disposición de los objetos— contribuyen a reducir el sobreajuste a artefactos del simulador y a mejorar la robustez en la transferencia de simulación a realidad [8, 9].

Dos cuestiones siguen siendo centrales para la accesibilidad urbana y la navegación asistida. En primer lugar, persiste un marcado desbalance de clases: las categorías dominantes, como carreteras, cielo y edificios, ocupan la mayoría de los píxeles, mientras que los dispositivos de movilidad y los elementos de infraestructura estrechos —incluidas rampas de bordillo, bolardos y señales de tráfico— representan solo una fracción mínima del total [3, 5]. En ausencia de contramedidas específicas, los modelos de segmentación tienden a sobreajustarse a las clases mayoritarias e infraajustarse a las categorías minoritarias que resultan críticas para la accesibilidad. Para mitigar este efecto, se emplean comúnmente funciones de pérdida que reajustan el peso de ejemplos difíciles de clasificar o subrepresentados, como la pérdida focal y la pérdida Tversky, junto con estrategias de aumento de datos sensibles a las clases.

En segundo lugar, las arquitecturas de segmentación deben equilibrar el contexto global con la preservación del detalle fino. Las arquitecturas CNN de tipo codificador–decodificador, como U-Net y DeepLab, capturan eficazmente la estructura local, pero presentan campos receptivos limitados. Por el contrario, los modelos basados en Transformers ofrecen una capacidad sólida de razonamiento global, aunque pueden tener dificultades para mantener límites precisos a nivel de píxel. Con el fin de abordar esta dicotomía, las arquitecturas híbridas CNN–Transformer combinan respaldos convolucionales con módulos de autoatención, permitiendo modelar dependencias de largo alcance sin sacrificar la definición de contornos en escenas urbanas congestionadas [10–12].

La mayoría de los conjuntos de datos sintéticos existentes y los flujos de trabajo de segmentación se centran principalmente en aplicaciones de conducción autónoma y no abordan de forma explícita la accesibilidad urbana. Sus taxonomías priorizan participantes genéricos del tráfico y elementos de infraestructura de carácter grueso, e incluyen rara vez etiquetas detalladas para ayudas de movilidad o componentes a nivel de acera. En consecuencia, estos conjuntos de datos no resultan directamente adecuados para evaluar la segmentación orientada a la accesibilidad ni para respaldar sistemas de navegación asistida.

En contraste, el conjunto de datos SYNTHUA-DT (Synthetic Urban Accessibility – Digital Twin) [13] se centra explícitamente en la accesibilidad urbana. Este conjunto modela un entorno urbano realista mediante Unreal Engine 5.1 y proporciona anotaciones semánticas perfectas a nivel de píxel para 22 clases, incluyendo múltiples categorías de dispositivos de movilidad —como sillas de ruedas, andadores y bastones—, así como peatones y elementos de infraestructura a nivel de acera. SYNTHUA-DT fue diseñado para abordar esta brecha en los datos orientados a la accesibilidad, ofreciendo un corpus controlable en el que las

ayudas de movilidad y las estructuras peatonales están representadas de manera sistemática y reproducible.

1.1. Trabajos relacionados

Los datos sintéticos para la comprensión de escenas urbanas han sido ampliamente explorados mediante el uso de videojuegos comerciales y simuladores dedicados. Kamimura *et al.* extrajeron anotaciones densas a partir de GTA-V y demostraron que el preentrenamiento con datos sintéticos puede mejorar el rendimiento de los modelos cuando se combina con un ajuste fino sobre datos del mundo real [7]. De manera similar, flujos de trabajo basados en CARLA se han utilizado para ampliar conjuntos de datos como Cityscapes con escenarios de tráfico adicionales y condiciones climáticas adversas, incrementando la robustez frente a configuraciones poco frecuentes [1]. Los entornos de gemelo digital, como UrbanSyn, integran modelos urbanos tridimensionales realistas con técnicas de adaptación de dominio y transferencia de estilo con el objetivo de reducir la brecha entre imágenes simuladas y reales [4,6]. En conjunto, estos estudios reportan de forma consistente mejoras en la precisión y la capacidad de generalización; sin embargo, sus espacios de etiquetas se orientan principalmente a participantes genéricos del tráfico y no abordan de manera sistemática los elementos relacionados con la accesibilidad urbana.

El desbalance de clases y la segmentación de objetos pequeños también han sido objeto de un análisis extensivo. Azad *et al.* y Liu *et al.* estudiaron cómo las distribuciones de etiquetas de cola larga degradan el rendimiento en clases minoritarias y evaluaron funciones de pérdida, como la pérdida focal y la pérdida Tversky, para reajustar el peso de ejemplos difíciles de clasificar o subrepresentados [3, 5]. Arquitecturas como U-Net y DeepLabv3+ se utilizan comúnmente como líneas base debido a su equilibrio entre precisión y costo computacional [14, 15]. Asimismo, se han propuesto enfoques híbridos CNN–Transformer, incluidos codificadores basados en Swin Transformer y decodificadores aumentados con mecanismos de autoatención, para capturar dependencias de largo alcance al tiempo que se preservan detalles estructurales en escenas urbanas complejas [10–12]. No obstante, la mayoría de estas evaluaciones se apoyan en conjuntos de datos en los que los peatones con discapacidad de movilidad y la infraestructura de accesibilidad están ausentes o severamente subrepresentados, lo que limita su aplicabilidad directa a sistemas de navegación asistida.

1.2. Contribuciones

Basándose en el conjunto de datos SYNTHUA-DT, este trabajo investiga si los datos sintéticos de alta fidelidad, combinados con estrategias de entrenamiento sensibles al desbalance de clases, son suficientes para alcanzar un desempeño de segmentación apto para el despliegue en clases críticas para la accesibilidad urbana. Asimismo, se cuantifican las mejoras obtenidas en comparación con una línea base basada en U-Net.

Las principales contribuciones de este trabajo se resumen a continuación:

- **Uso de un conjunto de datos sintético orientado a la accesibilidad.** El conjunto de datos SYNTHUA-DT [13], generado mediante Unreal Engine 5.1, proporciona 5036 imágenes de alta resolución con anotaciones semánticas perfectas a nivel de píxel en 22 clases, incluyendo explícitamente múltiples dispositivos de movilidad y elementos de infraestructura a nivel de acera.
- **Canalización de preprocesamiento y estructuración del conjunto de datos.** Se propone una canalización que convierte máscaras codificadas por color en supervisión multicanal, aplica refinamientos morfológicos simples sensibles a la clase y genera particiones de entrenamiento, validación y prueba que respaldan un entrenamiento y una evaluación robustos.
- **Evaluación comparativa de arquitecturas de segmentación.** Las arquitecturas U-Net y DeepLabv3+ se entrenan y evalúan empleando funciones de pérdida sensibles al desbalance y técnicas de aumento de datos. El análisis incluye métricas globales y por clase, con énfasis particular en ayudas de movilidad y aceras.
- **Marco orientado al despliegue en sistemas de accesibilidad.** La canalización sintética propuesta, junto con el análisis de calibración, sienta las bases para extensiones futuras mediante técnicas de adaptación de dominio hacia conjuntos de datos del mundo real y su integración en sistemas de ciudades inteligentes y navegación asistida.

El resto del artículo se organiza de la siguiente manera. La Sección 2 describe la metodología propuesta; la Sección 3 presenta los resultados experimentales y el análisis por clases; la Sección 4 discute las limitaciones y las líneas de trabajo futuro; y la Sección 5 indica las conclusiones principales.

2. Materiales y métodos

El enfoque propuesto comprende cinco componentes principales: (i) la generación del conjunto de datos sintético, (ii) el preprocesamiento de imágenes y máscaras semánticas, (iii) el diseño de las arquitecturas de los modelos, (iv) la estrategia de entrenamiento, incluido el diseño de la función de pérdida, y (v) las métricas

de evaluación. En conjunto, estos elementos conforman una canalización sistemática para el entrenamiento y la evaluación de modelos de segmentación semántica en escenas urbanas orientadas a la accesibilidad.

