

Open-Vocabulary Annotation for Pedestrian Scene Understanding under Domain Shift

Justine Buss¹, Michael Kreutzer¹, and Seyed Eghbal Ghobadi¹

Technische Hochschule Mittelhessen – University of Applied Sciences
Gießen, Germany

{justine.buss,michael.kreutzer,seyed.eghbal.ghobadi}@mni.thm.de

Abstract. The ability to safely navigate complex urban environments remains a major challenge for people with visual impairments in everyday pedestrian settings. While AI-based assistive systems offer promising support, their development is often constrained by the lack of domain-specific annotated datasets. Existing urban benchmarks rarely capture the pedestrian-centered scenes required for accessibility applications.

This paper investigates whether open-vocabulary grounding and segmentation models can be used to generate pixel-level annotations under strong domain shift in pedestrian scenes. We combine language-guided object localization with prompt-based segmentation and introduce an image-level filtering strategy to improve annotation consistency while reducing manual effort.

The resulting pseudo-labeled subset is used to train and evaluate multiple semantic segmentation architectures for binary sidewalk segmentation and road-sidewalk segmentation, two perception tasks relevant to assistive pedestrian navigation. High-capacity models exceed 80 % mIoU for binary sidewalk segmentation, while lightweight models remain competitive, suggesting that the generated labels preserve large-scale scene structure for low-resource assistive vision.

Overall, these findings suggest that foundation-model-based annotation pipelines offer a practical approach to dataset initialization in assistive vision domains, while semantic grounding under domain shift remains a key limitation.

Keywords: semantic segmentation · open-vocabulary annotation · foundation models · assistive vision · pedestrian scene domain shift

1 Introduction

Visually impaired individuals face significant challenges when navigating urban environments due to dense infrastructure, dynamic traffic conditions, and small or partially occluded obstacles [1, 11]. While assistive tools such as white canes or guide dogs provide reliable support, they are limited in sensing range and lack holistic scene understanding [10, 11].

Camera-based assistive vision systems aim to overcome these limitations by enabling contextual scene understanding through computer vision models [1,

21]. A key requirement for such systems is dense semantic scene understanding, commonly formulated as pixel-wise semantic segmentation [6].

However, progress in this direction is often constrained not by model capacity, but by the availability of suitable training data. Existing urban scene datasets are predominantly vehicle-centric and fail to adequately represent pedestrian-centered, near-field environments, including sidewalks, curb boundaries, and small obstacles [3, 7, 13]. As a result, researchers developing accessibility-oriented systems frequently face the need to construct task-specific datasets, making annotation a central bottleneck.

Recent advances in open-vocabulary vision models have demonstrated that pretrained foundation models can significantly reduce annotation effort. In particular, promptable segmentation models such as Segment Anything (SAM) [12] and open-vocabulary grounding models such as Grounding DINO [14] enable scalable weak supervision for dense annotation. Their combination, as explored in Grounded-SAM [17], provides a general paradigm for automatic dataset construction.

In this work, we investigate whether this paradigm transfers to pedestrian-centered assistive vision applications. We propose an automated annotation pipeline that combines open-vocabulary grounding with prompt-based segmentation to generate dense labels for urban pedestrian scenes. To improve robustness under domain shift, we further introduce a conservative postprocessing strategy that filters inconsistent predictions while maintaining low manual overhead.

We construct a domain-specific dataset for pedestrian scene understanding in assistive vision and evaluate its effectiveness by training multiple semantic segmentation architectures. In addition, we analyze how domain shift in pedestrian scenes affects grounding quality in open-vocabulary pipelines and how such errors propagate to downstream segmentation performance. Our contributions are summarized as follows:

- We investigate foundation-model-based annotation as a low-effort dataset creation strategy for pedestrian-centered assistive vision under strong near-field domain shift.
- We construct a domain-specific pedestrian-perspective dataset (BGH) and an automatically annotated and filtered variant (BGH-A).
- We introduce a conservative postprocessing strategy that improves annotation consistency with minimal manual effort.
- We analyze annotation fidelity and downstream segmentation performance across binary sidewalk segmentation and road-sidewalk segmentation settings, identifying key challenges arising from domain shift in pedestrian scenes in assistive vision scenarios.

2 Related Work

2.1 Assistive Systems for Visually Impaired Navigation

Assistive technologies for visually impaired pedestrians have evolved from sensor-based devices toward vision-driven approaches. Early systems, such as enhanced

white canes with ultrasonic or infrared sensors, extend local obstacle perception but remain limited to short-range detection and lack global scene understanding [11].

