Tactile Path Guidance via Weakly Supervised Visual Attention

Suayder M. Costa*, Rafael J. P. Damaceno*, Henrique Morimitsu†, Roberto M. Cesar Jr.*
*Institute of Mathematics and Statistics, University of São Paulo, São Paulo, SP, Brazil,
†School of Computer and Communication Engineering, University of Science and Technology Beijing, China
Email: suayder@ime.usp.br, rafael.damaceno@ime.usp.br, hmori@ustb.edu.cn, rmcesar@usp.br

***Abstract*—Tactile paving is a structure available on sidewalks that supports visually impaired people in walking independently. Maintaining this and other structures in the urban environment is essential for pedestrians' safety and well-being. Computational solutions for their assessment, an essential part of urban infrastructure maintainability, require the availability of specific data, which is costly and time-consuming to obtain. In this context, this work proposes using the SAM2 segmentation model as a basis to enhance saliency detection models. These saliency models can then detect important urban features more accurately and with lower costs, making them ideal for deploying on mobile devices. This paper illustrates how this approach can improve the identification of tactile paving to aid the mobility of visually impaired users while also collecting summarized data about the conditions of these structures.**

I. Introduction

Tactile paving assists visually impaired individuals in navigating public spaces. This urban infrastructure element alerts pedestrians to changes in direction, the presence of curbs, and other important features. To ensure its effectiveness, it's essential to maintain tactile paving, sidewalks, and other urban infrastructures in good condition.

Regular inspections are vital for promptly identifying deteriorating sidewalk structures and objects that can pose obstacles to pedestrians. While essential, these assessments can be costly and time-consuming, often leading to inadequate maintenance, which hinders walkability. As reported in various studies, the lack of maintenance, resulting in deteriorated sidewalks, increases the risk of falls and consequently leads to higher rates of hospital admissions [1], [2].

To streamline the assessment of sidewalks and related structures, computational tools aimed at describing and evaluating the built environment have been developed [3]. Moreover, a few studies have generated datasets highlighting obstacles to walkability [4], [5], [6], [7], which were specifically designed to improve the detection of objects in the context of people with disabilities (e.g., blind individuals and those with reduced mobility).

However, there is a lack of data regarding ground conditions that specifically support the walking of visually impaired people, such as tactile paving. Moreover, many studies utilize object detection techniques to identify the contents of a scene. On the one hand, these solutions can facilitate the detection and counting of obstacles; on the other hand, they are unsuitable for characterizing the ground itself, such as determining if some object represents a safety issue [8] and detecting the conditions of tactile paving.

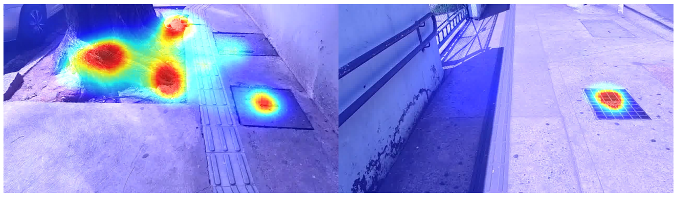

Fig. 1. Salience map obtained from the SUM model proposed by Hosseini et al. [9].

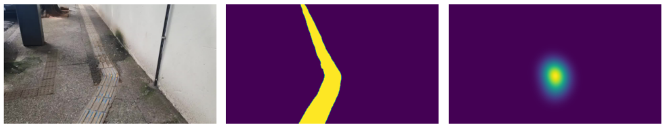

Fig. 2. Example of generated pseudo-labels: the first image shows the original frame, the middle image displays the segmentation mask, and the last is the salience map, which highlights a curve in the scene to assist pedestrian navigation.

A possible strategy to address this issue, but one that requires the existence of high-quality annotated data, is the use of visual salience models, a class of neural networks capable of identifying what is visually relevant to a person [9], [10]. However, most of them lack specialization for pedestrians and their specific needs.

Figure 1 illustrates how saliency models identify important regions of pedestrian footage images from the SideSeeing Hospital Dataset [11].