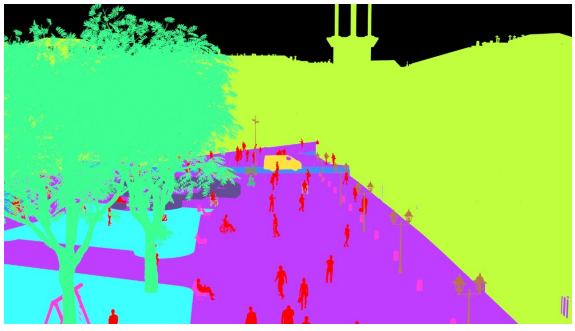
2.1. Generación del conjunto de datos sintético

Este trabajo se basa en el conjunto de datos SYNTHUA-DT (Synthetic Urban Accessibility – Digital Twin) [13, 16], un corpus sintético compuesto por 5036 imágenes urbanas de alta resolución (1920 × 1080 px) generado mediante Unreal Engine 5.1. Para la síntesis de las escenas se empleó una canalización de renderizado físicamente basada (PBR), con el objetivo de aproximar de manera realista la iluminación y las propiedades de los materiales del entorno urbano. Cada escena incluye disposiciones arquitectónicas diversas —como fachadas históricas y rascacielos modernos—, junto con mobiliario urbano, actores dinámicos (peatones, ciclistas y vehículos) y múltiples condiciones climáticas.

Las máscaras semánticas perfectas a nivel de píxel correspondientes a 22 clases —entre ellas aceras, carreteras, pasos peatonales, peatones, dispositivos de movilidad, vegetación, vehículos, señalización y otros elementos de infraestructura— se generaron automáticamente durante el proceso de renderizado [16, 17]. La taxonomía de clases fue diseñada explícitamente para resaltar componentes relevantes para la accesibilidad urbana, incluyendo múltiples categorías de sillas de ruedas, andadores, bastones y estructuras a nivel de acera, las cuales suelen estar ausentes o subrepresentadas en conjuntos de datos urbanos convencionales.

Con el fin de mejorar la robustez frente a cambios de dominio, se aplicaron técnicas de aleatorización de dominio mediante la variación de la iluminación, los parámetros de la cámara, las condiciones ambientales —como escenarios despejados, nublados y lluviosos— y los atributos de los objetos, incluyendo texturas de pavimentos, fachadas, vehículos y vestimenta [18]. Esta diversidad generada de forma procedimental promueve la invariancia de características durante la transferencia de simulación a realidad, como se ilustra en la Figura 1. La Tabla 1 resume el esquema de codificación de colores empleado para el etiquetado semántico de las clases en SYNTHUA-DT.

(a) Renderización ultrarrealista de SYNTHUA-DT



(b) Máscara de segmentación semántica correspondiente

Figura 1. Ejemplo de un par SYNTHUA-DT: (a) imagen RGB de alta fidelidad; (b) máscara semántica perfecta a nivel de píxel.

Tabla 1. Esquema de codificación de colores para las clases de segmentación semántica en SYNTHUA-DT

Color	Elementos	Categoría
	Edificios	Edificio
	Silla de ruedas motorizada	Dispositivos de movilidad
	Muletas	Dispositivos de movilidad
	Andador	Dispositivos de movilidad
	Silla de ruedas	Dispositivos de movilidad
	Bastón ortopédico	Dispositivos de movilidad
	Bastón	Dispositivos de movilidad
	Muleta ortopédica	Dispositivos de movilidad
	Césped	Naturaleza
	Árboles, plantas	Naturaleza
	Personas	Transeúntes
	Perros	Transeúntes
	Bolardo, banco, papelera pública Lata, columpio, sombrilla, panel publicitario	Mobiliario urbano
	Fuente, monumentos, lugar turístico	Mobiliario urbano
	Coche, autobús, vehículos	Transporte
	Bicicleta	Transporte
	Motocicleta, scooter	Transporte
	Poste de alumbrado público	Infraestructura urbana
	Señal de límite de velocidad, señal de estacionamiento con límite de tiempo	Infraestructura urbana
	Poste de semáforo	Infraestructura urbana
	Aceras	Infraestructura urbana

2.2. Preprocesamiento de imágenes y máscaras semánticas

El preprocesamiento consta de dos etapas principales: (i) el redimensionamiento y la normalización de las imágenes RGB de entrada y (ii) la descomposición de las máscaras semánticas codificadas por color en tensores de etiquetas multicanal, utilizados como supervisión para el entrenamiento de los modelos de segmentación.

2.2.1. Redimensionamiento y normalización

Con el fin de equilibrar la eficiencia computacional y la fidelidad en la representación de objetos pequeños, todas las imágenes se redimensionaron a una resolución de 512 × 512 px utilizando la interpolación INTER_AREA de OpenCV, la cual resulta adecuada para tareas de segmentación al preservar el detalle de los bordes durante la reducción de escala [17].

Experimentos preliminares mostraron que una resolución inferior de 256 × 256 px produjo de manera consistente valores de IoU más bajos para clases de infraestructura fina, como bordillos y señalización, lo que justifica la resolución adoptada. Finalmente, las intensidades RGB se normalizaron al rango [0, 1] mediante una división por 255 [18].

2.2.2. Descomposición de máscaras semánticas

Las máscaras semánticas codificadas por color se convirtieron en un tensor binario multicanal de dimensión 512 × 512 × 22 mediante umbralización en el espacio de color HSV, aislando el rango de tono correspondiente a cada clase. Para manejar el solapamiento de tonos en el límite circular 0°/180°, se empleó un algoritmo de crecimiento de regiones, seguido de operaciones morfológicas de apertura y cierre con tamaños de kernel específicos por clase, orientadas a eliminar artefactos y reforzar la coherencia regional.

Asimismo, los componentes conectados con un área inferior a un umbral mínimo predefinido se filtraron, obteniéndose finalmente máscaras one-hot adecuadas para la supervisión durante el entrenamiento [7, 19].

El Algoritmo 1 formaliza este procedimiento. Durante el entrenamiento, el tensor multicanal de 22 clases se colapsa en un único mapa de etiquetas mutuamente excluyentes, el cual es procesado por una cabeza de segmentación con activación *softmax*. La representación multicanal intermedia se conserva con fines de diagnóstico y para posibles extensiones futuras hacia escenarios de segmentación multietiqueta.

Algoritmo 1 Descomposición semántica de máscaras basada en el color

Entrada: Máscara semántica RGB $M \in \mathbb{R}^{H\times W\times 3}$, parámetros específicos de clase $P = \{P_1, \ldots, P_{22}\}$
Salida: Tensor binario multicanal $T \in \{0,1\}^{H\times W\times 22}$

Convertir M a HSV: $M_{\text{HSV}} \leftarrow \texttt{ConvertToHSV}(M)$ Inicializar $T \leftarrow \mathbf{0} \in \{0,1\}^{H\times W\times 22}$
for $i = 1$ **to** 22 **do**
  $(l_i, u_i) \leftarrow P_i$.Umbrales HSV $a_i \leftarrow P_i$.área mínima $k_i \leftarrow P_i$.tamaño del núcleo $g_i \leftarrow P_i$.crecimiento de la región $m_i \leftarrow P_i$.operaciones morfológicas
  Aplicar umbral HSV: $B_i \leftarrow \mathbb{1}[M_{\text{HSV}} \in [l_i, u_i]]$
  Crecimiento de la región: $R_i \leftarrow \texttt{RegionGrow}(B_i, M_{\text{HSV}}, g_i)$
  Filtrado morfológico: $M_i \leftarrow \texttt{Morph}(R_i, k_i, m_i)$
  Filtrado de área: $F_i \leftarrow \mathbb{1}[\texttt{Area}(M_i) \geq a_i]$
  Almacenar resultado: $T[:,:,i] \leftarrow F_i$
return T

2.3. Arquitecturas de los modelos

Se evalúan comparativamente las siguientes arquitecturas de segmentación:

- **U-Net:** arquitectura simétrica de tipo codificador–decodificador con conexiones de salto, ampliamente utilizada en tareas de segmentación —especialmente en dominios con conjuntos de datos limitados— debido a su capacidad para preservar y recuperar detalles espaciales finos [14]
- **DeepLabv3+:** modelo de tipo codificador–decodificador que integra Atrous Spatial Pyramid Pooling (ASPP) para la captura de contexto multiescala, junto con un decodificador ligero orientado al refinamiento espacial. En esta implementación se emplea un codificador ResNet-101 con convoluciones *atrous* [15].