More recent approaches use camera-based perception systems combined with deep learning to identify obstacles, pedestrians, or traffic elements and provide auditory or haptic feedback [1, 21]. While effective for object-level recognition, these methods remain fundamentally local and do not explicitly model traversable space or holistic scene structure required for reliable navigation.

This limitation has motivated a shift toward structured scene understanding in pedestrian environments. However, progress in this direction is tightly coupled to the availability of annotated data in relevant viewpoints.

2.2 Semantic Segmentation for Urban Scene Understanding

Semantic segmentation provides a standard formulation for dense scene understanding by assigning per-pixel labels. It supports tasks such as free-space estimation, scene layout reasoning, and navigation-relevant perception [6].

Large-scale datasets such as Cityscapes [3], KITTI [13], and A2D2 [7] have driven progress in urban perception. However, these datasets are primarily vehicle-centric and exhibit strong viewpoint bias, limiting their applicability to near-field pedestrian scenarios. Models trained on them often suffer from domain shift in scale, perspective, and ground-plane structure [20].

While pedestrian-centered datasets such as Urban Footpath Image Dataset [5] and Sidewalk Semantic [4] have become available, they remain limited in scale, semantic coverage, or annotation consistency, leaving a persistent gap between available data and the requirements of assistive vision systems.

2.3 Annotation Cost and Dataset Limitations

Pixel-level annotation in pedestrian environments is costly due to fine-grained boundaries, ambiguous class definitions, and cluttered scenes [3, 13]. This makes dataset creation a major bottleneck for assistive vision applications requiring dense supervision.

Semi-automatic annotation tools and synthetic data generation can reduce effort but typically require manual refinement or fail to generalize across diverse real-world conditions [15]. This limits their effectiveness for scalable dataset creation in domain-specific settings.

2.4 Foundation Models for Open-Vocabulary Annotation

Recent vision foundation models provide a promising direction for reducing annotation cost in low-resource domains. SAM [12] enables prompt-based mask generation, while Grounding DINO [14] performs open-vocabulary object localization using natural language queries.

Their combination in pipelines such as Grounded-SAM [17] enables automatic image annotation without task-specific training.

However, most existing work evaluates such pipelines on generic object-centric benchmarks and does not systematically analyze their behavior under strong pedestrian near-field domain shift. In these settings, small object scale, ambiguous ground-plane structure, and viewpoint bias may significantly affect grounding reliability and segmentation consistency [9].

In this work, we study the applicability of foundation-model-based annotation pipelines for assistive vision dataset creation and analyze their failure modes under domain shift in pedestrian scenes.

3 Methods

3.1 Problem Setting

Pedestrian-centered urban environments differ substantially from standard urban perception benchmarks. Compared to vehicle-centered datasets, they exhibit stronger near-field effects, smaller object instances, frequent occlusions, and less well-defined boundaries between navigable and non-navigable regions [3, 7, 13]. These characteristics introduce a strong domain shift for open-vocabulary grounding models, particularly affecting semantic consistency across visually similar ground-plane classes [9, 20].

The input consists of unlabeled pedestrian-perspective images captured in urban and semi-urban environments. The scene domain focuses on near-field urban perception relevant to pedestrian navigation.

Table 1: Semantic classes used for annotation and evaluation.

Idx	Class	Idx	Class
0	road	5	vegetation
1	sidewalk	6	person
2	building	7	car
3	traffic light	8	bicycle
4	traffic sign	9	trash can

A compact semantic label space of ten classes is defined in Table 1. Each class is associated with fixed natural-language prompts used for open-vocabulary localization and evaluation.

3.2 Foundation-Model Annotation Pipeline

The annotation pipeline, illustrated in Fig. 1, consists of three stages: open-vocabulary localization, prompt-based segmentation, and mask aggregation.

Grounding DINO [14] serves as the semantic localization component of the pipeline. Conditioned on predefined textual prompts derived from the class

schema in Tab. 1, it identifies and localizes semantically relevant regions in the input image. This yields a set of candidate detections with associated class labels and confidence scores. Low-confidence predictions are filtered, and non-maximum suppression is applied to remove redundant detections. The resulting bounding boxes define region prompts for the segmentation stage.

SAM [12] generates pixel-accurate masks conditioned on these bounding boxes. Each mask inherits the semantic label from the corresponding detection. Instance-level masks are merged into a single semantic label map using a confidence-based selection strategy in overlapping regions.