The examples in Figure 1 indicate that saliency maps are useful for identifying regions of interest on the ground, such as manhole covers and pavement failures. These results suggest a potential strategy for using such models to focus on sidewalk structures or problems and generate videos that summarize accessibility issues, for instance.

In this study, we propose a visual attention method to efficiently locate the tactile paving in urban scene footage. We leverage the SAM2 model [12] to facilitate the generation of labeled video data and then use it to train a specialized saliency detector for tactile pavements. The saliency map can identify the tactile paving and even pinpoint important sections, such as changes in direction (see Figure 2). Saliency detection may also require fewer resources than complex segmentation models, which makes them more suitable for mobile applications. The advantages of the resulting solution are twofold: (1) helping visually impaired users navigate complex environments by providing guidance from portable devices, and (2) facilitating the evaluation of sidewalk conditions by focusing only on more relevant data around the saliency points (e.g., by removing unimportant data with video cropping [13]).

II. Background

A. Pseudo-labels

The success of deep learning relies heavily on large-scale, high-quality annotations, which can be a laborious task. To address this challenge, a common approach is to generate weakly labeled data, balancing annotation costs with accuracy [14].

This strategy enables researchers, for instance, to annotate samples of a dataset using large language models and other frameworks, such as segmentation models, which associate each pixel of an image or video frame with a label. To train salience detection with weakly labeled data, recent works leverage edge maps and class activation maps to generate pseudo-labels as feedback to self-supervised learning [15].

Similarly to those approaches, our work relies on weakly labeling data to fine-tune visual saliency models.

B. Tactile Paving Detection

Many approaches can be used to detect tactile paving, with object detection being a common method for tackling this problem. In this context, several studies have proposed using MobileNetV3 due to its suitability for smartphone deployment, enabling solutions to provide real-time detection and user feedback [16].

Ito et al. [17] presented a detection method based on dynamic statistical thresholding in the HSV color space. In a broader context, Xia et al. [5] developed a new dataset called WOTR, an acronym for "Walk On The Road", and YOLO-based models to better detect the 15 categories of obstacles they proposed (one of these categories is tactile paving). The work by Ghilardi et al. [18] explored image processing techniques based on Canny and blur filters, followed by applying Hough Lines Transform to detect tactile paving.

Unlike these studies, we use segmentation masks of tactile paving obtained through the SAM2 model and employ image processing techniques such as Hough Lines Transform and lines intersection detection to generate density maps.

III. Method

This section introduces the dataset used in this work, along with the pipeline adopted for data annotation, training, and evaluation of salience detection models. We also discuss the metrics employed to evaluate the models. The goal is to leverage the segmentation masks quickly generated through the SAM2 model and process them to feed the training of saliency detection models with a context of interest.

TABLE I
Total duration (s), and total number of frames of each sample extracted from the SideSeeing Hospital Dataset.

ID	Duration (s)	Total frames
JUNDIAI-HSV#BLOCK01	241.94	7,259
JUNDIAI-HSV#ROUTE01	139.82	4,195
JUNDIAI-HSV#ROUTE02	69.66	2,090
SANTOS-HM#BLOCK01	321.43	9,644
All	772.85	23,188

A. Dataset

The data utilized in this research is a subset of a dataset generated as part of our ongoing project named SideSeeing [11], an initiative that aims to facilitate the collection and analysis of multimodal content pertaining to sidewalks. The SideSeeing collection framework utilizes chest-mounted smartphones to record videos (and sensor data) focused on ground-level features. This approach enables us to capture a comprehensive perspective of sidewalk conditions and surrounding environments.

The subset used in our study consists of four videos totaling 23 thousand frames, filmed in two Brazilian cities (see Table I). The videos were recorded at 30 frames per second with a resolution of 1280 by 720 pixels and depict routes traveled by pedestrians near hospitals.

B. Pipeline

The pipeline adopted in this work is composed of four stages, as follows: a) Data Preparation; b) Pseudo-Labeling; c) Model Training; and d) Model Evaluation (see Figure 3).

1) Data Preparation: The first stage is based on a tool developed to collect mouse clicks while watching videos. The tool plays a video and prompts the user to click on a region in the screen every 60 frames. In this study, the region clicked must contain tactile paving. The output is a collection of (x,y) positions to be used as a prompt to SAM2.