Esta comparación permite analizar el impacto de la agregación explícita de contexto multiescala, característica de DeepLabv3+, frente a un diseño clásico codificador–decodificador como U-Net, en el desempeño de la segmentación semántica orientada a la accesibilidad urbana.

2.4. Estrategia de entrenamiento

2.4.1. Partición del conjunto de datos y aumentación

El conjunto de datos se dividió en un 80 % para entrenamiento (4028 imágenes), un 10 % para validación (503 imágenes) y un 10 % para prueba (505 imágenes), empleando muestreo estratificado con el fin de preservar la distribución de clases en cada partición [20]. Durante el entrenamiento se aplicó aumentación de datos en línea, que incluyó escalado aleatorio (0.5–2.0),

recorte aleatorio, volteo horizontal, pequeñas rotaciones ($\pm 10°$), ligeras variaciones de color, desenfoque gaussiano y técnicas de mezcla como ClassMix y CutMix, con énfasis en las clases minoritarias.

Las transformaciones geométricas se aplicaron de manera sincrónica tanto a las imágenes RGB como a las máscaras semánticas, mientras que las transformaciones fotométricas se aplicaron exclusivamente a las imágenes RGB [15].

2.4.2. Optimización y alcance del estudio

Ambos modelos se entrenaron utilizando el optimizador Adam con una tasa de aprendizaje inicial de 1×10^{-4} y un tamaño de lote de 8. Se empleó un esquema de decaimiento escalonado que redujo la tasa de aprendizaje a 1×10^{-5} tras 75 épocas. Adicionalmente, se utilizó entrenamiento con precisión mixta para optimizar el uso de la memoria de la GPU y acelerar el proceso de entrenamiento.

Durante el entrenamiento se aplicaron técnicas estándar de control y monitoreo, incluyendo *early stopping* (parada temprana) con una paciencia de 10 épocas, guardado de puntos de control del modelo, un programador ReduceLROnPlateau con un factor de 0.5 y una paciencia de 5 épocas, así como el uso de TensorBoard para el seguimiento de las métricas.

El estudio es estrictamente computacional; no se involucraron datos de sujetos humanos ni se aplicaron protocolos clínicos. Los experimentos se repitieron utilizando tres semillas aleatorias independientes, y las métricas se reportan como valores promedio acompañados de sus respectivos intervalos de confianza.

2.5. Funciones de pérdida

Las clases semánticas se consideran mutuamente excluyentes y la cabeza final de predicción emplea una activación *softmax* de 22 vías. En consecuencia, el objetivo de entrenamiento se define mediante una función de pérdida compuesta que combina una pérdida de Entropía Cruzada balanceada por clase con una pérdida Dice suavizada, con el fin de mitigar el desbalance de clases y mejorar la superposición espacial de las regiones segmentadas, ver ecuación (1):

$$L = \lambda_{\text{CE}}\,\text{CE}_{\text{balanced}} + \lambda_{\text{Dice}}\,\text{Dice}, \qquad \lambda_{\text{CE}} = \lambda_{\text{Dice}} = 0.5. \tag{1}$$

Los pesos de clase en la pérdida CE_balanced se definieron de manera inversamente proporcional a la frecuencia de píxeles de cada clase, con el objetivo de contrarrestar el desbalance inherente del conjunto de datos. De forma alternativa, se evaluó la pérdida focal ($\gamma = 2$) como reemplazo directo del término de entropía cruzada, observándose un comportamiento global comparable y una recuperación ligeramente superior en clases minoritarias, como se reporta en la Tabla 4.

La representación multicanal one-hot, derivada de las máscaras codificadas en HSV, se utiliza internamente para la decodificación de color y las operaciones de procesamiento morfológico, mientras que el entrenamiento de los modelos se realiza empleando un único mapa de etiquetas mutuamente excluyentes. Posibles extensiones hacia escenarios multietiqueta —como el modelado conjunto de peatones y dispositivos de asistencia— pueden abordarse mediante la incorporación de cabezas binarias auxiliares, siguiendo enfoques similares a los reportados en [21].

2.6. Métricas de evaluación

El desempeño de la segmentación y la calibración del modelo se evalúan mediante métricas cuantitativas estándar, calculadas tanto a nivel de clase como a nivel global del conjunto de datos.

Intersección sobre unión (IoU) y mIoU. Para la clase c, ver ecuación (2)

$$\text{IoU}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c + \text{FN}_c}, \qquad \text{mIoU} = \frac{1}{C}\sum_{c=1}^{C} \text{IoU}_c. \tag{2}$$

Precisión, *recall*, *F1-score* y exactitud balanceada. Para la clase c, ver ecuaciones (3), (4), (5)

$$\text{Precision}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}, \qquad \text{Recall}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}, \tag{3}$$

$$\text{F1}_c = \frac{2\,\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}, \tag{4}$$

$$\text{BA}_c = \frac{1}{2}\left(\frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} + \frac{\text{TN}_c}{\text{TN}_c + \text{FP}_c}\right). \tag{5}$$

Las puntuaciones globales se promedian de forma macro sobre las clases.

Calibración y métricas probabilísticas. Para evaluar la calibración probabilística de las predicciones, se calcularon el error de calibración esperado (ECE), el error máximo de calibración (MCE), la logverosimilitud negativa (NLL) y la puntuación de Brier. En particular, el ECE y el MCE se estiman dividiendo el rango de confianza [0, 1] en B intervalos (bins), donde n_b denota el número de predicciones cuya confianza cae dentro del intervalo b, ver ecuación (6):

$$\begin{aligned} \mathrm{ECE} &= \sum_{b=1}^{B} \frac{n_b}{N} \left|\mathrm{acc}(b) - \mathrm{conf}(b)\right|, \\ \mathrm{MCE} &= \max_b \left|\mathrm{acc}(b) - \mathrm{conf}(b)\right|. \end{aligned} \tag{6}$$

Dadas las probabilidades *softmax* p_i para la clase verdadera y_i, ver ecuación (7),

$$\begin{aligned} \mathrm{NLL} &= -\frac{1}{N}\sum_{i=1}^{N} \log p_i(y_i), \\ \mathrm{Brier} &= \frac{1}{N}\sum_{i=1}^{N} \|p_i - \mathbf{1}_{y_i}\|_2^2. \end{aligned} \tag{7}$$

Detalles de implementación

Backbone y decodificador. Se empleó DeepLabv3+ con un backbone ResNet-101 preentrenado en ImageNet. El codificador utiliza un output stride de 16, mientras que el decodificador emplea un stride de 4. El módulo Atrous Spatial Pyramid Pooling (ASPP) incorpora tasas de dilatación $\{1, 6, 12, 18\}$ junto con *image-level pooling*.

Entrada y recortes. Durante el entrenamiento se extrajeron recortes aleatorios de 512 × 512 píxeles a partir de imágenes originales de 1920 × 1080 píxeles. En la fase de inferencia se aplicó redimensionamiento bilineal para restaurar la resolución espacial.

Normalización. Se calcularon estadísticas de normalización por canal, en el rango [0, 1], exclusivamente sobre la partición de entrenamiento.