The output is a dense semantic segmentation map in which each pixel is assigned a class index from the predefined label set defined in Tab. 1.

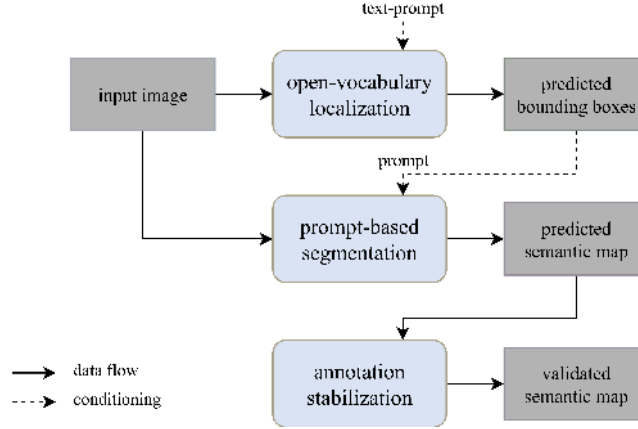


Fig. 1: Overview of the foundation-model-based annotation pipeline for generating semantic segmentation maps from *open-vocabulary localization* and *prompt-based segmentation*.

3.3 Annotation Stabilization

The generated annotations may contain errors originating from either localization or segmentation. To improve reliability, a lightweight post-hoc stabilization step is applied.

Each generated annotation map is inspected at image level and assigned a binary accept–reject decision. The decision is based on visible semantic consistency, with annotations rejected if major objects are missing, incorrectly labeled, or if large regions of the scene are not plausibly explained by the predicted classes. Accepted samples are retained without modification, while rejected samples are removed from the dataset. This procedure provides conservative quality control by filtering samples affected by strong grounding or segmentation failures under domain shift.

4 Experiments

4.1 Datasets

The experimental evaluation is based on the BGH dataset (Berlin–Gießen–Holzheim), which contains 293 pedestrian-perspective images captured during multiple recording sessions in diverse urban and semi-urban environments across Germany. These images serve as input to the annotation pipeline defined in Sec. 3.2, which produces pixel-level semantic labels.

After automated annotation and image-level filtering, 215 images remain for downstream experiments. This subset, referred to as BGH-A, represents the final curated dataset used for all learning-based experiments.

To evaluate the effect of additional supervision and increased data diversity, BGH-A is combined with the publicly available Sidewalk Semantic dataset (1,000 annotated images) [4]. Prior to merging, the original 35 semantic classes of Sidewalk Semantic are remapped to the proposed 10-class taxonomy, while source classes not represented in the target taxonomy are assigned to the ignore label. The resulting combined dataset is referred to as BBGH. While BBGH increases dataset scale and scene variability, it remains strictly separated from the annotation pipeline and is used only for downstream training and evaluation. Training and test splits are created after merging to maintain a consistent evaluation protocol across datasets.

A fixed subset consisting of the first 50 images of the BGH dataset was manually annotated using Labelbox following the same labeling protocol as the automatically generated annotations. These manual annotations serve exclusively as ground truth for evaluating annotation quality. All downstream training and downstream evaluation are conducted using the automatically generated annotations of the corresponding dataset splits.

4.2 Annotation Pipeline Setup

The annotation pipeline is configured using Grounding DINO (SwinT_OGC) for localization and SAM (ViT-H) for mask generation, both employed in pretrained form without task-specific fine-tuning.

The semantic class names defined in Tab. 1 are used directly as textual prompts for Grounding DINO. Bounding box predictions are generated using a box confidence threshold of 0.35 and a text confidence threshold of 0.25 before being forwarded to the segmentation stage. These thresholds are kept fixed throughout all experiments to ensure a consistent evaluation protocol and to isolate the effects of domain shift. Threshold optimization through ablation was not considered in this work.

For each detected bounding box, SAM generates multiple candidate masks, from which the highest-confidence prediction is retained. Generated masks inherit the semantic labels of their corresponding detections and are combined into a single semantic segmentation map using confidence-based overlap resolution.

All configuration parameters, including prompts, thresholds, and label mappings, remain unchanged across experiments to ensure reproducibility. The complete annotation process is executed offline with GPU acceleration during inference.

4.3 Evaluation Setup

Evaluation is structured along three complementary aspects designed to quantify the impact of domain shift in pedestrian scenes across (i) dataset construction stability, (ii) semantic grounding accuracy, and (iii) downstream segmentation performance.

