Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.5895
This work is licensed under a Creative Commons Attribution 4.0 International License.

Salience prediction methods for video cropping in sidewalk footage

Suayder M. Costa [**University of São Paulo** | ***suayder@ime.usp.br***]
Rafael J. P. Damaceno [**University of São Paulo** | ***rafael.damaceno@ime.usp.br***]
Henrique Morimitsu [**University of Science and Technology Beijing** | ***hmori@ustb.edu.cn***]
Roberto M. Cesar-Jr [**University of São Paulo** | ***rmcesar@usp.br***]

Institute of Mathematics and Statistics, University of São Paulo, R. do Matão, 1010, São Paulo, SP, 05508-090, Brazil.

Received: 12 April 2025 • **Accepted:** 29 October 2025 • **Published:** 15 April 2026

Abstract The condition of urban infrastructure is an important aspect in ensuring the safety and well-being of pedestrians. This is especially important around public health facilities, such as sidewalks surrounding hospitals. Computational tools have already demonstrated their potential in this context, including surface material classification and obstacle detection; however, most solutions require labeled data, which is costly and time-consuming. To address this gap, we propose two strategies for salience prediction in videos that reduce the dependence of manual labeling. The first leverages human visual attention, converting user clicks into attention maps. The second employs the SAM2 model to generate labeled video data more efficiently. The outputs of this process are used to train specialized saliency detectors to identify general cracks, surface defects, and key sections of tactile paving, such as directional changes. Also, we apply these saliency models to video cropping in order to highlight the most relevant areas within each frame. This approach enables content-aware video retargeting, supports object-focused attention, and facilitates sidewalk condition analysis by emphasizing defects and potential hazards. This work presents the following contributions: (1) development of a click-based video annotation tool, (2) development of two saliency detection strategies for sidewalks video cropping, (3) training and evaluation of saliency models for sidewalk structure analysis, and (4) successful application of these introduced methods for video cropping. Our experimental results showed that saliency models were able to highlight relevant information in urban environments, achieving an AUC of 0.582 in the best case for human-based attention and 0.914 for tactile-based attention, thereby enhancing assistive technologies for visually impaired individuals.

Keywords: Salience Prediction, Sidewalk, Tactile Paving, Video Cropping

Published under the Creative Commons Attribution 4.0 International Public License (CC BY 4.0)

1 Introduction

With the rise of social networks, smartphone usage and mobile video recording, there is an opportunity to leverage crowd-collected data to enhance the decision-making process. This is particularly relevant in Urban Informatics studies, where image-based datasets help provide information on land use [Miranda *et al*., 2020], sidewalk infrastructure [Saha *et al*., 2019], pedestrian obstacles [Park *et al*., 2020], among others.

An important aspect of urban infrastructure is its maintenance. Regular inspections are essential, as they help identify deteriorating conditions and objects that can hinder pedestrian movement. However, while necessary, these assessments can be costly and time-consuming, often resulting in inadequate maintenance and reduced walkability [Yussif *et al*., 2024]. In addition, deteriorated sidewalks increase the risk of falls, leading to higher rates of hospital admission [Abreu *et al*., 2018; Lee *et al*., 2022]. For visually impaired people, tactile paving serves as a guide to navigating public spaces, signaling changes in direction and the presence of curb ramps. However, to ensure its effectiveness, it must be properly maintained [Yussif *et al*., 2024].

In this context, many computational tools support the description and evaluation of the built environment [Shi *et al*., 2021; Zünd and Bettencourt, 2021]. Regarding sidewalks, studies have generated datasets that highlight obstacles for visually impaired pedestrians and individuals with reduced mobility [Park *et al*., 2020; Baba, 2021; Xia *et al*., 2023; Tang *et al*., 2023]. Despite the existence of these datasets, there is a lack of labeled data on ground conditions, especially based on egocentric videos, a category of footage recorded from the first-person point of view. 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 [Hosseini *et al*., 2024; Jain *et al*., 2021]. However, most of them lack specialization for the urban infrastructure context of pedestrians.

