Scientific Paper / Artículo Científico

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

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

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

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

Received: 11-07-2025, Received after review: 01-12-2025, Accepted: 09-12-2025, Published: 01-01-2026

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.

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

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

Suggested citation: S. F. Luna Romero, R. Gouveia and M. Abreu de Souza, "Enhancing semantic segmentation for urban accessibility using high-fidelity synthetic data," *Ingenius, Revista de Ciencia y Tecnología*, N.° 35, pp. 114-127, 2026, DOI: https://doi.org/10.17163/ings.n35.2026.09.

1. Introduction

Semantic segmentation, assigning a semantic label to every pixel in an image, is a core component for understanding complex urban scenes. It supports autonomous driving, traffic monitoring, augmented reality, and assistive navigation for pedestrians with mobility or visual impairments [1,2]. By distinguishing roads, sidewalks, buildings, vehicles, pedestrians, and accessibility-related landmarks, such as curb ramps and mobility aids, segmentation models provide the spatial awareness required for transportation planning, inclusive urban design, and real-time obstacle detection in assistive devices [1,2].

However, state-of-the-art deep neural networks (DNNs) remain highly dependent on large, pixel-level annotated datasets. Producing such labels is expensive and time-consuming, especially for fine structures and rare classes. This burden is particularly severe for safety-critical but underrepresented categories, such as pedestrians using wheelchairs, walkers, or canes, whose limited presence in existing datasets constrains generalization in real urban environments [3,4]. Mainstream urban scene datasets, including Cityscapes and KITTI, contain very few or no instances of mobility-impaired users, which leads to systematic biases in which models frequently miss these classes or absorb them into background [3,5].

High-fidelity simulation has emerged as an effective approach to mitigate reliance on large real-world datasets. Modern game engines, including Unreal Engine, and simulators, such as CARLA, can generate large volumes of photorealistic images with automatically rendered, pixel-accurate masks, substantially reducing annotation costs and enabling controlled experimentation on rare or safety-critical categories [1,4,6]. Pre-training segmentation networks on synthetic imagery and fine-tuning on smaller real datasets have been shown to improve performance in urban scenes [1,7]. Domain randomization, systematically varying lighting, textures, weather, camera viewpoints, and object configurations, further reduces overfitting to simulator-specific artifacts and enhances robustness for sim-to-real transfer [8,9].

Two issues remain central to urban accessibility and assistive navigation. First, pronounced class imbalance persists: dominant classes such as roads, sky, and buildings occupy most pixels, whereas mobility devices and narrow infrastructure elements, including curb ramps, bollards, and traffic signs, account for only a tiny fraction [3,5]. Without specific countermeasures, models tend to overfit majority classes and underfit minority categories that are critical for accessibility. Loss functions that reweight hard-to-classify or underrepresented examples, such as Focal loss and Tversky loss, along with class-aware augmentation, are commonly used to mitigate this effect. Second, segmentation architectures must balance global context and fine-grained detail. Encoder– decoder CNNs such as U-Net and DeepLab capture local structure but have limited receptive fields, whereas Transformer-based models offer strong global reasoning but may struggle to preserve precise boundaries. Hybrid CNN–Transformer architectures that combine convolutional backbones with self-attention modules aim to capture long-range dependencies while preserving detailed contours in cluttered scenes [10–12].

Most synthetic datasets and segmentation pipelines focus on autonomous driving and do not explicitly target accessibility. Their taxonomies emphasize generic traffic participants and coarse infrastructure, and they rarely include detailed labels for mobility aids or sidewalk-level elements. As a result, they are not directly suited to evaluating accessibility-oriented segmentation or supporting assistive navigation systems.

In contrast, the SYNTHUA-DT (Synthetic Urban Accessibility – Digital Twin) dataset [13] explicitly focuses on urban accessibility. It models a realistic urban environment in Unreal Engine 5.1 and provides pixel-perfect semantic annotations for 22 classes, including multiple categories of mobility devices such as wheelchairs, walkers, and canes, as well as pedestrians, and sidewalk-level infrastructure. SYNTHUA-DT is designed to address this gap in accessibility-oriented data by offering a controllable corpus in which mobility aids and sidewalk structures are systematically represented.

1.1. Related Work

Synthetic data for urban scene understanding has been explored using both commercial games and dedicated simulators. Kamimura et al. extracted dense annotations from GTA-V and demonstrated that synthetic pre-training can improve performance when combined with real-world fine-tuning [7]. CARLA-based pipelines have been used to augment Cityscapes with additional traffic scenarios and adverse weather conditions, thereby enhancing robustness to rare configurations [1]. Digital twin environments , such as UrbanSyn, further combine realistic 3D city models with domain-adaptation and style-transfer techniques to reduce the gap betwee simulated and real imagery [4,6]. These studies consistently report that synthetic data enhances accuracy and generalization, but their label spaces primarily target generic traffic participants and do not systematically address accessibility-related elements.

Class imbalance and small-object segmentation have also been studied extensively. Azad et al. And Liu et al. analyzed how long-tailed label distributions degrade performance on minority classes and evaluated loss functions such as Focal and Tversky loss to reweight hard-to-classify or underrepresented exam-

ples [3, 5]. U-Net and DeepLabv3+ are widely used baselines due to their balance between accuracy and computational cost [14, 15]. Hybrid CNN–Transformer approaches, including Swin Transformer encoders and transformer augmented decoders, have been proposed to better capture long-range context while preserving structural details in complex scenes [10–12]. However, most evaluations rely on datasets in which mobility-impaired pedestrians and accessibility infrastructure are absent or severely underrepresented, limiting their applicability to assistive navigation.