The first aspect focuses on the dataset resulting from the annotation pipeline. It analyzes how postprocessing and manual filtering affect dataset composition, including acceptance rates, per-image processing time, and changes in class-wise and pixel-wise distributions. Additionally, variations in mask quality and mask density under different model configurations (e.g., SAM variants) are used to assess the sensitivity of the annotation process.

The second evaluation stage targets annotation quality to quantify the impact of domain shift on semantic grounding accuracy independent of downstream training. It is assessed through spatial agreement between automatically generated labels and manually annotated ground truth on a subset of 50 images. This comparison is operationalized via Intersection-over-Union (IoU), reported as mean IoU (mIoU).

Finally, downstream utility evaluates whether annotations generated under domain shift are sufficiently reliable to support supervised learning of semantic segmentation models. Performance is measured using the same IoU-based metrics on held-out test splits, ensuring consistency across evaluation stages. For binary sidewalk segmentation, IoU is computed specifically for the sidewalk class and averaged over all test images. Results are additionally compared between BGH-A and BBGH. This comparison allows assessment of whether increased dataset scale can compensate for domain-shift-induced annotation noise.

4.4 Downstream Training Setup

To assess the usefulness of the generated annotations as training supervision, multiple semantic segmentation architectures are evaluated under a unified training protocol.

The selected models cover a range of computational complexity and represent commonly used architectures for dense prediction tasks. Four representative models are considered: U-Net [18], Fast-SCNN [16], DeepLabV3+ [2], and BiSeNetV2 [22]. U-Net and DeepLabV3+ employ ImageNet-pretrained encoders (ResNet-18 [8] and MobileNetV2 [19], respectively), while Fast-SCNN and BiSeNetV2 are implemented following their original architectures. All models are adapted to the shared semantic label space.

All models are trained for 20 epochs using the Adam optimizer with a learning rate of 1×10^{-3} and a batch size of 4. Input images are resized to 256×256 .

pixels and normalized prior to training. Training follows standard segmentation objectives, using binary cross-entropy (BCE) for binary sidewalk segmentation and categorical cross-entropy (CE) for road-sidewalk segmentation. In the binary setting, the models predict sidewalk versus all remaining pixels. In the road-sidewalk setting, only road and sidewalk are treated as valid classes, while all remaining semantic classes are assigned to the ignore label during training. Unless stated otherwise, all training settings are kept fixed across architectures and datasets to ensure comparability.

Each dataset is split at scene level into 85 % training and 15 % test images. Care is taken to ensure that test images originate from recording segments without spatial overlap with the training set, preventing evaluation on previously observed scenes.

To evaluate robustness, experiments are conducted both with and without offline data augmentation. The augmentation pipeline consists of geometric transformations, photometric changes, and additive noise applied during offline dataset generation.

The downstream experiments evaluate the utility and consistency of the generated supervision signal for training semantic segmentation models. While this provides insight into representation learning under domain shift, it does not assess full deployment scenarios such as real-world assistive navigation or user-centered evaluation, which remain beyond the scope of this work.

5 Results

5.1 Dataset Construction Analysis

Annotation Outcomes and Postprocessing The full BGH dataset comprising 293 pedestrian-perspective images is processed end-to-end by the open-vocabulary annotation pipeline. Every image passes through semantic grounding and mask generation without manual pixel-level intervention. A subsequent lightweight filtering stage retains only samples that satisfy basic semantic coherence and spatial plausibility, effectively acting as a conservative quality gate rather than a corrective step.

After filtering, 215 images remain, corresponding to an acceptance rate of 73 %. This rate indicates that a majority of scenes can be reliably handled under the current pipeline formulation, despite the strong domain shift in near-field pedestrian environments. Representative examples of accepted annotations are shown in Fig. 2a. At the same time, the non-trivial rejection fraction highlights systematic failure cases rather than random noise.

Rejected samples, as illustrated in Fig. 2b, are not uniformly distributed but concentrate in visually ambiguous street scenes. In particular, confusion arises in boundary regions between road and sidewalk, as well as in occluded or small-scale object configurations. This pattern suggests that the dominant limitation is not the absence of objects, but semantic ambiguity in structurally adjacent ground-plane categories.