In this study, we propose two strategies for salience prediction in egocentric videos. Figure 1 shows the pipeline of the proposed method, which is organized in two main phases: model learning and model prediction. Model learning involves multimodal data acquisition using the SideSeeing technology [Damaceno *et al*., 2024] and dataset annotation.

This paper explores two saliency map generation strategies: human attention and tactile paving attention, which are used to train saliency detection models. During the model prediction phase, the saliency maps produced by the previous models are employed for performing video cropping.

The first saliency generation strategy relies on human visual attention, applying post-processing techniques to transform user clicks into attention maps. The second strategy leverages the SAM2 model [Ravi *et al*., 2024] along with post-processing techniques to generate labeled video data more efficiently. In this approach, we consider the segmentation model's output as weak labels for the post-processing techniques. The outputs of both strategies are used to train specialized saliency detectors. In the first approach, the saliency map highlights general cracks, surface defects, or obstacles. In the second, it identifies tactile paving and marks key sections. These models can serve as the foundation for video cropping tools, leveraging salience maps to identify the most significant parts of videos and adjust frames to retain only the areas of interest.

The cropping process is a linear operation modeled as an optimization problem, aiming to maximize attention within a given bounding box shape. This process can be used, for example, to remove unnecessary parts of a video, change its orientation to suit different screen formats in images [Apostolidis and Mezaris, 2021] or video [Le *et al*., 2024], focus attention on specific actions in the scene [Jana *et al*., 2021], among other applications. A common challenge in this process is to determine the best cropping window in each frame, which can be subjective or guided by techniques such as object tracking [Taylor *et al*., 2016] and visual saliency [Ota *et al*., 2024].

This work extends the preliminary advances described in Costa *et al*. [2024]. The main contributions are summarized as: (1) development of a click-based video annotation tool, (2) development of two new saliency detection strategies for sidewalks video cropping, (3) training and evaluation of saliency models to identify sidewalk structures and associated issues, and (4) application of the generated models in a video cropping framework.

The paper is organized as follows. Section 2 reviews recent studies on salience prediction, video cropping, and sidewalk analysis. Section 3 details the dataset used in this work, presents the human attention and tactile paving methods for generating saliency maps, and explains the process of video cropping. Section 4 reports the outputs from both methods and provides a qualitative analysis of a video cropping application. Finally, Section 5 concludes the paper and suggests potential future work.

2 Related works

This section presents studies on salience prediction, video cropping, and sidewalk analysis. We focus particularly on salience models based on human visual attention and on tactile paving detection, explaining the rationale behind the strategies developed in our study.

2.1 Salience prediction and video cropping

Several studies on human visual attention have utilized machine-learning techniques for salience prediction. For instance, the models named SalEMA and SalCLSTM adapted neural networks to incorporate temporal information, demonstrating enhanced performance in handling video salience prediction [Linardos *et al*., 2019].

Later, ViNet [Jain *et al*., 2021] was proposed as an architecture for audio-visual and non-audio-visual saliency prediction and was successful in the task by using a 3D fully convolutional architecture design. This 3D architecture is also studied with hierarchical learning and domain adaptation to the same applications [Bellitto *et al*., 2021].

Another task related to visual human attention is movie editing. Strong correlations have been identified between movie editing annotations and spectators' gaze distributions [Bruckert *et al*., 2023], which could help optimize editing based on human visual attention. Similarly, our study investigates this concept by capturing mouse clicks to track and generate visual attention distributions. We then use this information to fine-tune models using a specific dataset related to sidewalk footage.

Visual attention has been used to assist in video cropping, which includes a task known as reframing (i.e., changing the video orientation from landscape to portrait and vice versa). For this task, a well-established model named SalCrop is based on spatio-temporal salience [Zhang *et al*., 2022]. Their proposed framework is built through four modules: video scene detection, video salience prediction, adaptive cropping, and video codec. The first module is responsible for splitting the data into short sequences; the second module identifies salient content in the frames; the third handles the cropping task, finding the optimal strategy; and the last module manages the encoding and decoding of the video content. This pipeline is well established and used in other works [Apostolidis and Mezaris, 2021; Tang *et al*., 2022; Ota *et al*., 2024].