2) Pseudo-labeling: In the second stage, the collected clicks from the previous step guide the SAM2 model in generating a segmentation mask that represents the tactile surface as part of our experiment. Attention maps assist in identifying the most relevant objects in an image by generating a density map. However, creating these maps involves a labor-intensive process of highlighting key objects. To make this faster, we leverage the segmentation masks generated by SAM2, as described in the previous step. We chose this model due to its zero-shot generalization capability, as highlighted in the work by Ravi et al. [12]. The segmentation masks generated by SAM2 indicate the location of the tactile paving in the images, as illustrated in Figure 4. In the next stage, we

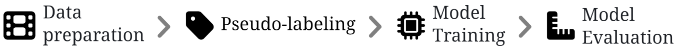

Fig. 3. Pipeline adopted in this work: a) data preparation; b) pseudo-labeling; c) model training; and d) model evaluation.

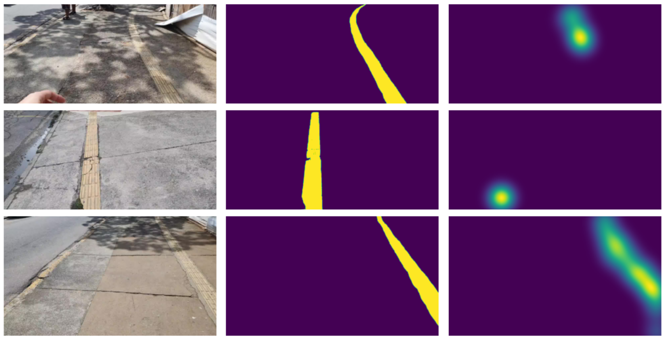

Fig. 4. For each row, three images are presented: the original image; the segmented tactile paving mask generated by SAM2; and the corresponding attention map. In the first row, the attention maps highlight curves. In the second and third rows, there are no curves in the scene, but ground irregularities are visible, especially in the second row.

use those masks to generate density maps that highlighting specific regions of the tactile paving - we focus on identifying curves and irregularities. We first extract the skeleton of the mask using a skeletonization technique. Next, we apply the Hough Transform to the extracted skeleton to detect straight lines. Following this, we identify intersections on the detected lines, which represent points of curvature or irregularities in the skeleton. If no intersection is found, we use the center of mass of the segmentation mask as the attention point. Finally, we employ the DBSCAN[1] algorithm to remove outlier points, retaining only dense point clouds. The resulting intersections are used to generate a Gaussian centered at each point, and the mean of these Gaussians is computed to create the final image corresponding to the pseudo labels.

3) Model Training: We use our generated pseudo-labels to train a saliency model to identify tactile paving. We chose the SUM [9] saliency model due to its recent success and good performance on five visual saliency benchmarks. The model was trained with default parameters, except for the number of epochs, which was changed to 20.

4) Model Evaluation: The metrics used to evaluate the trained model were Kullback–Leibler Divergence (KLD), correlation coefficient (CC), and Similarity (SIM), as is common in most visual saliency research.

IV. Results

This section presents some results and analysis of the trained SUM in tactile paving saliency data. The evaluation results demonstrated the model's effectiveness in detecting tactile salience through pseudo-labels. As shown in Figure 5, the model successfully highlights salient areas of the tactile path, which can help visually impaired users by indicating points of interest or critical navigational cues, such as changes in the tactile paving that guide movement through complex environments. Additionally, it enables the evaluation of sidewalk conditions by identifying areas where tactile paving may be missing or deteriorating.

We also assessed the model's performance quantitatively. The model achieved a CC of 0.842, indicating a strong positive linear relationship between predicted and actual values. However, the KLD value of 0.435 suggests a moderate divergence between the predicted and true mask distributions. Finally, a SIM of 0.533 reflects moderate similarity, highlighting discrepancies that could be addressed to improve the model's overall predictive accuracy and alignment with the true outcomes. Together, these metrics illustrate the model's strengths and areas for enhancement, paving the way for further refinement.