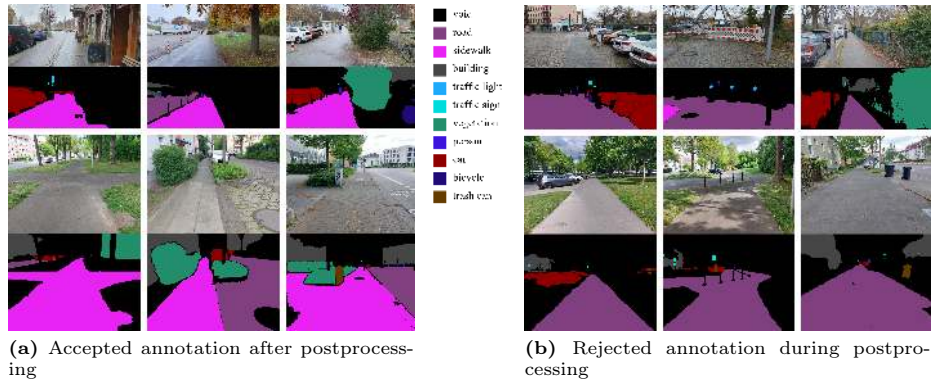


Fig. 2: Comparison of accepted and rejected annotations in the BGH dataset after postprocessing.

Efficiency and Human Effort Beyond annotation quality, the pipeline significantly alters the cost structure of dataset creation. After a one-time initialization overhead (3 min 41 s), full processing requires approximately 3.4–4.7 seconds per image, depending on scene complexity and the number of detected instances. The entire dataset is processed within roughly 18–21 minutes in an offline GPU setting.

Human involvement is reduced to binary acceptance decisions at image level. With an average inspection time of approximately 5 seconds per image, total manual effort remains around 25 minutes. Importantly, this effort is independent of pixel-level complexity and scales only with dataset size.

This stands in strong contrast to manual annotation requirements. A 50-image subset requires 571 minutes of dense labeling, corresponding to 11.42 minutes per image. Extrapolated to the full dataset, this would amount to approximately 55.8 hours. The resulting gap indicates a shift in annotation paradigm: from dense pixel supervision to lightweight validation of model-generated structure.

Dataset Composition The resulting dataset reflects both the underlying scene statistics in Tab. 2 and the inductive biases of the annotation pipeline. Large structural classes such as road, sidewalk, building, and car dominate image-level presence (> 80 % of images), indicating that the pipeline reliably captures primary urban context relevant for pedestrian navigation.

At pixel level, a smaller subset of classes accounts for the majority of annotated area. Road and sidewalk together form approximately one third of all labeled pixels, reinforcing their role as primary structure in near-field scenes. In contrast, fine-grained object categories such as traffic lights, bicycles, and trash cans remain extremely sparse, each contributing less than 0.2 % of pixels.

A notable fraction of pixels (approximately 47 %) is assigned to the void class. Rather than indicating labeling noise alone, this reflects a structural limitation of

open-vocabulary grounding under domain shift: large parts of pedestrian imagery either fall outside the predefined semantic schema or cannot be reliably grounded with sufficient confidence.

Filtering has limited impact on global class statistics (Tab. 2), suggesting that rejection operates at the level of entire scenes rather than redistributing class proportions. The most visible shifts occur for road and sidewalk, which also represent the most ambiguous boundary region in pedestrian-view imagery. This reinforces the observation that the main challenge lies in fine-grained discrimination between semantically adjacent traversable surfaces rather than in coarse scene coverage.

Table 2: Class distribution before and after filtering.

Class	Images		Pixel Share	
	Pre	Post	Pre	Post
road	241	169	15.49	11.83
sidewalk	240	208	17.45	21.67
building	242	178	10.19	9.66
traffic light	122	98	0.04	0.04
traffic sign	185	136	0.17	0.19
vegetation	149	112	6.02	6.34
person	148	105	0.06	0.06
car	265	203	2.98	2.95
bicycle	102	73	0.22	0.19
trash can	43	31	0.08	0.05
void	293	215	47.29	47.01

5.2 Annotation Quality

Agreement with Manual Ground Truth To assess how strongly open-vocabulary grounding transfers into pixel-level supervision under domain shift, automatically generated labels are compared against manually annotated ground truth on a 50-image subset.

The resulting mean IoU of 0.522 indicates a moderate but non-trivial alignment between model-generated annotations and human perception. While this remains substantially below fully supervised segmentation performance, it is notable given the zero-shot nature of the pipeline and the challenging domain shift.