Moreover, recent frameworks leverage the temporal component of videos Tang *et al*. [2022]; Imani and Islam [2024]. In this case, one such solution is based on a mechanism that detects jumping frames and smooths their importance, which arguably reduces the jitter of resized videos [Tang *et al*., 2022]. Similarly, Zhang *et al*. [2022] also includes an initial stage for scene detection to split the videos into short sequences, followed by a salience detection module.

Unlike these studies, our work focuses on fine-tuning models for egocentric urban footage to highlight salient components of urban environments that are most relevant to pedestrian experience and spatial navigation. In addition, as a case study, we apply these models to video cropping to emphasize important sidewalk segments for pavement analysis.

2.2 Sidewalk analysis and tactile paving

In computer vision, various methods can be used for sidewalk analysis, particularly to assess their features and conditions. One group of studies focuses on using image processing techniques or adapting deep learning models to detect tactile paving and other elements. Another group emphasizes creating datasets to support the identification of sidewalk struc-

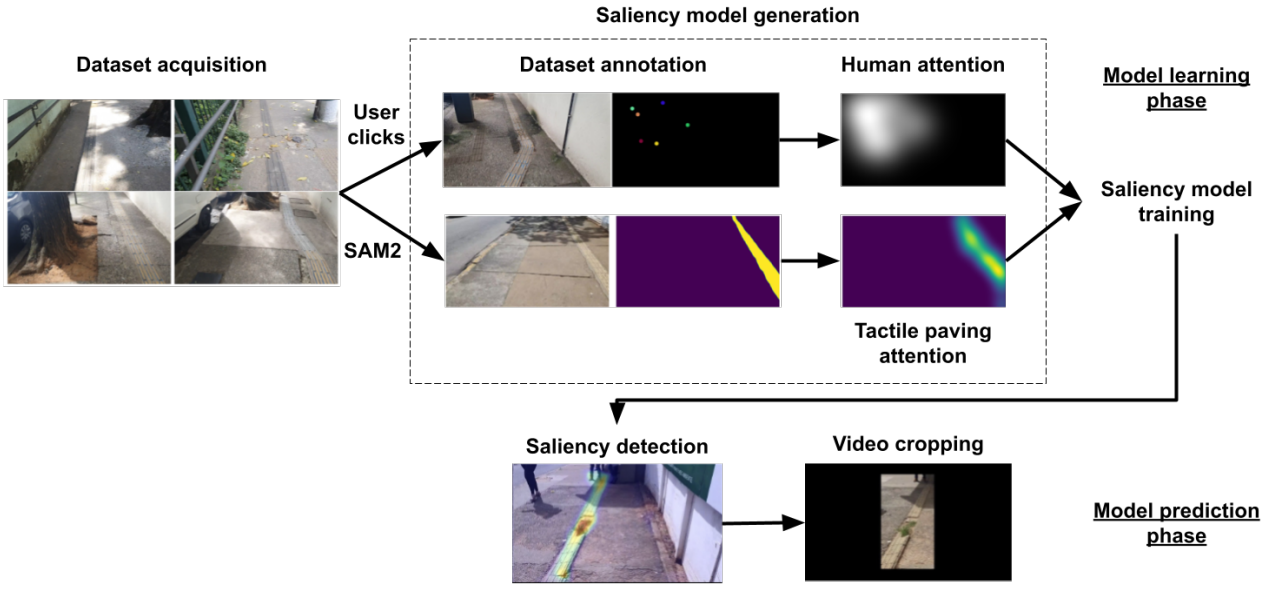


Figure 1. Overview of the proposed pipeline. There are two strategies for saliency generation. Human attention is generated from user clicks while tactile paving attention uses SAM2 segmentation masks. Models trained on these attention masks are used to detect saliency maps and applied for video cropping.

tures, including tactile paving, and objects that may hinder walkability.