1.2. Contributions

Building on SYNTHUA-DT, this work investigates whether high-fidelity synthetic data and imbalanceaware training are sufficient to achieve deploymentready segmentation performance for accessibility-critical classes, and it quantifies the gains relative to a U-Net baseline. The main contributions are as follows:

- **Accessibility-oriented synthetic dataset usage.** The SYNTHUA-DT dataset [13], generated with Unreal Engine 5.1, provides 5,036 high-resolution images with pixel-perfect annotations across 22 classes, explicitly including multiple mobility devices and sidewalk-level infrastructure.
- **Preprocessing and dataset structuring pipeline.** A pipeline is introduced that converts color-coded masks into multi-channel supervision, applies simple class-aware morphological refinements, and produces training, validation, and test splits to support robust training and evaluation.
- **Benchmarking of segmentation architectures.** U-Net and DeepLabv3+ are trained and evaluated using imbalance-aware loss functions and data augmentation. The analysis includes global and class-wise metrics, with emphasis on mobility aids and sidewalks.
- **Framework toward deployment in accessibility systems.** The synthetic pipeline, together with calibration analysis, supports future extensions through domain adaptation to real-world datasets and integration into smart-city and assistive navigation systems.

Section 2 details the methodology, Section 3 presents the experimental results and class-wise analysis, Section 4 discusses limitations and future work, and Section 5 summarizes the main findings.

2. Materials and Methods

The proposed approach comprises five components: synthetic dataset generation, preprocessing of images and semantic masks, model architectures, training strategy (including loss design), and evaluation metrics. Together, these elements define a pipeline for training and assessing segmentation models on accessibility-oriented urban scenes.

2.1. Synthetic Dataset Generation

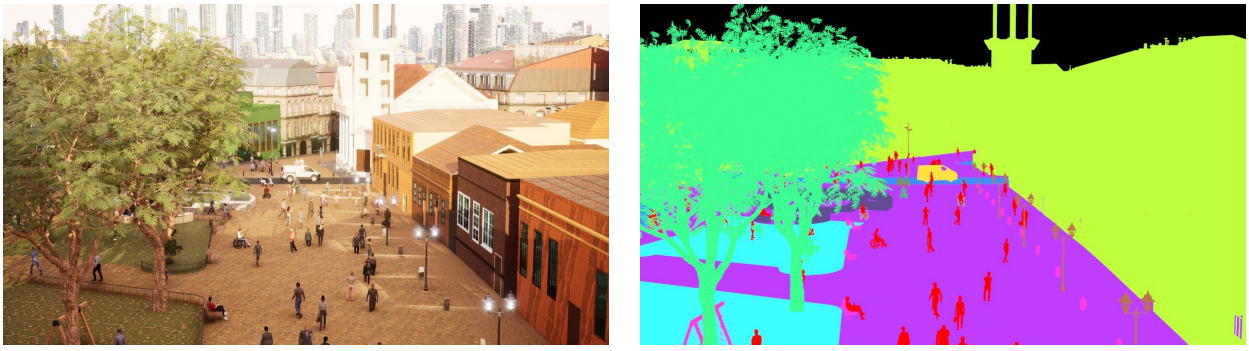
This work relies on the SYNTHUA-DT (Synthetic Urban Accessibility – Digital Twin) dataset [13, 16], a synthetic corpus of 5 036 high-resolution urban images (1920 × 1080 px) generated with Unreal Engine 5.1. A physically based rendering pipeline was used to approximate real-world illumination and material properties. Each scene includes diverse architectural layouts, such as historic façades and modern high-rises, along with stree tfurniture, dynamic actors (pedestrians, cyclists, vehicles), and multiple weather conditions.

Pixel-perfect semantic masks for 22 classes, including sidewalks, roads, crosswalks, pedestrians, mobility devices, vegetation, vehicles, signage, and other infrastructure, were automatically generated at render time [16, 17]. The class taxonomy was explicitly designed to highlight accessibility-related elements, including multiple categories of wheelchairs, walkers, canes, and sidewalk structures, which are rarely represented in conventional urban datasets.

To enhance domain robustness, domain randomization was applied by varying illumination, camera parameters, environmental factors such as clear, overcast, and rainy conditions, and object attributes, including textures for pavement, façades, vehicles, and clothing [18]. This procedurally generated diversity promotes feature invariance during sim-to-real transfer, as illustrated in Figure 1. Table 1 presents the color encoding scheme used for semantic class labeling in SYNTHUA-DT.

Table 1. Color encoding scheme for semantic segmentation classes in SYNTHUA-DT

Color	Elements	Category
	Buildings	Building
	Motorized Wheelchair	Mobility Devices
	Crutch	Mobility Devices
	Walker	Mobility Devices
	Wheelchair	Mobility Devices
	Orthopedic Cane	Mobility Devices
	Cane	Mobility Devices
	Orthopedic Crutch	Mobility Devices
	Grass	Nature
	Tree, Plants	Nature
	Humans	Passerby
	Dogs	Passerby
	Bollard, Bench, Public Trash Can, Swing, Parasol, Advertising Panel	Street Furniture
	Fountain, Monuments, Tourist Spot	Street Furniture
	Car, Bus, Vehicles	Transport
	Bike	Transport
	Motorcycle, Scooter	Transport
	Street Light Pole	Urban Infrastructure
	Streets	Urban Infrastructure
	Speed limit sign, Time-limited parking sign	Urban Infrastructure
	Traffic Light Pole	Urban Infrastructure
	Sidewalks	Urban Infrastructure



(a) Ultra-realistic render from SYNTHUA-DT (b) Corresponding semantic segmentation mask

Figure 1. Example SYNTHUA-DT pair: (a) high-fidelity RGB image; (b) pixel-perfect semantic mask.

2.2. Image and Semantic Mask Preprocessing

Preprocessing comprises two stages: (i) resizing and normalization of RGB images and (ii) decomposition of color-coded semantic masks into multi-channel label tensors.

2.2.1. Resizing and Normalization

To balance computational efficiency and small-object fidelity, all images were downsampled to 512 × 512 px using OpenCV's INTER_AREA interpolation, which preserves edge detail for segmentation tasks [17]. A lower resolution of 256×256 px consistently yielded lower IoU for curb and signage classes, justifying the chosen resolution. RGB intensities were normalized to the range [0, 1] by dividing by 255 [18].