V. Conclusion

In this paper, we employed SAM2 model to generate pseudo labels for visual saliency task aimed at detecting tactile paving. Our approach involved selectively highlighting important areas within the tactile paving, including irregularities like curves or breaks, or just the center of the tactile if it does not presents irregularities.

The findings indicate that our method not only improves an accurate saliency model but also has the potential to inform maintenance strategies for urban infrastructure as well as guide visually impaired pedestrians by highlighting important regions that affect the walk. Future work will explore different deep-learning architectures and extend this solution to identify other sidewalk issues, such as potholes and surface irregularities, further contributing to urban safety and pedestrian well-being.

Acknowledgment

The authors would like to thank FAPESP (grants #2022/12204-9, #2022/15304-4), CNPq, CAPES, FINEP, USP, UGPN and MCTI PPI-SOFTEX (TIC 13 DOU 01245.010222/2022-44) for their financial support for this research. HM is supported by the Beijing Natural Science Foundation (IS23060) and the Fundamental Research Funds for the Central Universities (FRF-TP-22-048A1).

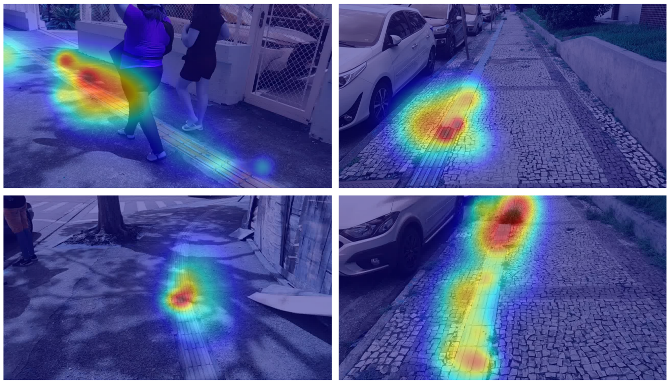

Fig. 5. Inference images from the fine-tuned model. The inference shows the model is accurate in recognizing tactile in the scene.

[1] https://scikit-learn.org/stable/modules/clustering.html#dbscan, accessed on September 27, 2024.

References

[1] D. R. d. O. M. Abreu, E. S. Novaes, R. R. d. Oliveira, T. A. d. F. Mathias, and S. S. Marcon, “Internação e mortalidade por quedas em idosos no brasil: análise de tendência,” *Ciência & Saúde Coletiva*, vol. 23, no. 4, p. 1131–1141, Apr 2018. [Online]. Available: https://doi.org/10.1590/1413-81232018234.09962016

[2] S. Lee, X. Ye, J. W. Nam, and K. Zhang, “The association between tree canopy cover over streets and elderly pedestrian falls: A health disparity study in urban areas,” *Social Science & Medicine*, vol. 306, p. 115169, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0277953622004750

[3] W. Shi, M. F. Goodchild, M. Batty, M.-P. Kwan, A. Zhang *et al.*, *Urban informatics*. Springer, 2021.

[4] T. Baba, “Vidvip: Dataset for object detection during sidewalk travel,” *Journal of Robotics and Mechatronics*, vol. 33, no. 5, pp. 1135–1143, 2021.

[5] H. Xia, C. Yao, Y. Tan, and S. Song, “A dataset for the visually impaired walk on the road,” *Displays*, vol. 79, p. 102486, 2023.

[6] W. Tang, D.-e. Liu, X. Zhao, Z. Chen, and C. Zhao, “A dataset for the recognition of obstacles on blind sidewalk,” *Universal Access in the Information Society*, vol. 22, no. 1, pp. 69–82, mar 2023. [Online]. Available: https://doi.org/10.1007/s10209-021-00837-9

[7] K. Park, Y. Oh, S. Ham, K. Joo, H. Kim, H. Kum, and I. S. Kweon, “Sideguide:a large-scale sidewalk dataset for guiding impaired people,” in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 10 022–10 029.

[8] I. Vozniak, P. Müller, L. Hell, N. Lipp, A. Abouelazm, and C. Müller, “Context-empowered visual attention prediction in pedestrian scenarios,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 950–960.