The work by Ghilardi *et al*. [2016], for example, explored Canny and blur filters, followed by the application of the Hough Line Transform to detect tactile paving. The authors' approach achieved an accuracy of 88.48% on a set of images captured specifically for the study. The images were taken by users using a smartphone camera positioned 1 meter above the ground and angled at 45 degrees.

Ito *et al*. [2021] presented a tactile paving detection method based on dynamic statistical thresholding in the HSV color space. The authors designed a device composed of a depth camera and a microcontroller board. The images captured by the camera are cropped by 30% on all sides. The straight border of the tactile paving is detected using the Hough Line Transform. The study reports an accuracy of 91.65% for detecting tactile paving in a set of 870 images captured worldwide.

Regarding deep learning models, the work by Chen *et al*. [2023], for example, proposes a system based on MobileNet to detect tactile paving in real time using a smartphone camera, helping pedestrians navigate by providing information through text messages. Niu and Bao [2024] proposed an architecture based on the Fast-SCNN to more effectively address the negative impact of lighting conditions on the segmentation masks of tactile paving. The study by Li *et al*. [2024] introduced the TPSegmentDiff, a diffusion-based model that includes a voting mechanism for segmenting tactile paving.

Another part of studies focus on the creation of datasets. This is the case of the work by Theodosiou *et al*. [2020], which developed a dataset of egocentric images focused on pedestrian walking barriers, such as cracks, potholes, tree, mail box, broken pavement, among others. Using this dataset, the authors trained and evaluated several deep neural networks, achieving an overall accuracy of 88.4% with the best-performing architecture.

Saha *et al*. [2019] created Project Sidewalk, a web-based initiative that enables users to label images from Google Street View while virtually exploring streetscapes. The available labels include: curb ramps, missing curb ramps, obstacles, no sidewalk, and surface problems. Users are invited to navigate the tool and click to mark where these issues occur.

The WOTR dataset, an acronym for Walk On The Road, was created by Xia *et al*. [2023]. The authors adapted YOLO-based models to classify 15 categories of objects related to pedestrians with visual impairments. One of these categories is tactile paving, which, depending on the sidewalk, may be inaccessible due to the presence of other objects that can act as obstacles, such as parked vehicles or roadblocks.

In addition, Tenji10K, introduced in Takano *et al*. [2024], is a dataset containing 10,000 path images organized into 20 sequences of first-person videos featuring tactile paving in Japan. The videos were recorded at a resolution of 640×480 pixels and 30 frames per second, using smart glasses positioned 1.7 meters above the ground. Each image in a sequence is accompanied by a mask representing the tactile paving, annotated using two boundary lines. This approach is possible because, in the first-person view with the user walking on the tactile paving, it appears as a region extending vertically from the bottom to the top of the image, making it representable by two lines.

Our study focuses on proposing robust methods for processing challenging egocentric videos captured by cameras mounted on pedestrian bodies to analyze pavement conditions. We identify the most relevant parts of the video using user clicks and segmentation models.

3 Materials and methods

We propose techniques to detect prominence in videos, whose results can be used in video cropping tools. Our work falls within urban informatics and utilizes a dataset of sidewalk footage recorded by pedestrians, focusing on ground-level features. Section 3.1 describes the dataset used in this study.

To gather saliency data, we developed two strategies, both initiated by user clicks, to identify regions of interest within the video. The first strategy is based on human visual attention, focusing on the subjectivity of users as they determine which elements in the video frames may pose obstacles or hinder pedestrian safety. Participants were instructed to watch the videos and click on elements that met these criteria. Section 3.2 describes this method in detail.

In the second strategy, described in Section 3.3, we combined segmentation models with pseudo-labeling techniques to enhance the identification of relevant areas in the scene. The goal is to detect specific structures and potentially identify ground irregularities. In this approach, participants were instructed to click exclusively on the tactile paving.

We apply our saliency maps to a cropping algorithm that selects the most informative region in each frame. By maximizing attention within fixed crop dimensions, the method ensures key elements – such as obstacles or tactile paving – remain visible, supporting safety-focused video tools. This new approach is described in Section 3.4.