A more detailed breakdown in Tab. 3 reveals a clear dependence on structural scale and visual ambiguity. Large, well-defined classes such as sidewalk (0.792), car (0.703), and building (0.625) are consistently well captured. These categories typically exhibit stable spatial extents and distinct visual cues, which aligns well with both grounding and segmentation priors of foundation models.

In contrast, performance degrades sharply for small or visually ambiguous classes such as traffic light (0.091) and trash can (0.241). These failures are

not isolated but systematically linked to near-field clutter and low-resolution object representations, where both localization and class disambiguation become unstable.

Table 3: Per-class IoU between automatically generated annotations and manual ground-truth annotations on a 50-image subset. Bold indicates the highest IoU, and italic the lowest.

Class	IoU	Class	IoU
road	0.456	vegetation	0.401
sidewalk	0.792	person	0.594
building	0.625	car	0.703
traffic light	<i>0.091</i>	bicycle	0.460
traffic sign	0.437	trash can	0.241

Detection and Error Analysis A closer inspection of error structure reveals that inaccuracies are dominated by semantic misclassification rather than failures of spatial localization. The segmentation masks produced by SAM remain largely spatially consistent, while errors primarily originate from incorrect class assignments during open-vocabulary grounding.

This is reflected in the scale of class confusion, which affects 24.35 % of correctly localized object pixels. This indicates that a dominant failure mode lies in semantic grounding under domain shift, rather than in mask quality or geometric alignment.

Errors are strongly concentrated in semantically adjacent ground-plane categories, especially road and sidewalk. This aligns with observations from both dataset construction and per-class IoU analysis, suggesting a consistent failure mode: visually similar traversable surfaces are frequently interchangeable under open-vocabulary labeling.

When accounting for class importance via weighted IoU, performance increases to 0.679. The weighted IoU is computed as the weighted average of per-class IoUs,

$$wIoU = \frac{\sum_i w_i \cdot IoU_i}{\sum_i w_i},$$

where class weights (w_i) are manually assigned according to their relevance for pedestrian navigation: 1.2 for sidewalks, 1.0 for buildings and cars, 0.5 for medium-relevance classes, and 0.1 for small or rare objects. While road–sidewalk confusion remains the primary safety-relevant failure mode, many remaining errors affect small or rare classes with limited impact on overall scene structure. Thus, weighting classes by functional relevance yields a higher overall score while highlighting the comparatively stable navigation-relevant scene structure.

Spatial Sensitivity Analysis To investigate whether domain shift induces spatial bias in the annotation pipeline, we analyze spatially weighted IoU variants. This serves as a diagnostic tool to detect potential region-dependent failures in grounding or segmentation, e.g., systematic preference for near-field regions or central image areas.

Spatially weighted IoU variants (horizontal, vertical, center-biased, and corridor-focused) remain tightly clustered between 0.525 and 0.527. The absence of significant variation across weighting schemes suggests that performance is not systematically biased toward specific image regions.

This stability implies that the pipeline does not preferentially degrade in near-field versus far-field regions. Instead, errors are primarily semantic in nature and largely independent of spatial position. In other words, domain shift affects what is recognized rather than where it is recognized.

5.3 Downstream Utility

Binary Segmentation The utility of automatically generated annotations is ultimately reflected in their ability to support supervised learning. Binary sidewalk segmentation provides a controlled setting to evaluate whether the generated labels preserve sufficient structural signal for downstream optimization.

Across architectures, training on the BGH-A dataset yields consistently strong performance, as summarized in Tab. 4. High-capacity models such as DeepLabV3+ and U-Net exceed 80 % mIoU, while lightweight models such as Fast-SCNN and BiSeNetV2 remain in the 70 % range. This consistency suggests that despite moderate annotation noise, the dataset preserves robust structural supervision for large-scale scene components.

A comparison with the larger BBGH dataset, also reported in Tab. 4, reveals nearly identical average performance (78.59 % vs. 78.64 %). This suggests that performance does not consistently improve with increased dataset size in this mixed-supervision setting. However, BBGH differs from BGH-A in both scale and composition, including differences in geographic region and annotation source. Overall, dataset scaling alone is not the primary driver of performance under these conditions, with annotation consistency and domain alignment likely playing a more important role.

Data augmentation introduces model-dependent effects. Lightweight architectures benefit noticeably from increased variability, while high-capacity models remain largely unaffected. This asymmetry suggests that augmentation primarily compensates for representational limitations rather than correcting annotation noise.