Semillas y hardware. Los experimentos se realizaron utilizando semillas aleatorias $\{11, 23, 37\}$ y se ejecutaron en una GPU NVIDIA RTX 3090 (24 GB). Se empleó un tamaño de lote de 8 y entrenamiento con precisión mixta. El tiempo de entrenamiento fue de aproximadamente 5.2 h por semilla para 100 épocas.

Reproducibilidad. El conjunto de datos SYNTHUA-DT, junto con los scripts de preprocesamiento, las configuraciones de los modelos y el código de entrenamiento, se publicarán tras la aceptación del artículo en el repositorio público [22].

3. Resultados y discusión

Todos los resultados reportados corresponden a un escenario synthetic-to-synthetic, en el cual tanto el entrenamiento como la evaluación se realizaron exclusivamente sobre el conjunto de datos SYNTHUA-DT. El desempeño global y por clase se presenta junto con un análisis detallado de calibración, así como con una discusión de sus implicaciones para aplicaciones de navegación asistida, con especial énfasis en los dispositivos de movilidad y la infraestructura a nivel de acera.

3.1. Rendimiento global

La línea base U-Net alcanzó un mIoU de 0.0626 (IC del 95 %: 0.058–0.067), con valores macro-promediados de precisión, *recall* y *F1-score* de 0.1328, 0.0985 y 0.0872, respectivamente, lo que indica una capacidad limitada para capturar la estructura semántica global más allá de unas pocas clases dominantes. En contraste, DeepLabv3+ obtuvo un mIoU de 0.8400 (IC del 95 %: 0.828–0.852), junto con métricas macro-promediadas de precisión, *recall* y *F1-score* de 0.9085, 0.9145 y 0.9106, respectivamente, como se resume en la Tabla 3.

Tabla 2. Clases semánticas en SYNTHUA-DT y sus categorías de alto nivel

Clase	Elementos	Categoría
1	Edificios	Estructura
2	Silla de ruedas motorizada	Dispositivos de movilidad
3	Muletas	Dispositivos de movilidad
4	Andador	Dispositivos de movilidad
5	Silla de ruedas	Dispositivos de movilidad
6	Bastón ortopédico	Dispositivos de movilidad
7	Bastón	Dispositivos de movilidad
8	Muleta ortopédica	Dispositivos de movilidad
9	Césped	Naturaleza
10	Árboles, plantas	Naturaleza
11	Personas	Transeúntes
12	Perros	Transeúntes
13	Elementos del paisaje urbano	Mobiliario urbano
14	Lugares turísticos	Mobiliario urbano
15	Coche, autobús, vehículos	Transporte
16	Bicicleta	Transporte
17	Motocicleta, scooter	Transporte
18	Farola	Infraestructura urbana
19	Calles	Infraestructura urbana
20	Señales	Infraestructura urbana
21	Semáforos	Infraestructura urbana
22	Aceras	Infraestructura urbana

Estos resultados representan mejoras relativas de aproximadamente 13.4× en mIoU, 6.8× en precisión, 9.3× en *recall* y 10.4× en *F1-score* en comparación con U-Net. Adicionalmente, el valor de d de Cohen fue muy superior a 2 para el mIoU, confirmando un tamaño del efecto grande.

Las 22 clases semánticas se agrupan en siete categorías funcionales, como se presenta en la Tabla 2, reflejando los requisitos de accesibilidad urbana al diferenciar explícitamente dispositivos de movilidad, infraestructura, elementos naturales y clases relacionadas con peatones.

Con el fin de aislar el impacto de la arquitectura y del diseño de la función de pérdida, la Tabla 4 presenta un estudio de ablación que compara U-Net, DeepLabv3+ sin preentrenamiento sintético y DeepLabv3+ con preentrenamiento sintético, empleando pérdidas compuestas BCE–Dice y basadas en focal. DeepLabv3+ sin preentrenamiento ya supera a U-Net, alcanzando un mIoU de 0.291, lo que evidencia la ventaja de la arquitectura en términos de agregación

de contexto. La incorporación del preentrenamiento sintético incrementa el mIoU hasta 0.840 y mejora de manera sustancial el *recall* de la clase aceras, que pasa de 0.531 a 0.921. La variante basada en pérdida focal obtiene un mIoU comparable de 0.823 y presenta un *recall* ligeramente superior en algunas clases poco frecuentes; sin embargo, la combinación BCE–Dice ofrece el mejor equilibrio global entre desempeño agregado y estabilidad entre clases.

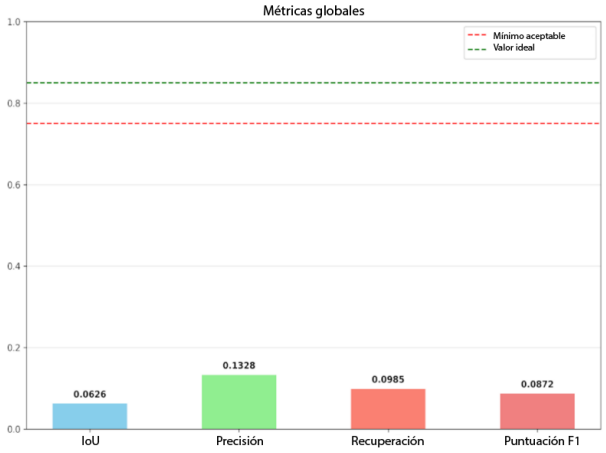
Las Figuras 2(a) y 2(b) corroboran visualmente estas tendencias: mientras que U-Net genera máscaras fragmentadas y espacialmente inconsistentes, DeepLabv3+ produce predicciones más coherentes y de mayor resolución espacial, con límites de objeto claramente definidos.

Tabla 3. Comparación del rendimiento global de U-Net y DeepLabv3+ en SYNTHUA-DT (conjunto de prueba)

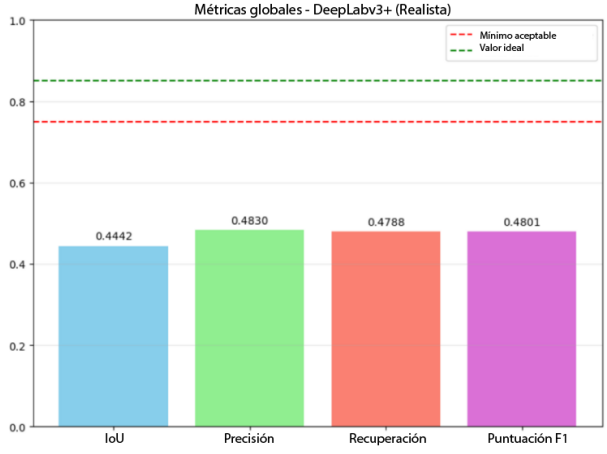
Modelo	mIoU	Precisión	Recuperación	Puntuación F1
U-Net	0.0626	0.1328	0.0985	0.0872
DeepLabv3+	0.8400	0.9085	0.9145	0.9106

Tabla 4. Estudio de ablación: impacto del preentrenamiento sintético y de las funciones de pérdida (métricas definidas en la Sección 2.6)

Modelo	mIoU	Precisión	Recuperación	Puntuación F1
U-Net	0.0626	0.1328	0.0985	0.0872
DeepLabv3+	0.8400	0.9085	0.9145	0.9106



(a) Resultados cualitativos de segmentación para U-Net



(b) Resultados cualitativos de segmentación para DeepLabv3+

Figura 2. Comparación de salidas de segmentación cualitativas para U-Net y DeepLabv3+.

3.2. Análisis por clases

Para evitar enmascarar el comportamiento de las clases minoritarias, se examinan las métricas y la prevalencia por clase.