3.1 Dataset

The data utilized in this research is part of the SideSeeing project[Damaceno *et al*., 2024], which aims to facilitate the collection and analysis of multimodal content related to sidewalks. The SideSeeing collection framework employs smartphones mounted on chest supports to capture video, audio, and sensor data (e.g., accelerometer, magnetometer) while walking through various environments. This approach allows for the detailed documentation of ground-level features and sidewalk conditions. The most recent dataset, in particular, focuses on video recordings taken during walks near hospitals and transportation hubs, providing valuable insights into diverse, real-world settings.

This dataset is compelling for several reasons: (1) it presents samples with real challenging conditions, such as motion blur and abrupt motions, caused by the pedestrian walking; (2) the ground-facing camera conditions make it ideal to capture and analyze paving conditions. Figure 2 displays some sample frames from this dataset to better illustrate the characteristics of the scene.

Moreover, this strategy differs from the traditional method of acquiring urban data, which is often collected from a car-centric perspective [Park *et al*., 2020]. By using video footage recorded by pedestrians, we obtain a more detailed view of pedestrian pathways, enabling an in-depth analysis of ground and sidewalk features.

The subset used in this study consists of nine video files recorded in three Brazilian cities. As shown in Table 1, the dataset includes a total of 65,000 frames (36 minutes of video) for the human visual attention-based method, and 32,000 frames (13.5 minutes of video) for the tactile paving-based method. The videos were recorded at 30 frames per second with a resolution of 1280 × 720 pixels. Seven videos were used in the human attention-based method, and five in the tactile paving-based method. Three videos overlap between the two approaches; in these cases, only the segments containing tactile paving were included in the tactile paving method.

Table 1. Dataset Summary Table - Total duration in seconds and total number of frames for each video sample extracted from the dataset, annotated using human attention and tactile paving methods.

	Human attention		Tactile paving	
ID	Duration	Frames	Duration	Frames
J-HSV-B01	241.94	7,259	241.94	7,259
J-HSV-R01	-	-	139.82	4,195
J-HSV-R02	-	-	69.66	2,090
S-CHE-B01	330.32	9,910	-	-
S-HM-B01	321.43	9,644	321.43	9,644
SP-HC-R01	239.77	7,194	-	-
SP-HC-R02	689.88	20,680	40	1,200
SP-HU-R01	190.71	5,722	-	-
SP-HU-R02	151.08	4,533	-	-
All	2,165.13	64,942	812.85	23,388

3.2 Saliency based on human attention

In everyday situations, many elements can capture human attention, and most studies explore this in generic videos and contexts [Jiang *et al*., 2015; Gitman *et al*., 2014]. But what specifically draws a person's attention while walking? We address this question by identifying attention points through saliency maps.

The goal is to develop a method based on human attention that can be applied more broadly, allowing the results to be used for extracting key information from egocentric videos. The following sections present details on the participants and labeling protocol, as well as salience map generation, model selection, and evaluation metrics.

Participants and labeling protocol

We developed an application to annotate videos by clicking on points of interest while following the video walkthrough. Eight participants annotated seven video files, generating a total of 10,685 clicks across different frames. We chose to register at least one click for every 60 video frames, inspired by previous studies on salience detection [Jiang *et al*., 2015; Kim *et al*., 2017]. The goal was for all annotators to label the same videos, expecting that their clicks converged and demonstrated consistency in attention patterns.

Each click produces a coordinate pair (x, y), representing the mouse position within the video frame at the time of the click. This position indicates the location identified as significant by the annotator. Our methodology is informed by studies exploring the relationship between mouse clicks and eye gaze in visual attention research [Chen *et al*., 2001; Zhu *et al*., 2023], as well as research suggesting that discrete clicks provide a clearer record of points of interest [Kim *et al*., 2017].