Overall, the results indicate that the dataset encodes reliable large-scale structure, while remaining limited by label consistency rather than data volume.

Two-Class Segmentation Increasing task complexity to road-sidewalk segmentation (two-class setting) exposes the limits of the annotation pipeline more

Table 4: Binary sidewalk segmentation mIoU (%) on BGH-A and BBGH, without (–) and with (✓) offline augmentation. Bold indicates the best performance per column.

Architecture	BGH-A		BBGH	
	–	✓	–	✓
U-Net	81.1	78.9	84.2	84.9
Fast-SCNN	69.5	75.8	75.3	70.1
DeepLabV3+	83.0	82.7	81.9	86.9
BiSeNetV2	76.7	81.1	66.6	79.2
Average	78.6	79.6	77.0	80.3

Table 5: Two-class road–sidewalk segmentation mIoU (%) on BGH-A, without (–) and with (✓) offline augmentation. Bold indicates the best performance per column.

Architecture	BGH-A	
	–	✓
U-Net	59.2	66.6
Fast-SCNN	58.4	47.8
DeepLabV3+	72.9	74.3
BiSeNetV2	62.1	56.0
Average	62.2	61.2

clearly. Average performance decreases to 62.16 % mIoU, as reported in Tab. 5, reflecting increased ambiguity in class separation rather than failure of spatial learning.

Despite this drop, relative model rankings remain stable. This suggests that architectural capacity continues to determine performance ordering even under increased semantic noise, indicating that the dataset preserves consistent learning signals across models.

Augmentation effects become less predictable in this setting. Some architectures benefit, while others degrade slightly, implying that variability in training data cannot fully compensate for inherent class ambiguity in ground-plane segmentation.

Qualitative patterns remain consistent: road and sidewalk remain structurally well separated, while errors concentrate in minority or visually ambiguous classes. This reinforces the conclusion that the primary limitation lies in fine-grained semantic resolution rather than global scene understanding.

Despite reduced absolute scores, all models converge reliably, indicating that the annotation pipeline provides a stable though imperfect supervisory signal even under two-class conditions.

Overall, road–sidewalk results highlight a central trade-off: the pipeline is robust in capturing structural layout, but limited in resolving fine-grained semantic boundaries under domain shift in pedestrian scenes.

6 Conclusion

This work investigated whether open-vocabulary grounding and prompt-based segmentation models can generate pixel-level annotations for pedestrian-centered urban scenes under strong domain shift, with a focus on assistive vision scenarios where labeled data is scarce. We combine language-guided localization with promptable segmentation and introduce a conservative filtering strategy to improve annotation reliability while minimizing manual supervision.

Our results show that foundation-model-based annotation pipelines are viable under substantial domain shift in pedestrian scenes. The generated labels preserve large-scale scene structure and enable strong downstream performance in binary sidewalk segmentation across architectures, despite only moderate agreement with manually annotated ground truth. This indicates that foundation models can provide useful supervisory signals for initializing semantic understanding systems in low-resource, accessibility-oriented settings.

We further observe that increasing dataset size does not consistently improve performance under mixed supervision, suggesting that annotation quality and domain alignment are at least as important as scale in this setting. This is particularly relevant for assistive vision applications, where data collection is inherently limited and scaling cannot easily compensate for label noise. Error analysis indicates that remaining limitations mainly stem from semantic grounding under domain shift, particularly in visually ambiguous ground-plane regions and small-scale objects. Since the annotation pipeline intentionally uses only simple class-name prompts throughout all experiments, these results isolate the effects of domain shift without introducing prompt engineering as an additional optimization variable. More descriptive prompting strategies may improve semantic grounding and reduce rejection rates, but remain future work.

From an application perspective, the most relevant limitation is semantic confusion between adjacent traversable regions, especially road and sidewalk. Such errors are safety-relevant in assistive navigation, as they directly affect the interpretation of walkable space. While the proposed conservative postprocessing step removes inconsistent predictions at image level, residual ambiguity persists in the road-sidewalk setting. Despite these limitations, the approach provides a practical trade-off between annotation cost and supervision quality. It significantly reduces manual effort while still producing datasets that are effective for training segmentation models.

Overall, open-vocabulary annotation provides a scalable mechanism for constructing pedestrian scene datasets under data scarcity. Its effectiveness is constrained by the reliability of semantic grounding under domain shift, particularly for visually similar traversable regions. This makes it well-suited for dataset initialization in pedestrian vision pipelines, but insufficient as a standalone solution for safety-critical deployment.