2.2.2. Semantic Mask Decomposition

Color-coded masks were converted into a multichannel binary tensor of shape (512 × 512 × 22) using HSV-based thresholding to isolate each class's hue range. A region-growing algorithm handled hue wraparound at the 0°/180° boundary, followed by morphological opening and closing with class-specific kernel sizes to remove artifacts and enforce region coherence. Small connected components below a minimum area threshold were filtered out, yielding one-hot masks for supervision [7, 19].

Algorithm 1 formalizes this procedure. For training, the 22-channel tensor is collapsed into a single, mutually exclusive 22-class label map used by a softmax segmentation head. The intermediate representation is retained for diagnostics and potential multi-label extensions.

Algoritmo 1 Color-Based Semantic Mask Decomposition

Input: RGB semantic mask $M \in \mathbb{R}^{H\times W\times 3}$, class-specific parameters $P = \{P_1, \ldots, P_{22}\}$
Output: Multichannel binary tensor $T \in \{0,1\}^{H\times W\times 22}$

Convert M to HSV: $M_{\text{HSV}} \leftarrow \texttt{ConvertToHSV}(M)$ Initialize $T \leftarrow \mathbf{0} \in \{0,1\}^{H\times W\times 22}$
for $i = 1$ **to** 22 **do**
  $(l_i, u_i) \leftarrow P_i$.HSV thresholds $a_i \leftarrow P_i$.min area $k_i \leftarrow P_i$.kernel size $g_i \leftarrow P_i$.region growth $m_i \leftarrow P_i$.morph operations
  Apply HSV thresholding: $B_i \leftarrow \mathbb{1}[M_{\text{HSV}} \in [l_i, u_i]]$
  Region growing: $R_i \leftarrow \texttt{RegionGrow}(B_i, M_{\text{HSV}}, g_i)$
  Morphological filtering: $M_i \leftarrow \texttt{Morph}(R_i, k_i, m_i)$
  Area filtering: $F_i \leftarrow \mathbb{1}[\texttt{Area}(M_i) \geq a_i]$
  Store result: $T[:, :, i] \leftarrow F_i$
return T

2.3. Model Architectures

The following architectures are benchmarked:

- **U-Net:** A symmetric encoder–decoder architecture with skip connections, widely used for biomedical and small-dataset segmentation due to its strong boundary recovery [14]
- **DeepLabv3+:** An encoder–decoder model that combines Atrous Spatial Pyramid Pooling (ASPP) for multi-scale context with a lightweight decoder for spatial refinement, using a ResNet-101 encoder with atrous convolutions [15].

This comparison clarifies the impact of multi-scale context aggregation (DeepLabv3+) versus a classical encoder–decoder design (U-Net) in accessibilityfocused segmentation.

2.4. Training Strategy

2.4.1. Dataset Splitting and Augmentation

The dataset was divided into 80% training (4,028 images), 10% validation (503 images), and 10% test (505 images) using stratified sampling to preserve the class distribution across splits [20]. Online augmentation during training included random scaling (0.5–2.0), random cropping, horizontal flipping, small rotations ($\pm 10°$), mild color jitter, Gaussian blur, and ClassMix/-CutMix with emphasis on minority classes. Geometric transforms were applied synchronously to RGB images and masks, whereas photometric transforms were applied to RGB images only [15].

2.4.2. Optimization and Study Scope

Both models were trained using Adam with an initial learning rate of 1×10^{-4} and a batch size of 8. A step-decay schedule reduced the learning rate to 1×10^{-5} after 75 epochs. Mixed-precision training was used to optimize GPU memory usage and training speed. Early stopping with a patience of 10 epochs, model checkpointing, ReduceLROnPlateau with a factor of 0.5 and a patience of 5 epochs, and TensorBoard were employed.

The study is strictly computational; no human-subject data or clinical protocols were involved. Experiments were repeated using three random seeds, and metrics are reported as averages with confidence intervals.

2.5. Loss Functions

Semantic classes are mutually exclusive, and the final prediction head uses a 22-way softmax. The primary training objective combines a class-balanced Cross-Entropy loss with a soft Dice loss, see equation (1):

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

Class weights in CE_balanced are set inversely proportional to pixel frequency to counteract class imbalance. Focal loss ($\gamma = 2$) was additionally benchmarked as a drop-in replacement for the Cross-Entropy term, yielding similar global behavior but slightly higher recall for minority classes, as reported in Table 4. The HSV onehot tensor is used internally for color decoding and morphological processing, whereas; training uses a single label map. Potential multi-label extensions, such as jointly modeling humans and assistive gear, can be modeled via auxiliary binary heads [21].

2.6. Evaluation Metrics

Segmentation performance and calibration are evaluated using standard metrics at both the class and dataset level.

Intersection over Union (IoU) and mIoU. For class c, see equation (2)

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

Precision, Recall, F1-score, and Balanced Accuracy. For class c, see equations (3), (4), (5)

$$\begin{aligned} \text{Precision}_c &= \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}, \\ \text{Recall}_c &= \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}, \end{aligned} \tag{3}$$

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

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

Global scores are macro-averaged over classes.

Calibration and probabilistic metrics. Expected Calibration Error (ECE), Maximum Calibration Error (MCE), Negative Log-Likelihood (NLL), and Brier score are computed. With B confidence bins and nb predictions in bin b, see equation (6):

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

Given softmax probabilities p_i for true class y_i, see equation (7),

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

2.7. Implementation Details

Backbone and decoder: DeepLabv3+ with a ResNet-101 backbone pretrained on ImageNet, an encoder output stride of 16 a decoder stride of 4, and ASPP dilation rates of $\{1, 6, 12, 18\}$ with image-level pooling.

Input and crops: Random 512 × 512 crops were extracted from 1920 × 1080 images, and bilinear resizing was applied at inference.

Normalization: Per-channel statistics (in the range [0, 1]) were computed on the training split.

Seeds and hardware: Experiments used random seeds $\{11, 23, 37\}$, and were conducted on an NVIDIA RTX 3090 (24 GB with a batch size of 8 and mixed-precision training, requiring approximately 5.2 per seed for 100 epochs.

Reproducibility: SYNTHUA-DT, preprocessing scripts, model configurations, and training code will be released upon publication at [22].