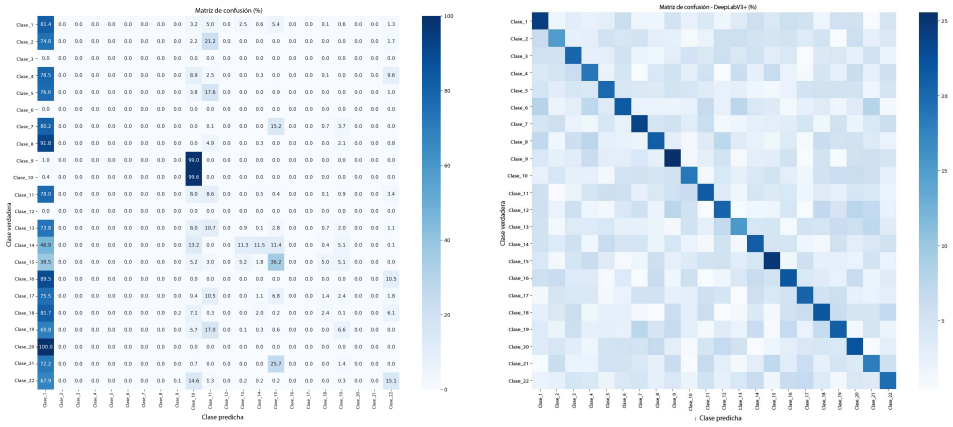
3.2.1. Prevalencia del conjunto de datos

La Tabla 5 reporta la prevalencia a nivel de píxel de cada clase en la partición de prueba, normalizada al 100 %. Las clases dominantes —incluidas edificios, calzadas y vegetación adyacente a la vía— concentran la mayor proporción de píxeles, mientras que los dispositivos de movilidad y los elementos de infraestructura de pequeña escala representan una fracción marginal del conjunto de datos.

3.2.2. Línea base U-Net

La Tabla 6 resume el desempeño por clase del modelo U-Net. El modelo muestra un rendimiento nulo en las clases correspondientes a dispositivos de movilidad (clases 2–8), con valores de IoU, precisión, *recall* y *F1-score* iguales a cero. Asimismo, las clases de aceras y varios elementos de infraestructura presentan valores muy bajos de IoU y *recall*; por ejemplo, el *recall* de la clase aceras alcanza únicamente 0.153.

La matriz de confusión mostrada en la Figura 3(a) evidencia una clasificación errónea recurrente de las clases minoritarias como fondo o como categorías dominantes, lo que es consistente con los efectos del desbalance extremo de clases reportados en [5].



(a) Matriz de confusión de U-net (b) Matriz de confusión de DeepLabv3+

Figura 3. Comparación de matrices de confusión entre U-Net y DeepLabv3+. U-Net muestra un fuerte sesgo hacia las clases dominantes, mientras que DeepLabv3+ produce predicciones más precisas y diferenciadas.

Tabla 5. Prevalencia del conjunto de datos (proporción de píxeles por clase en el conjunto de prueba; normalizada para sumar 100 %)

Clase	Prevalencia (%)
Edificios	8.28
Silla de ruedas motorizada	4.70
Muletas	1.67
Andador	6.80
Silla de ruedas	7.23
Bastón ortopédico	5.50
Bastón	7.29
Muleta ortopédica	4.88
Césped	5.13
Árboles. plantas	4.33
Personas	0.87
Perros	1.55
Elementos del paisaje urbano	0.87
Lugares turísticos	6.12
Coches. autobuses. vehículos	3.34
Bicicletas	5.01
Motocicletas. scooters	8.47
Farolas	2.78
Calles	4.14
Señales	7.17
Semáforos	2.60
Aceras	1.30

Nota: Los píxeles se calculan globalmente en la división de la prueba y se normalizan para sumar el 100 %.

Tabla 6. Métricas de rendimiento por clase para U-Net en SYNTHUA-DT (conjunto de prueba)

Clase	IoU	Precisión	Recuperación	Especificidad	Puntuación F1	Acierto equilibrado
Edificios	0.554	0.819	0.632	0.595	0.710	0.613
Silla de ruedas motorizada	0.000	0.000	0.000	1.000	0.000	0.500
Muletas	0.000	0.000	0.000	1.000	0.000	0.500
Andador	0.000	0.000	0.000	1.000	0.000	0.500
Silla de ruedas	0.000	0.000	0.000	1.000	0.000	0.500
Bastón ortopédico	0.000	0.000	0.000	1.000	0.000	0.500
Bastón	0.000	0.000	0.000	1.000	0.000	0.500
Muleta ortopédica	0.000	0.000	0.000	1.000	0.000	0.500
Césped	0.000	0.000	0.000	1.000	0.000	0.500
Árboles. plantas	0.507	0.510	0.983	0.949	0.662	0.966
Humanos	0.011	0.018	0.042	0.953	0.022	0.497
Perros	0.000	0.000	0.000	1.000	0.000	0.500
Elementos del paisaje urbano*	0.001	0.002	0.004	0.993	0.002	0.499
Lugares turísticos**	0.004	0.106	0.004	1.000	0.007	0.502
Coches. autobuses. vehículos	0.077	0.160	0.249	0.972	0.137	0.611
Bicicletas	0.000	0.000	0.000	1.000	0.000	0.500
Motocicleta. scooter	0.000	0.094	0.000	1.000	0.000	0.500
Poste de alumbrado público	0.013	0.091	0.015	0.999	0.025	0.507
Calles	0.057	0.550	0.057	0.999	0.102	0.528
Señales***	0.000	0.000	0.000	1.000	0.000	0.500
Poste de semáforo	0.000	0.000	0.000	1.000	0.000	0.500
Aceras	0.141	0.681	0.152	0.989	0.245	0.570

*Bolardo, banco, papelera pública, columpio, sombrilla, panel publicitario
**Fuente, monumentos, lugar turístico
***Señal de límite de velocidad, señal de estacionamiento con límite de tiempo
Nota: La prevalencia de la clase es independiente del modelo y se indica una vez en la tabla 5.

3.2.3. DeepLabv3+

DeepLabv3+ exhibe un desempeño elevado en las clases asociadas a dispositivos de movilidad y a la infraestructura urbana. Como se muestra en la Tabla 7 y en la Figura 5, todas las clases de ayudas de movilidad superan un IoU de 0.75, destacándose la clase silla de ruedas motorizada con un IoU de 0.94. Asimismo, las clases humanas y de animales (perros), que rara vez son detectadas por U-Net, alcanzan valores de IoU de 0.754 y 0.944, respectivamente.

Las clases de infraestructura —incluidas aceras, calles, señales, postes de semáforo y postes de alumbrado público— también superan de manera consistente un IoU de 0.75. DeepLabv3+ mantiene un equilibrio adecuado entre precisión y *recall* en las ayudas de movilidad, como se ilustra en la Figura 4, mientras que la distribución del *F1-score* presentada en la Figura 6 indica que la mayoría de las clases alcanzan valores superiores a 0.85, umbral comúnmente considerado indicativo de un desempeño apto para el despliegue.

En particular, el *recall* de la clase aceras aumenta de 0.152 con U-Net a 0.921 con DeepLabv3+, lo que representa una mejora aproximada de 6× y reduce de manera sustancial la presencia de falsos negativos en regiones críticas para la navegación asistida.

Las matrices de confusión mostradas en la Figura 3 evidencian claramente el contraste entre ambos modelos. U-Net presenta un fuerte sesgo hacia las clases dominantes, con una clasificación errónea generalizada de las clases minoritarias, que son absorbidas principalmente por el fondo o por categorías estructurales frecuentes.

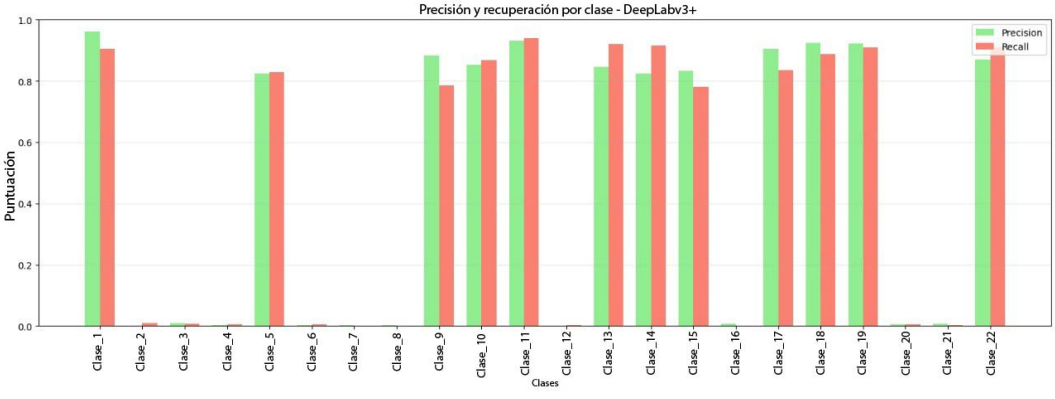


Figura 4. DeepLabv3+: precisión y *recall* por clase.