We adopted the following instructions for labeling the videos. "While viewing the videos, identify objects or areas that may impede the safe movement of pedestrians. To do this, click on these objects or areas on the screen. Consider diverse pedestrian profiles, including people with disabilities, the elderly, and children. At least one click must be made every 60 frames, a task that is automatically enforced by the tool, which pauses the video to prompt a click. Examples of obstacles include: a) Temporary objects: boxes, garbage bags, traffic cones, and similar items; b) Improperly placed

Figure 2. Eight frames showing the ground of sidewalks near hospitals in three Brazilian cities.

fixed elements: poorly positioned poles, signs, trash bins, and others; c) Holes and uneven surfaces, which are also classified as obstacles; and d) Other obstacles: bicycles on sidewalks, debris, or any object that forces pedestrians to alter their path."

These labeling instructions ensure that the annotations reflect a comprehensive view of pedestrian accessibility, acknowledging that the concept of an obstacle is not limited to just one category of object but encompasses a broader range of real-world challenges.

Salience map generation, model selection, and evaluation metrics

Our annotation protocol generates clicks, which are insufficient for training salience prediction models. Therefore, we adopt a strategy to convert these clicks into salience maps, which serve as ground truth for each frame in the dataset's videos [Jiang *et al*., 2015; Zhang *et al*., 2019].

First, we applied two-dimensional cubic interpolation to propagate each coordinate pair (x, y) across the video frames, ensuring a smoother spatial distribution of annotations. Next, we removed outlier points using a z-score filter, excluding points beyond 1.2 standard deviations from the mean. Finally, we generated the final salience maps using a Gaussian Mixture Model with a standard deviation of 6% of the video width, based on the methodology outlined in Gitman *et al*. [2014] and Lyudvichenko and Vatolin [2019]. Specifically, the salience map at each frame is computed as a sum of Gaussian functions centered at the annotated points, as formulated in Equations (1) and (2):

$$G(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x - x_0)^2 + (y - y_0)^2}{2\sigma^2}\right) \quad (1)$$

$$I(x, y) = \frac{\sum_i G_i(x, y)}{\max\left(\sum_i G_i(x, y)\right)} \quad (2)$$

where (x_0, y_0) is the center of a Gaussian, σ is the standard deviation, (x, y) are pixel coordinates in the frame grid, $G_i(x, y)$ represents the Gaussian function for the i-th center and $I(x, y)$ is the sum of the Gaussians – this sum is scaled to be in the range $[0, 1]$. This approach refines the raw annotations into continuous salience representations, effectively capturing the distribution of human attention. Figure 3 illustrates sample frames along with their corresponding salience maps.

Figure 3. Examples of annotated frames from our dataset. The first column displays the original video frame, the second column shows the click positions recorded by each annotator, and the third column presents the corresponding ground truth salience map.

To predict salience maps from a video, we adopted well-established model architectures from the literature, specifically leveraging encoder-decoder convolutional networks. We experimented with three top-performing saliency models: 1) ViNet [Jain *et al*., 2021] uses a simple encoder for feature extraction and the decoder infers a saliency map via trilinear interpolation and 3D convolutions, combining features from multiple hierarchies. 2) TMFI-Net [Zhou *et al*., 2023] uses transformer-based multi-scale spatiotemporal features. Its semantic-guided encoder enriches features with high-level context, while the hierarchical decoder refines predictions using multi-dimensional attention, and 3) STSANet [Wang *et al*., 2023] uses Spatio-Temporal Self-Attention 3D blocks at multiple levels to extract long-range relations in video.

3.3 Saliency based on tactile paving

Motivated by the idea of focusing more attention on specific objects, we proposed a second approach. In this case, the attention is specifically directed toward urban accessibility, targeting tactile paving. Unlike the previous method, which relied on general human attention, this new one focuses on the recognition of tactile paving features, such as curves and irregularities, which are critical for safe navigation. This reduces subjectivity and leverages the SAM2 [Ravi *et al*., 2024] segmentation model to locate tactile surfaces with minimum human intervention. The next sections provide details on the labeling protocol and pseudo-label generation, as well as on model selection and evaluation metrics.