References

1. Bhole, S., Dhok, A.: Deep learning based object detection and recognition framework for the visually-impaired. In: 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC). pp. 725–728 (2020). <https://doi.org/10.1109/ICCMC48092.2020.ICCMC-000135>
2. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Eur. Conf. Comput. Vis. pp. 833–851. Springer International Publishing (2018). https://doi.org/10.1007/978-3-030-01234-2_49
3. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes dataset for semantic urban scene understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3213–3223 (2016). <https://doi.org/10.1109/CVPR.2016.350>
4. De Brabandere, B., Debals, O., Staelens, T.: Sidewalk semantic dataset (2025), <https://huggingface.co/datasets/segments/sidewalk-semantic>
5. G M, V., Pereira, B., Little, S.: Urban footpath image dataset to assess pedestrian mobility. In: Proceedings of the 1st International Workshop on Multimedia Computing for Urban Data. pp. 23–30. Association for Computing Machinery (2021). <https://doi.org/10.1145/3475721.3484313>
6. Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Garcia Rodriguez, J.: A review on deep learning techniques applied to semantic segmentation. arXiv preprint arXiv:1704.06857 (2017). <https://doi.org/10.48550/arXiv.1704.06857>
7. Geyer, J., Kassahun, Y., Mahmudi, M., Ricou, X., Durgesh, R., Chung, A.S., Hauswald, L., Pham, V.H., Mühlegg, M., Dorn, S., Fernandez, T., Jänicke, M., Mirashi, S., Savani, C., Sturm, M., Vorobiov, O., Oelker, M., Garreis, S., Schubert, P.: A2D2: Audi autonomous driving dataset. arXiv preprint arXiv:2004.06320 (2020). <https://doi.org/10.48550/arXiv.2004.06320>
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 770–778 (2016). <https://doi.org/10.1109/CVPR.2016.90>
9. Hoyer, L., Dai, D., Van Gool, L.: DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9924–9935 (2022). <https://doi.org/10.1109/CVPR52688.2022.00969>
10. Hwang, H., Jung, H.T., Giudice, N.A., Biswas, J., Lee, S.I., Kim, D.: Towards robotic companions: Understanding handler-guide dog interactions for informed guide dog robot design. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (2024). <https://doi.org/10.1145/3613904.3642181>
11. Kim, S.Y., Cho, K.: Usability and design guidelines of smart canes for users with visual impairments. *International Journal of Design* **7**, 99–110 (2013)
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: Int. Conf. Comput. Vis. pp. 3992–4003 (2023). <https://doi.org/10.1109/ICCV51070.2023.00371>
13. Liao, Y., Xie, J., Geiger, A.: KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(3), 3292–3310 (2023). <https://doi.org/10.1109/TPAMI.2022.3179507>

14. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: Eur. Conf. Comput. Vis. pp. 38–55. Springer Nature Switzerland (2024). https://doi.org/10.1007/978-3-031-72970-6_3
15. Mumuni, A., Mumuni, F., Gerrar, N.K.: A survey of synthetic data augmentation methods in computer vision. *Machine Intelligence Research* **21**, 1–39 (2024). <https://doi.org/10.1007/s11633-022-1411-7>
16. Poudel, R.P., Liwicki, S., Cipolla, R.: Fast-SCNN: Fast semantic segmentation network. arXiv preprint arXiv:1902.04502 (2019). <https://doi.org/10.48550/arXiv.1902.04502>
17. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded SAM: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024). <https://doi.org/10.48550/arXiv.2401.14159>
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A. (eds.) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. pp. 234–241. Springer International Publishing (2015). https://doi.org/10.1007/978-3-319-24574-4_28
19. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: Inverted residuals and linear bottlenecks. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 4510–4520 (2018). <https://doi.org/10.1109/CVPR.2018.00474>
20. Wang, M., Deng, W.: Deep visual domain adaptation: A survey. *Neurocomputing* **312**, 135–153 (2018). <https://doi.org/10.1016/j.neucom.2018.05.083>
21. Wang, W., Jing, B., Yu, X., Sun, Y., Yang, L., Wang, C.: YOLO-OD: Obstacle detection for visually impaired navigation assistance. *Sensors* **24**(23) (2024). <https://doi.org/10.3390/s24237621>
22. Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., Sang, N.: BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation. *Int. J. Comput. Vis.* **129**(11), 3051–3068 (2021). <https://doi.org/10.1007/s11263-021-01515-2>