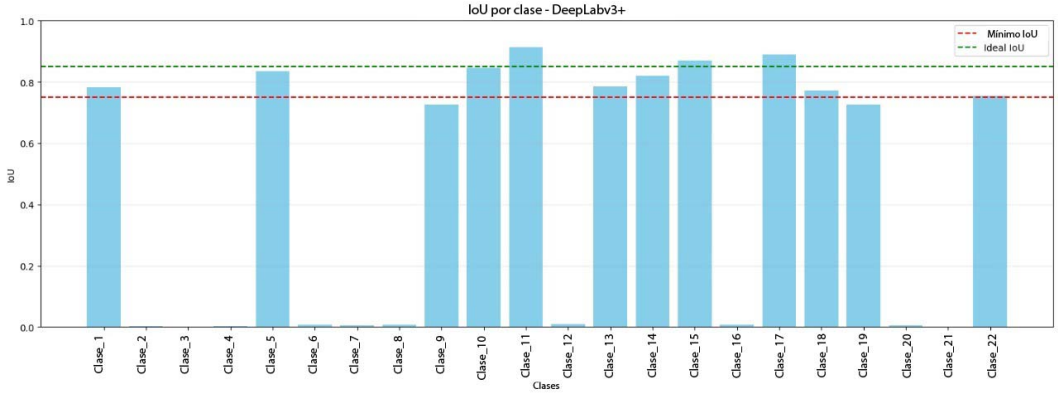


Figura 5. DeepLabv3+: IoU por clase (métricas según la Sección 2.6).

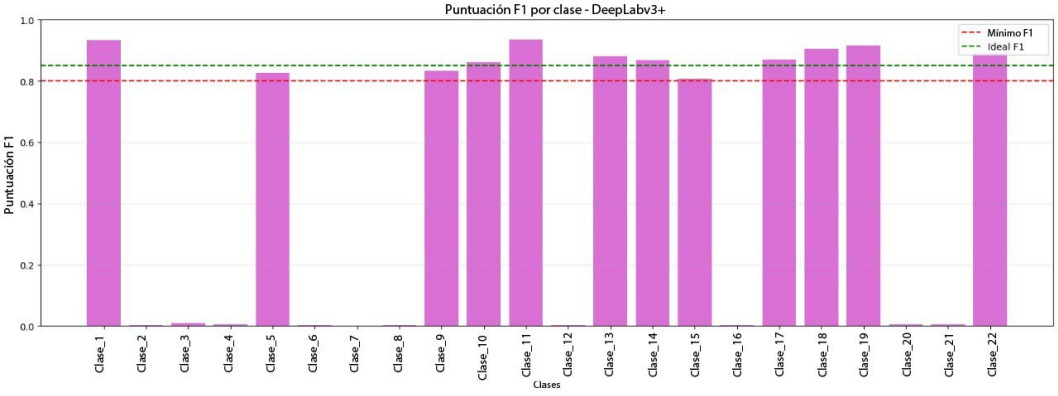


Figura 6. DeepLabv3+: *F1-score* por clase.

Tabla 7. Métricas de rendimiento por clase para DeepLabv3+ en SYNTHUA-DT (conjunto de prueba)

Clase	IoU	Precisión	Recuperación	Especificidad	Puntuación F1	Acierto equilibrado
Edificios	0,825	0,888	0,884	0,913	0,886	0,898
Silla de ruedas motorizada	0,940	0,898	0,936	0,972	0,916	0,954
Muletas	0,896	0,909	0,891	0,907	0,900	0,899
Andador	0,870	0,952	0,918	0,989	0,935	0,953
Silla de ruedas	0,781	0,876	0,921	0,970	0,898	0,945
Bastón ortopédico	0,781	0,917	0,874	0,918	0,895	0,896
Bastón	0,762	0,927	0,976	0,900	0,951	0,938
Muleta ortopédica	0,923	0,856	0,951	0,973	0,901	0,962
Césped	0,870	0,929	0,972	0,964	0,950	0,968
Árboles, plantas	0,892	0,872	0,966	0,966	0,917	0,966
Humanos	0,754	0,858	0,928	0,969	0,892	0,949
Perros	0,944	0,973	0,970	0,907	0,972	0,938
Elementos del paisaje urbano*	0,916	0,976	0,862	0,932	0,915	0,897
Lugares turísticos**	0,792	0,955	0,875	0,910	0,914	0,893
Coches, autobuses, vehículos	0,786	0,890	0,856	0,978	0,872	0,917
Bicicletas	0,787	0,863	0,892	0,956	0,877	0,924
Motocicleta, scooter	0,811	0,939	0,901	0,930	0,919	0,915
Poste de alumbrado público	0,855	0,907	0,885	0,906	0,896	0,895
Calles	0,836	0,866	0,958	0,928	0,909	0,943
Señales de tráfico***	0,808	0,914	0,896	0,929	0,905	0,913
Postes de semáforos	0,872	0,854	0,887	0,966	0,870	0,926
Aceras	0,778	0,968	0,921	0,957	0,944	0,939

*Bolardo, banco, papelera pública, columpio, sombrilla, panel publicitario
**Fuente, monumentos, lugar turístico
***Señal de límite de velocidad, señal de estacionamiento con límite de tiempo
Nota: La prevalencia de la clase es independiente del modelo y se indica una vez en la Tabla 5.

3.3. Calibración del modelo

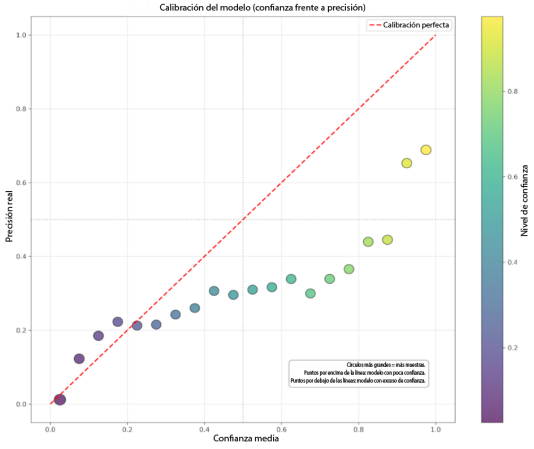
La exactitud por sí sola es insuficiente para sistemas críticos para la seguridad; las estimaciones de confianza también deben ser fiables.

3.3.1. U-Net

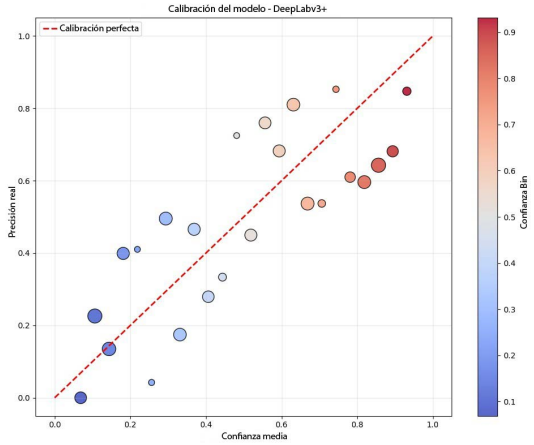
La Figura 7(a) muestra el diagrama de fiabilidad del modelo U-Net. La curva se sitúa por debajo de la línea de calibración ideal para niveles de confianza superiores a 0.5, lo que indica una marcada sobreconfianza. En particular, las predicciones con una confianza reportada entre 0.6 y 0.8 corresponden a una precisión empírica de solo 0.2–0.4. Esta deficiente calibración resulta especialmente problemática para aplicaciones de navegación asistida, donde errores asociados a altos niveles de confianza pueden conducir a decisiones de orientación inseguras.