[9] A. Hosseini, A. Kazerouni, S. Akhavan, M. Brudno, and B. Taati, “Sum: Saliency unification through mamba for visual attention modeling,” *arXiv preprint arXiv:2406.17815*, 2024.

[10] S. Jain, P. Yarlagadda, S. Jyoti, S. Karthik, R. Subramanian, and V. Gandhi, “Vinet: Pushing the limits of visual modality for audio-visual saliency prediction,” 2021.

[11] R. Damaceno, L. Ferreira, F. Miranda, M. Hosseini, and R. Cesar Jr, “Sideseeing: A multimodal dataset and collection of tools for sidewalk assessment,” *arXiv preprint arXiv:2407.06464*, 2024.

[12] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, “Sam 2: Segment anything in images and videos,” *arXiv preprint arXiv:2408.00714*, 2024.

[13] S. Costa, R. Damaceno, and R. Cesar Jr, “Video cropping using salience maps: a case study on a sidewalk dataset,” *2024 37th SIBGRAPI-WiP conference on Graphics, Patterns and Images, Workshop of Works in Progress*, 2024.

[14] Z. Ren, S. Wang, and Y. Zhang, “Weakly supervised machine learning,” *CAAI Transactions on Intelligence Technology*, vol. 8, no. 3, pp. 549–580, 2023.

[15] R. Yasarla, R. Weng, W. Choi, V. M. Patel, and A. Sadeghian, “3sd: Self-supervised saliency detection with no labels,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 313–322.

[16] W. Chen, Z. Xie, P. Yuan, R. Wang, H. Chen, and B. Xiao, “A mobile intelligent guide system for visually impaired pedestrian,” *Journal of Systems and Software*, vol. 195, p. 111546, 2023.

[17] Y. Ito, C. Premachandra, S. Sumathipala, H. W. H. Premachandra, and B. Sudantha, “Tactile paving detection by dynamic thresholding based on hsv space analysis for developing a walking support system,” *IEEE Access*, vol. 9, pp. 20 358–20 367, 2021.

[18] M. C. Ghilardi, R. C. Macedo, and I. H. Manssour, “A new approach for automatic detection of tactile paving surfaces in sidewalks,” *Procedia Computer Science*, vol. 80, pp. 662–672, 2016, international Conference on Computational Science 2016, ICCS 2016, 6-8 June 2016, San Diego, California, USA.

Bio

Suayder M. Costa Final-year student of Master’s degree in Computer Science at University of São Paulo. Since graduation, he has worked with computer vision, and then worked on developing computer vision solution in companies. Currently, he works with Urban Informatics in his Master’s degree and also works at research institute with development of computer vision solutions to recognize daily family interactions.

Rafael J. P. Damaceno received his PhD in Computer Science from the Federal University of ABC. He has experience in the fields of Human-Computer Interaction and Scientometrics. Since his master’s, he has participated in studies on mobile device accessibility. Currently, he is a postdoctoral researcher at the University of São Paulo, working on Computer Vision and Deep Learning for Urban Informatics.

Henrique Morimitsu received the MSc and PhD degrees in Computer Science from the Institute of Mathematics and Statistics, University of São Paulo, Brazil. He is currently a lecturer at the School of Computer and Communication Engineering at the University of Science and Technology Beijing. His research interests include motion estimation and the development of efficient models for edge devices.

Roberto M. Cesar Jr. is a Full Professor of Computer Science at the University of São Paulo. His research interests include computer vision, pattern recognition, image processing, bioinformatics, and eScience. He has a Ph.D. in Physics and has served as a member of various academic committees and as the director of the eScience Research Center at USP.

Rationale

Sidewalk inspections rely heavily on manual methods, which are time-consuming, costly, and often insufficient to ensure the accessibility of urban environments. Well-maintained sidewalks play an important role in promoting pedestrian safety and enhancing the overall quality of life. By leveraging visual saliency models, this project aims to develop a method for accurately identifying sidewalk structures such as tactile paving. The goal is to enable the generation of informative videos that highlight potential hazards, such as deteriorated tactile paving or obstructions, empowering policymakers to make informed decisions.