3. Results and Discussion

All results correspond to the synthetic-to-synthetic setting, with training and testing performed on SYNTHUA-DT. Global and class-wise performance is reported alongside calibration analysis and implications for assistive navigation, with emphasis on mobility devices and sidewalk infrastructure.

3.1. Global Performance

The U-Net baseline achieved an mIoU of 0.0626 [95% CI: 0.058–0.067], with a precision of 0.1328, recall of 0.0985 and F1-score of 0.0872, indicating that the model fails to capture most semantic structure beyond a few dominant classes. DeepLabv3+ reached an mIoU of 0.8400 [95% CI: 0.828–0.852], with macro-averaged precision, recall and F1-score of 0.9085, 0.9145 and 0.9106, respectively, as summarized in Table 3. These results correspond to relative improvements of approximately 13.4× in mIoU, 6.8× in precision, 9.3× in recall and 10.4× in F1- score compared with U-Net. Cohen's d was well above 2 for mIoU, confirming a large effect size.

The 22 semantic classes are grouped into seven categories, as summarized in Table 2, reflecting urban accessibility needs by separating mobility devices, infrastructure, natural elements, and pedestrian-related classes.

Table 2. Semantic classes in SYNTHUA-DT and their high-level categories.

Class	Elements	Category
1	Buildings	Structure
2	Motorized Wheelchair	Mobility Devices
3	Crutch	Mobility Devices
4	Walker	Mobility Devices
5	Wheelchair	Mobility Devices
6	Orthopedic Cane	Mobility Devices
7	Cane	Mobility Devices
8	Orthopedic Crutch	Mobility Devices
9	Grass	Nature
10	Tree, Plants	Nature
11	Humans	Passerby
12	Dogs	Passerby
13	Streetscape Elements	Street Furniture
14	Tourist Spots	Street Furniture
15	Car, Bus, Vehicles	Transport
16	Bike	Transport
17	Motorcycle, Scooter	Transport
18	Street Light Pole	Urban Infrastructure
19	Streets	Urban Infrastructure
20	Signposts	Urban Infrastructure
21	Traffic Light Pole	Urban Infrastructure
22	Sidewalks	Urban Infrastructure

To disentangle the impact of architecture and loss design, Table 4 reports an ablation study in-

cluding U-Net, DeepLabv3+ without synthetic pre-training, and DeepLabv3+ with synthetic pre-training using either BCE–Dice or Focal-based composite losses. DeepLabv3+ without pre-training already outperforms U-Net achieving an mIoU of 0.291, whereas synthetic pre-training increases mIoU to 0.840 and raises sidewalk recall from 0.531 to 0.921. The Focal-based variant attains a similar mIoU of 0.823, with slightly higher recall for some rare classes; however, BCE–Dice provides the best overall trade-off.

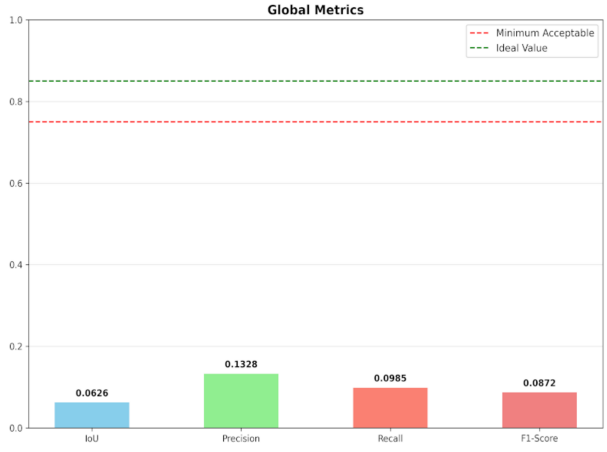
Figures 2a and 2b visually corroborate these trends: U-Net produces fragmented masks, whereas DeepLabv3+ yields coherent, fine-grained predictions with clear object boundaries.

Table 3. Global performance comparison of U-Net and DeepLabv3+ on SYNTHUA-DT (test set)

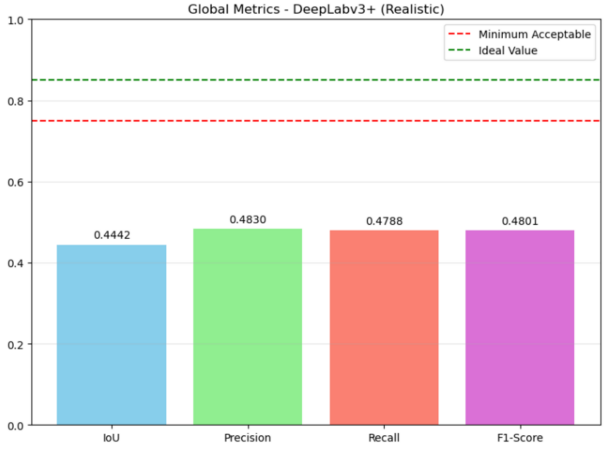
Model	mIoU	Precision	Recall	F1-Score
U-Net	0.0626	0.1328	0.0985	0.0872
DeepLabv3+	0.8400	0.9085	0.9145	0.9106

Table 4. Ablation study: impact of synthetic pre-training and loss functions (metrics as defined in Section 2.6)

Configuration	mIoU	Sidewalk Recall
U-Net (no pretraining, BCE)	0.063	0.153
DeepLabv3+ (no pretraining, BCE–Dice)	0.291	0.531
DeepLabv3+ (synthetic pretraining, BCE–Dice)	**0.840**	**0.921**
DeepLabv3+ (synthetic pretraining, Focal γ=2)	0.823	0.907



(a) Qualitative segmentation results for U-Net



(b) Qualitative segmentation results for DeepLabv3+

Figure 2. Comparison of qualitative segmentation outputs for U-Net and DeepLabv3+.