3.3.2. DeepLabv3+ y escalado de temperatura

DeepLabv3+ presenta una calibración notablemente superior en comparación con U-Net, aunque persiste un grado moderado de sobreconfianza en los niveles más altos de confianza, como se observa en la Figura 7(b). La Tabla 8 resume cuantitativamente las métricas de calibración obtenidas antes y después de aplicar el escalado de temperatura.



(a) Curva de calibración de U-Net



(b) Curva de calibración de DeepLabv3+ (antes/después del escalado de temperatura)

Figura 7. Comparación de la calibración del modelo para U-Net y DeepLabv3+. Las curvas relacionan la confianza predicha con la precisión empírica (véanse las definiciones de ECE/MCE en la Sección 2.6).

Tabla 8. Métricas de calibración para DeepLabv3+ en el conjunto de prueba (media ± IC del 95 % mediante bootstrap a nivel de imagen, 10 000 iteraciones)

Configuración	ECE (%)	MCE (%)	NLL	Brier
Precalibración	8.5 ± 0.7	23.1 ± 1.9	0.693 ± 0.018	0.162 ± 0.004
Escalado de temperatura	**3.3** ± 0.5	**9.8** ± 1.3	**0.612** ± 0.015	**0.148** ± 0.003

El escalado de temperatura reduce el ECE en aproximadamente un 61 % y el MCE en un 58 %, y mejora asimismo la NLL y la puntuación de Brier. La sobreconfianza queda en gran medida confinada al intervalo de mayor confianza (> 0.8), y la incidencia de clasificaciones erróneas con alta confianza es sustancialmente menor que en U-Net. Esta mejora resulta especialmente relevante para los módulos posteriores que deben tomar decisiones conscientes del riesgo.

3.4. Clases listas para el despliegue e impacto práctico

La Tabla 9 resume las clases que cumplen el umbral de preparación para el despliegue, definido como IoU ≥ 0.75. Con DeepLabv3+, las 22 clases superan este umbral, incluidas las correspondientes a dispositivos de movilidad, infraestructura, elementos naturales y transeúntes. En contraste, ninguna de las clases alcanza este nivel de desempeño con U-Net.

Tabla 9. Clases que cumplen el umbral de preparación para el despliegue (IoU $\geq$ 0,75) con DeepLabv3+

Categoría	Umbral de coincidencia de clases	Rango IoU
Dispositivos de movilidad	7/7 clases	0.762-0.940
Infraestructura urbana	5/5 clases	0.778-0.872
Naturaleza	2/2 clases	0.870–0.892
Estructura	1/1 clase	0.825
Transeúntes	2/2 clases	0.754–0.944
Mobiliario urbano	2/2 clases	0.792–0.916
Transporte	3/3 clases	0.786–0.811

Todas las clases de dispositivos de movilidad —incluidas sillas de ruedas, andadores, bastones, muletas y variantes ortopédicas— alcanzan valores de IoU entre 0.762 y 0.940, lo que indica una detección fiable a través de distintos puntos de vista y condiciones de iluminación. La clase aceras alcanza un IoU de 0.778 con un *recall* de 0.921, en comparación con un *recall* de 0.153 obtenido por U-Net, lo que supone una reducción de los falsos negativos de aceras superior al 80 %. Desde una perspectiva aplicada, este resultado disminuye de forma significativa la probabilidad de sugerir terreno no transitable como aceras a usuarios de sillas de ruedas.

4. Limitaciones y trabajo futuro

A pesar de las sólidas mejoras alcanzadas con SYNTHUA-DT y DeepLabv3+, deben reconocerse varias limitaciones:

- **Evaluación exclusivamente sintética y brecha de dominio.** Todos los experimentos se realizaron en un escenario sintético a sintético; el desempeño en condiciones reales —con variaciones de iluminación, desenfoque por movimiento, ruido del sensor y oclusiones— permanece sin evaluar. El trabajo futuro incorporará conjuntos de datos reales orientados a la accesibilidad y aplicará técnicas de adaptación de dominio, como alineación adversaria, transferencia de estilo y autoentrenamiento, para reducir la brecha entre lo sintético y lo real.

- **Detalles de clases de muy baja prevalencia y precisión en los límites.** Los errores residuales se concentran en estructuras delgadas, como las puntas de bastones y los radios de las sillas de ruedas, así como en las transiciones calle–acera, donde el IoU de contorno (≈ 0.68) se mantiene por debajo del umbral objetivo de 0.75. Se explorarán pérdidas sensibles a bordes, módulos de atención centrados en límites y recortes de mayor resolución para mejorar la consistencia geométrica de grano fino.

- **Diversidad de escenas y escenarios de accesibilidad.** Las escenas actuales se centran en entornos exteriores con un conjunto fijo de ayudas de movilidad y tipos de infraestructura. Escenarios relevantes —como transiciones interiores entre rampas y ascensores, obstáculos temporales (obras) y cruces congestionados— no están representados. Las futuras extensiones de SYNTHUA-DT incorporarán disposiciones más diversas, flujos dinámicos de peatones y configuraciones de accesibilidad poco frecuentes.

- **Familias de modelos y aprendizaje multitarea.** Solo se evaluaron U-Net y DeepLabv3+. No se analizaron otras familias de modelos, incluidos decodificadores basados en transformers, respaldos híbridos CNN–Transformer ni modelos ligeros en tiempo real. Dado que los sistemas operativos suelen requerir estimación conjunta de profundidad, segmentación por instancias o detección de rampas de bordillo, el trabajo futuro explorará arquitecturas multitarea que equilibren precisión, calibración y rendimiento en tiempo real.

- **Calibración y decisiones conscientes de la incertidumbre.** Incluso tras el escalado de temperatura, DeepLabv3+ mantiene una leve sobreconfianza a niveles altos de probabilidad. Investigaciones futuras integrarán métodos conscientes de la incertidumbre —incluidos ensamblados, dropout Monte Carlo y aprendizaje profundo evidencial— junto con reglas de decisión sensibles al riesgo, para que la planificación de rutas y las alertas de obstáculos consideren explícitamente la incertidumbre de la segmentación.

En conjunto, SYNTHUA-DT y el marco de entrenamiento propuesto constituyen un primer paso hacia la segmentación semántica sintética orientada a la accesibilidad. Los trabajos futuros combinarán generación de datos sintéticos, recopilación de datos del mundo real, técnicas avanzadas de adaptación de dominio y modelado consciente de la incertidumbre, con el objetivo de desarrollar módulos de percepción robustos para la navegación urbana inclusiva.

5. Conclusiones

Este estudio demuestra que la generación de datos sintéticos de alta fidelidad mediante Unreal Engine 5.1, combinada con un entrenamiento consciente del desbalance de clases y una arquitectura moderna de tipo codificador–decodificador, puede mejorar de forma sustancial la segmentación semántica en escenarios de accesibilidad urbana. Utilizando el conjunto de datos SYNTHUA-DT, el preentrenamiento de DeepLabv3+ con 5036 imágenes anotadas produjo un incremento de 13.4× en el mIoU global, de 0.0626 a 0.84, junto con mejoras aproximadas de 6.8× en precisión, 9.3× en recall y 10.4× en F1-score, en comparación con la línea base U-Net.