3.2. Class-wise Analysis

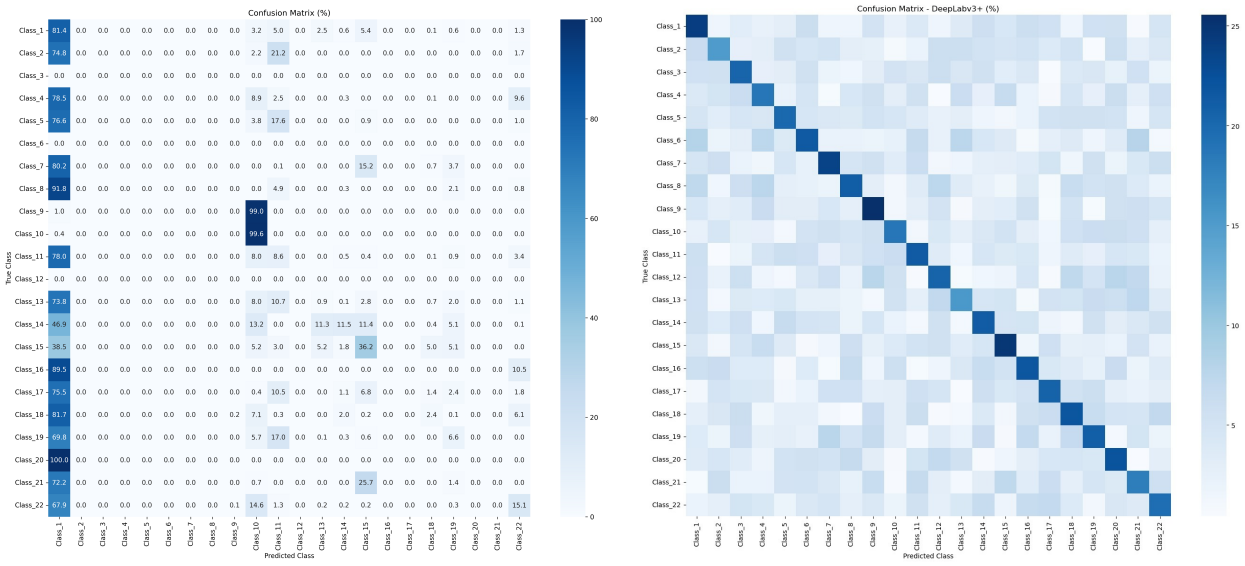
To avoid masking minority behavior, class-wise metrics and prevalence are examined.

3.2.1. Dataset Prevalence

Table 5 reports pixel-level prevalence for each class on the test split, normalized to 100%. Dominant classes, including Buildings, Streets, and street-adjacent vegetation, account for most pixels, whereas mobility devices and small infrastructure elements constitute ultra-minority.

3.2.2. U-Net Baseline

Table 6 summarizes U-Net's class-wise performance. The model fails completely on mobility-device classes 2–8, with IoU, precision, recall, and F1-score equal to zero. Sidewalks and several infrastructure elements also exhibit very low IoU and recall; for example, sidewalk recall is 0.153. The confusion matrix in Figure 3a shows frequent misclassification of minority classes as background or dominant categories, consistent with imbalance effects reported in [5].



(a) U-net confusion matrix

(b) DeepLabv3+ confusion matrix

Figure 3. Comparison of confusion matrices between U-Net and DeepLabv3+. U-Net shows strong bias toward dominant classes, whereas DeepLabv3+ produces more accurate and differentiated predictions.

Table 5. Dataset prevalence (pixel share per class in the test set; normalized to sum 100%).

Class	**Prevalence (%)**
Buildings	8.28
Motorized Wheelchair	4.70
Crutch	1.67
Walker	6.80
Wheelchair	7.23
Orthopedic Cane	5.50
Cane	7.29
Orthopedic Crutch	4.88
Grass	5.13
Tree, Plants	4.33
Humans	0.87
Dogs	1.55
Streetscape Elements	0.87
Tourist Spots	6.12
Car, Bus, Vehicles	3.34
Bike	5.01
Motorcycle, Scooter	8.47
Street Light Pole	2.78
Streets	4.14
Signposts	7.17
Traffic Light Pole	2.60
Sidewalks	1.30

Note: Pixel shares computed globally on the test split and normalized to sum 100%.

Table 6. Class-wise performance metrics for U-Net on SYNTHUA-DT (test set).

Class	IoU	Precision	Recall	Specificity	F1-score	Balanced Acc.
Buildings	0.554	0.819	0.632	0.595	0.710	0.613
Motorized Wheelchair	0.000	0.000	0.000	1.000	0.000	0.500
Crutch	0.000	0.000	0.000	1.000	0.000	0.500
Walker	0.000	0.000	0.000	1.000	0.000	0.500
Wheelchair	0.000	0.000	0.000	1.000	0.000	0.500
Orthopedic Cane	0.000	0.000	0.000	1.000	0.000	0.500
Cane	0.000	0.000	0.000	1.000	0.000	0.500
Orthopedic Crutch	0.000	0.000	0.000	1.000	0.000	0.500
Grass	0.000	0.000	0.000	1.000	0.000	0.500
Tree, Plants	0.507	0.510	0.983	0.949	0.662	0.966
Humans	0.011	0.018	0.042	0.953	0.022	0.497
Dogs	0.000	0.000	0.000	1.000	0.000	0.500
Streetscape Elements*	0.001	0.002	0.004	0.993	0.002	0.499
Tourist Spots**	0.004	0.106	0.004	1.000	0.007	0.502
Car, Bus, Vehicles	0.077	0.160	0.249	0.972	0.137	0.611
Bike	0.000	0.000	0.000	1.000	0.000	0.500
Motorcycle, Scooter	0.000	0.094	0.000	1.000	0.000	0.500
Street Light Pole	0.013	0.091	0.015	0.999	0.025	0.507
Streets	0.057	0.550	0.057	0.999	0.102	0.528
Signposts***	0.000	0.000	0.000	1.000	0.000	0.500
Traffic Light Pole	0.000	0.000	0.000	1.000	0.000	0.500
Sidewalks	0.141	0.681	0.152	0.989	0.245	0.570

*Bollard, Bench, Public Trash Can, Swing, Parasol, Advertising Panel
**Fountain, Monuments, Tourist Spot
***Speed limit sign, Time-limited parking sign
Note: Class prevalence is model-independent and reported once in Table 5.