A nivel de clase, DeepLabv3+ detectó con éxito todas las categorías críticas para la accesibilidad, alcanzando un IoU ≥ 0.75 para cada ayuda de movilidad presente en SYNTHUA-DT. Las sillas de ruedas motorizadas alcanzaron un IoU de 0.94, mientras que las sillas de ruedas convencionales y los andadores obtuvieron valores de IoU de 0.78 y 0.87, respectivamente. El recall en la detección de aceras aumentó de 0.153 con U-Net a 0.921 con DeepLabv3+, lo que redujo los falsos negativos de aceras en más del 80 % y mejoró de manera significativa la fiabilidad en la identificación de trayectorias para la navegación asistida. En conjunto, las 22 clases semánticas superaron el umbral de IoU de 0.75, lo que indica un rendimiento sólido y consistente tanto en clases dominantes como minoritarias.

El análisis de calibración mostró que DeepLabv3+ con escalado de temperatura reduce el error de calibración esperado y el error máximo de calibración en aproximadamente un 60 %, disminuyendo el riesgo de clasificaciones erróneas con alta confianza en decisiones críticas para la seguridad. Esta combinación de un IoU elevado por clase y una calibración probabilística mejorada resulta especialmente relevante para los módulos posteriores que deben razonar sobre la seguridad de las rutas y la evitación de obstáculos bajo incertidumbre.

Persisten errores residuales en estructuras delgadas y en los límites, como las puntas de bastones, los radios de las sillas de ruedas y las transiciones de bordillo, para los cuales las métricas sensibles a contornos permanecen por debajo de 0.75. Como se discute en la Sección 4, el trabajo futuro abordará estas limitaciones mediante objetivos conscientes de los bordes, un muestreo sintético enriquecido de configuraciones poco frecuentes, familias de modelos más amplias —incluidas arquitecturas híbridas CNN–Transformer— y una adaptación explícita de dominio a imágenes urbanas del mundo real.

En contraste con los conjuntos de datos sintéticos orientados principalmente a la conducción autónoma, el marco SYNTHUA-DT modela y evalúa explícitamente las ayudas de movilidad y la infraestructura a nivel de acera, proporcionando un recurso centrado en la accesibilidad y un punto de referencia reproducible para futuras investigaciones en navegación urbana inclusiva y percepción para ciudades inteligentes.

Rol de autores

- **Santiago Felipe Luna Romero:** conceptualización, curación de datos, análisis formal, investigación, metodología, desarrollo de software, supervisión, validación, visualización, redacción - revisión y edición.
- **Renato Gouveia:** investigación, desarrollo de software, procesamiento de datos, redacción - borrador original.
- **Mauren Abreu de Souza:** administración de proyecto, obtención de financiación, recursos, supervisión.

Referencias

[1] M. Ivanovs, K. Ozols, A. Dobrajs, and R. Kadikis, "Improving semantic segmentation of urban scenes for self-driving cars with synthetic images," *Sensors*, vol. 22, no. 6, p. 2252, Mar. 2022. [Online]. Available: http://doi.org/10.3390/s22062252

[2] E. Mohamed, K. Sirlantzis, and G. Howells, "Indoor/outdoor semantic segmentation using deep learning for visually impaired wheelchair users," *IEEE Access*, vol. 9, pp. 147 914–147 932, 2021. [Online]. Available: http://doi.org/10.1109/access.2021.3123952

[3] R. Azad, M. Heidary, K. Yilmaz, M. Hüttemann, S. Karimijafarbigloo, Y. Wu, A. Schmeink, and D. Merhof, "Loss functions in the era of semantic segmentation: A survey and outlook," *arXiv preprint*, 2023. [Online]. Available: http://doi.org/10.48550/ARXIV.2312.05391

[4] J. L. Gómez, M. Silva, A. Seoane, A. Borrás, M. Noriega, G. Ros, J. A. Iglesias-Guitian, and A. M. López, "All for one, and one for all: Urbansyn dataset, the third musketeer of synthetic driving scenes," 2023. [Online]. Available: http://doi.org/10.48550/ARXIV.2312.12176

[5] J. Tian, N. Mithun, Z. Seymour, H.-P. Chiu, and Z. Kira, "Striking the right balance: Recall loss for semantic segmentation," *arXiv preprint*, 2021. [Online]. Available: http://doi.org/10.48550/ARXIV.2106.14917

[6] Z. Song, Z. He, X. Li, Q. Ma, R. Ming, Z. Mao, H. Pei, L. Peng, J. Hu, D. Yao, and Y. Zhang, "Synthetic datasets for autonomous driving: A survey," 2023. [Online]. Available: http://doi.org/10.48550/ARXIV.2304.12205

[7] R. Kamimura, "Information-theoretic enhancement learning and its application to visualization of self-organizing maps," *Neurocomputing*, vol. 73, no. 13–15, pp. 2642–2664, Aug. 2010. [Online]. Available: http://doi.org/10.1016/j.neucom.2010.05.013

[8] Q. Wu and H. Liu, "Unsupervised domain adaptation for semantic segmentation using depth distribution," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 14 374–14 387. [Online]. Available: https://upsalesiana.ec/ing35ar9r1

[9] S. F. Luna-Romero, C. R. Stempniak, M. Abreu de Souza, and G. Reynoso-Meza, *Urban Digital Twins for Synthetic Data of Individuals with Mobility Aids in Curitiba, Brazil, to Drive Highly Accurate AI Models for Inclusivity.* Springer Nature Switzerland, 2024, pp. 116–125. [Online]. Available: http://doi.org/10.1007/978-3-031-52090-7_12

[10] Y. Yuan, Y. Du, Y. Ma, and H. Lv, "DSC-Net: enhancing blind road semantic segmentation with visual sensor using a dual-branch Swin-CNN architecture," *Sensors*, vol. 24, no. 18, p. 6075, Sep. 2024. [Online]. Available: http://doi.org/10.3390/s24186075

[11] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Álvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *arXiv preprint*, 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2105.15203

[12] S. F. Luna Romero, C. R. Stempniak, M. Abreu de Souza, and G. Reynoso-Meza, "A transfer learning model proposal for country border security using aerial thermal images," in *Procedings do XXIV Congresso Brasileiro de Automática*, ser. CBA2022. SBA Sociedade Brasileira de Automática, Oct. 2022. [Online]. Available: http://doi.org/10.20906/cba2022/3341

[13] S. F. L. Romero, M. A. d. Souza, and L. S. Andrade, "Synthua-dt: A methodological framework for synthetic dataset generation and automatic annotation from digital twins in urban accessibility applications," *Technologies*, vol. 13, no. 8, p. 359, Aug. 2025. [Online]. Available: http://doi.org/10.3390/technologies13080359

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *arXiv preprint*, 2015. [Online]. Available: http://doi.org/10.48550/ARXIV.1505.04597

[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *arXiv preprint*, 2018. [Online]. Available: http://doi.org/10.48550/ARXIV.1802.02611

[16] S. F. Luna-Romero, M. Abreu de Souza, and L. Serpa Andrade, "Artificial vision systems for mobility impairment detection: Integrating synthetic data, ethical considerations, and real-world applications," *Technologies*, vol. 13, no. 5, p. 198, May 2025. [Online]. Available: http://doi.org/10.3390/technologies13050198

[17] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," *arXiv preprint*, 2018. [Online]. Available: http://doi.org/10.48550/ARXIV.1804.06516

[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015. [Online]. Available: http://doi.org/10.48550/ARXIV.1502.03167

[19] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," *arXiv preprint*, 2017. [Online]. Available: http://doi.org/10.48550/ARXIV.1703.06870

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *arXiv preprint*, 2017. [Online]. Available: http://doi.org/10.48550/ARXIV.1708.02002

[21] J. Brewer, K. Rajagopal, A. Sadofyev, and W. van der Schee, "Evolution of the mean jet shape and dijet asymmetry distribution of an ensemble of holographic jets in strongly coupled plasma," *Journal of High Energy Physics*, vol. 2018, no. 2, Feb. 2018. [Online]. Available: http://doi.org/10.1007/jhep02(2018)015

[22] R. Gouveia. (2025) Pibiti semantic segmentation. Github, Inc. [Online]. Available: https://upsalesiana.ec/ing35ar9r3