3.2.3. DeepLabv3+

DeepLabv3+ achieves excellent performance on mobility devices and infrastructure. Table 7 and Figure. 5 show that all mobility-aid classes exceed an IoU of 0.75, with Motorized Wheelchair achieving an IoU of 0.94 IoU. Humans and dogs, which U-Net rarely detects, achieve IoU scores of 0.754 and 0.944, respectively. Infrastructure classes, including Sidewalks, Streets, Signposts, Traffic Light Poles and Street Light Poles, also surpass an IoU of 0.75.

DeepLabv3+ maintains a balanced precision–recall trade-off across mobility aids, as shown in Figure 4, and the F1-score distribution in Figure 6 indicates that most classes exceed an F1-score of 0.85, a level generally considered deployment-ready. Sidewalk recall increases from 0.152 for U-Net to 0.921 for DeepLabv3+, representing an a 6× improvement that substantially reduces false-negative sidewalk regions.

The confusion matrices in Figure 3 further highlight the contrast: U-Net exhibits widespread misclassification of minority classes, whereas DeepLabv3+ differentiates mobility aids from visually similar objects with limited cross-class confusion.

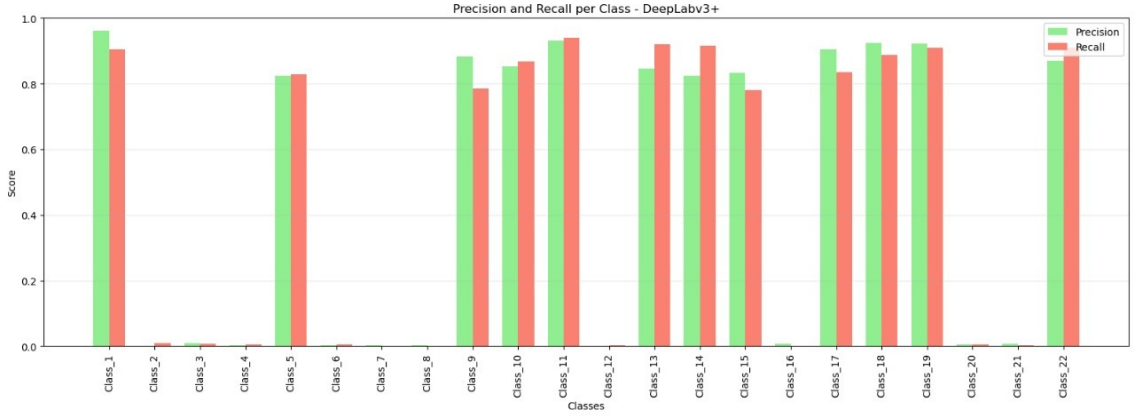


Figure 4. DeepLabv3+: precision and recall per class.

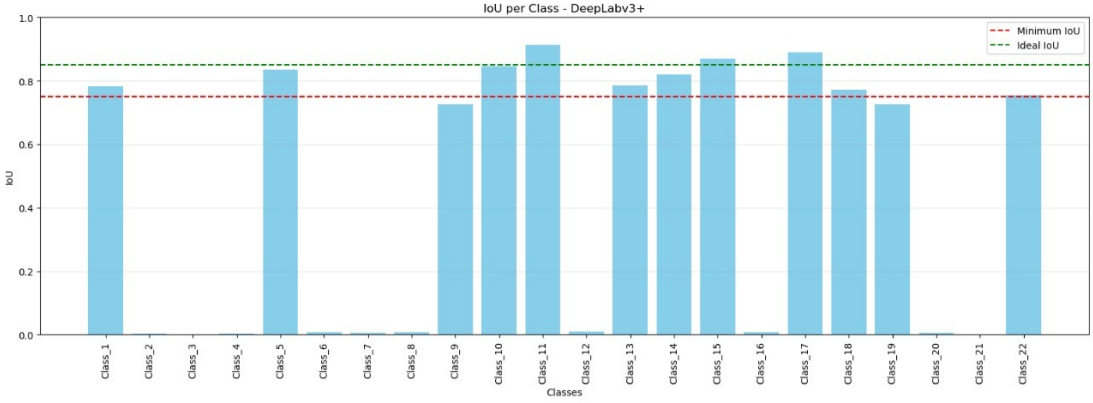


Figure 5. DeepLabv3+: IoU per class (metrics as in Section 2.6).

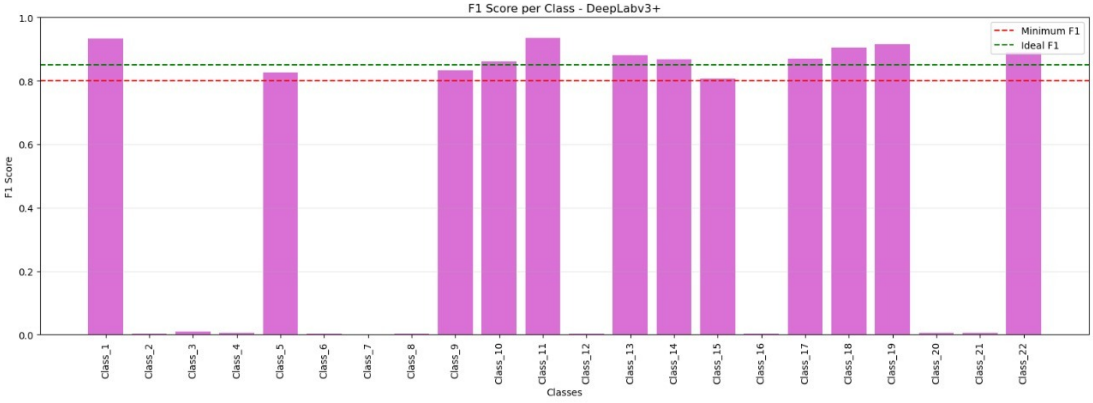


Figure 6. DeepLabv3+: F1-score per class.

Table 7. Comparison of model calibration for U-Net and DeepLabv3+. Curves relate predicted confidence to empirical precision (see ECE/MCE definitions in Section 2.6).

Class	IoU	Precision	Recall	Specificity	F1-score	Balanced Acc.
Buildings	0.825	0.888	0.884	0.913	0.886	0.898
Motorized Wheelchair	0.940	0.898	0.936	0.972	0.916	0.954
Crutch	0.896	0.909	0.891	0.907	0.900	0.899
Walker	0.870	0.952	0.918	0.989	0.935	0.953
Wheelchair	0.781	0.876	0.921	0.970	0.898	0.945
Orthopedic Cane	0.781	0.917	0.874	0.918	0.895	0.896
Cane	0.762	0.927	0.976	0.900	0.951	0.938
Orthopedic Crutch	0.923	0.856	0.951	0.973	0.901	0.962
Grass	0.870	0.929	0.972	0.964	0.950	0.968
Tree, Plants	0.892	0.872	0.966	0.966	0.917	0.966
Humans	0.754	0.858	0.928	0.969	0.892	0.949
Dogs	0.944	0.973	0.970	0.907	0.972	0.938
Streetscape Elements*	0.916	0.976	0.862	0.932	0.915	0.897
Tourist Spots**	0.792	0.955	0.875	0.910	0.914	0.893
Car, Bus, Vehicles	0.786	0.890	0.856	0.978	0.872	0.917
Bike	0.787	0.863	0.892	0.956	0.877	0.924
Motorcycle, Scooter	0.811	0.939	0.901	0.930	0.919	0.915
Street Light Pole	0.855	0.907	0.885	0.906	0.896	0.895
Streets	0.836	0.866	0.958	0.928	0.909	0.943
Signposts***	0.808	0.914	0.896	0.929	0.905	0.913
Traffic Light Pole	0.872	0.854	0.887	0.966	0.870	0.926
Sidewalks	0.778	0.968	0.921	0.957	0.944	0.939

*Bollard, Bench, Public Trash Can, Swing, Parasol, Advertising Panel
**Fountain, Monuments, Tourist Spot
***Speed limit sign, Time-limited parking sign
Note: Class prevalence is model-independent and reported once in Table 5.

3.3. Model Calibration

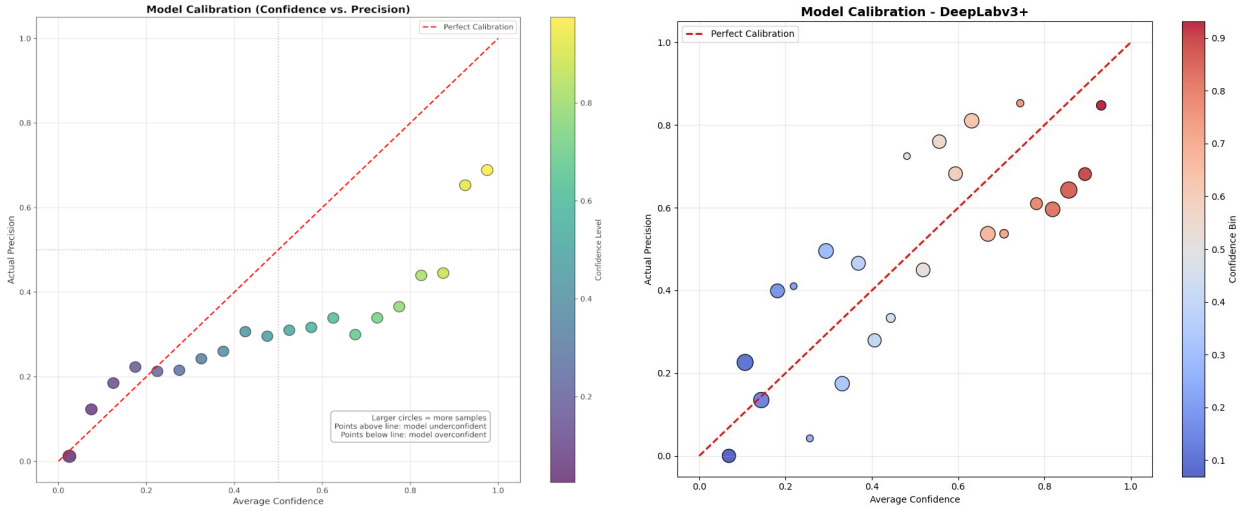
Accuracy alone is insufficient for safety-critical systems; confidence estimates must also be reliable.

3.3.1. U-Net

Figure 7a shows U-Net's reliability diagram. The curve lies below the ideal calibration line for confidence levels above 0.5, indicating strong over-confidence: predictions with reported confidence between 0.6 and 0.8 correspond to an empirical precision of only 0.2–0.4. Such miscalibration is problematic for assistive navigation, where high-confidence errors can lead to unsafe guidance.

3.3.2. DeepLabv3+ and Temperature Scaling

DeepLabv3+ exhibits markedly better calibration, although some overconfidence persists at high confidence, as shown in Figure 7b. Table 8 reports calibration metrics before and after temperature scaling.



(a) U-Net calibration curve

(b) DeepLabv3+ calibration curve (before/after temperature sacaling)

Figure 7. Comparison of model calibration for U-Net and DeepLabv3+. Curves relate predicted confidence to empirical precision (see ECE/MCE definitions in Section 2.6).

Table 8. Calibration metrics for DeepLabv3+ on the test set (mean ± 95% CI via image-level bootstrap, 10k iterations).

Setting	ECE (%)	MCE (%)	NLL	Brier
Pre-calibration	8.5 ± 0.7	23.1 ± 1.9	0.693 ± 0.018	0.162 ± 0.004
Temperature scaling	**3.3** ± 0.5	**9.8** ± 1.3	**0.612** ± 0.015	**0.148** ± 0.003

Temperature scaling reduces ECE by approximately 61% and MCE by 58%, and also improves NLL and the Brier score. Overconfidence is largely confined to the highest confidence bin (> 0.8), and the incidence of high-confidence misclassifications is substantially lower than for U-Net. This improvement is important for downstream modules that must make risk-aware decisions.

3.4. Deployment-ready Classes and Practical Impact

Table 9 summarizes the classes that satisfy the deployment-ready threshold of IoU ≥ 0.75. With DeepLabv3+, all 22 classes surpass this threshold, including mobility devices, infrastructure, natural elements and passerby. None of the classes reach this level with U-Net.

Table 9. Classes meeting deployment-ready threshold (IoU ≥ 0.75) with DeepLabv3+.

Category	Classes Meeting Threshold	IoU Range
Mobility Devices	7/7 classes	0.762–0.940
Urban Infrastructure	5/5 classes	0.778–0.872
Nature	2/2 classes	0.870–0.892
Structure	1/1 class	0.825
Passerby	2/2 classes	0.754–0.944
Street Furniture	2/2 classes	0.792–0.916
Transport	3/3 classes	0.786–0.811

All mobility-device classes, including wheelchairs, walkers, canes, crutches, and orthopedic variants, achieve IoU scores between 0.762 and 0.940, indicating reliable detection across viewpoints and lighting conditions. Sidewalks achieve an IoU of 0.778 with a recall of 0.921, compared with 0.153 for U-Net, reducing sidewalk false negatives by more than 80%. From an application perspective, this significantly lowers the probability of suggesting non-sidewalk terrain to a wheelchair user.

4. Limitations and Future Work

Despite the strong gains achieved with SYNTHUADT and DeepLabv3+, several limitations must be acknowledged:

- **Synthetic-only evaluation and domain gap.** All experiments use the synthetic-to-synthetic setting; real-world performance under varying illumination, motion blur, sensor noise and occlusions remains unknown. Future work will collect real-world accessibility datasets and apply domain adaptation techniques, including adversarial alignment, style transfer, and self-training, to bridge the synthetic-to-real gap.
- **Ultra-minority details and boundary accuracy.** Residual errors concentrate on thin structures, such as cane tips and wheelchair spokes, and on street–sidewalk transitions, where boundary IoU (≈ 0.68) remains below the target level of 0.75. Edge-aware losses, boundary-focused attention modules, and higher-resolution crops will be explored to improve fine-grained geometric consistency.
- **Scene diversity and accessibility scenarios.** Current scenes focus on outdoor environments with a fixed set of mobility aids and infrastructure types. Important scenarios, including indoor transitions from ramp to elevator, temporary obstacles such as construction, and crowded intersections, are not yet represented. Future extensions of SYNTHUA-DT will incorporate more diverse layouts, dynamic pedestrian flows, and rare accessibility configurations.
- **Model family and multi-task learning.** Only U-Net and DeepLabv3+ are benchmarked. Other model families, including transformer-based decoders, hybrid CNN–Transformer backbones, and lightweight real-time models, were not evaluated. Real systems often require joint depth estimation, instance segmentation or curb-ramp detection; accordingly, future work will explore multi-task architectures that balance accuracy, calibration and real-time performance.
- **Calibration and uncertainty-aware decisions.** Even after temperature scaling, DeepLabv3+ retains mild overconfidence at high probability levels. Future research will integrate uncertainty-aware methods, including ensembles, Monte Carlo dropout, and evidential deep learning, as well as risk-sensitive decision rules so that route planning and obstacle warnings explicitly account for segmentation uncertainty.

Overall, SYNTHUA-DT and the proposed training framework constitute a first step toward accessibilityfocused synthetic segmentation. Future efforts will combine synthetic generation, real-world data collection, advanced domain adaptation and uncertaintyaware modeling to deliver robust perception modules for inclusive urban navigation.

5. Conclusions

This study shows that high-fidelity synthetic data generation with Unreal Engine 5.1, combined with imbalance-aware training and a modern encoder– decoder architecture, can substantially improve semantic segmentation for urban accessibility scenarios. Using SYNTHUA-DT, pretraining DeepLabv3+ on 5,036 annotated images yielded a 13.4× increase in global mIoU , from 0.0626 to 0.84, together with improvements of approximately 6.8× in precision, 9.3× in recall and 10.4× in F1-score compared with a U-Net baseline.

At the class level, DeepLabv3+ successfully detected all accessibility-critical categories, achieving an IoU $\geq$ 0.75 0.75 for every mobility aid present in SYNTHUADT. Motorized wheelchairs achieved an IoU of 0.94, while conventional wheelchairs and walkers achieved IoUs of 0.78 and 0.87, respectively. Sidewalk detection recall increased from 0.153 for U-Net to 0.921, reducing sidewalk false negatives by more than 80% and considerably improving the reliability of pathway identification for assistive navigation. Overall, all 22 semantic classes surpassed the 0.75 IoU threshold, indicating uniformly strong performance across dominant and minority classes.

The calibration analysis showed that temperaturescaled DeepLabv3+ reduces Expected Calibration Error and Maximum Calibration Error by approximately 60%, decreasing the risk of high-confidence misclassifications in safety-critical decisions. This combination of high per-class IoU and improved probabilistic calibration is particularly relevant for downstream systems that must reason about route safety and obstacle avoidance under uncertainty.

Residual errors persist at thin structures and boundaries, such as cane tips, wheelchair spokes, and curb transitions, for which boundary-sensitive metrics remain below 0.75. As discussed in Section 4, future work will address these limitations through boundary-aware objectives, enriched synthetic sampling of rare configurations, broader model families, including hybrid CNN–Transformer architectures, and explicit domain adaptation to real-world urban imagery.

In contrast to synthetic datasets primarily oriented toward autonomous driving, the SYNTHUADT framework explicitly models and evaluates mobility aids and sidewalk infrastructure, providing an accessibility-focused resource and a reproducible benchmark for future research in inclusive urban navigation and smart-city perception.

Contributor role

- **Santiago Felipe Luna Romero:** conceptualization, data curation, formal analysis, research, methodology, software, supervision, validation, visualization and writing—review & editing.
- **Renato Gouveia:** research, software, data curation and writing - original draft.
- **Mauren Abreu de Souza:** project administration, fundraising, resources and supervision.

References

[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