SAM2

This section provides a brief explanation about the SAM2 [Ravi *et al*., 2024], a transformer architecture with streaming memory designed for real-time video processing. Its zero-shot generalization capability for new data, combined with its high-performance on the Promptable Visual Segmentation (PVS) task, made it suitable for use in this work. PVS allows to prompt the place of objects through mask, points, or boxes. This information is encoded and passed to a mask decoder, which generates the corresponding segmentation masks. In the context of this work, we aim to generate prompts using our click tool along with the tactile input video and perform segmentation based on these prompts.

Labeling protocol and pseudo-label generation

The first stage uses the same tool developed and described in Section 3.2 to collect mouse clicks while watching videos. The tool plays a video and prompts the user to click on a region of the screen every 60 frames. Unlike the previous approach, the clicked region must contain tactile paving. The output is a collection of (x, y) positions to be used as a prompt to SAM2.

In the second stage, the collected clicks from the previous step guide the generation of attention maps that represent tactile surfaces. 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, which, with minimal input, efficiently isolate the tactile paving.

The segmentation masks indicate the location of the tactile paving in the images, as illustrated in Figure 4. In the next stage, we use those masks to generate density maps that highlight specific regions of the tactile paving – we focus on identifying curves and irregularities.

Tactile paving saliency identification

To generate attention points, we propose two approaches. The first approach uses only the segmentation mask of the tactile paving, resulting in attention points at the borders of the tactile area. The second approach uses the skeleton of the mask, which centers the point of analysis on the center of the tactile paving.

For the first, we apply the Hough Line Transform to the mask to detect straight lines. Following this, we identify intersections on the detected lines, which represent points of curvature or irregularities in the border of the mask. 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.

[1] https://scikit-learn.org/stable/modules/clustering.html#dbscan, accessed on March 31, 2025.

A second variation of this approach was to use the skeletons from the tactile mask instead of the mask itself. We then followed the same steps in extracting straight lines with Hough Transform, finding the line intersections, and filtering outliers points. The advantage of this approach is that it centers the density map on the tactile area, reducing emphasis on the borders.

As illustrated in Figure 5, the first row uses the segmentation mask to determine the attention points, it results in a greater variance in the attention points and, consequently, more variance in the final attention maps. The second row represents the second approach, where the skeleton is used instead. This results in the points of interest being more centered within the mask, reducing variance in the attention maps.

3.4 Cropping application

Video cropping trims video frames to highlight key areas. It is used in many applications, including removing unwanted regions [Quang Minh Khiem *et al*., 2010] and adjusting orientation for different screens (video retargeting) [Apostolidis and Mezaris, 2021; Imani and Islam, 2024]. The cropping process is formulated as a linear optimization problem that seeks to maximize the amount of saliency captured within a fixed-size bounding box. Given a salience map for each frame, the algorithm identifies the region with the highest accumulated attention values, ensuring that the cropped area retains the most informative content. The problem can be formulated as follows:

- A saliency map $S \in \mathbb{R}^{H \times W}$, where $S(x, y)$ represents the saliency value at pixel (x, y).
- Desired crop dimensions w (width) and h (height).
- $\mathcal{R}$ as the set of all possible $w \times h$ rectangular regions within the frame, that is $\mathcal{R} = \{[x, y, x + w, y + h] | x, y \geq 0, x + w \leq W, y + h \leq H\}$.

We want to find the optimal crop region $R^* \in \mathcal{R}$ that maximizes the total saliency:

$$R^* = \arg\max_{R \in \mathcal{R}} \sum_{(x,y) \in R} S(x, y) \tag{3}$$

4 Results and discussion

This section presents the performance of models for saliency prediction strategies in human attention and tactile paving. Additionally, we provide visual examples to illustrate the models' effectiveness in saliency prediction and cropping. A video demo of the proposed approach is available in the project Github.

The results are calculated using four metrics to compare the predictions with the Gaussian ground truth: a) Similarity (SIM), which measures the overlap between the predicted and ground truth salience maps; b) Linear Correlation Coefficient (CC), which assesses the linear relationship between the predicted and ground truth maps; c) Kullback-Leibler Divergence (KLDiv), which quantifies the difference between the predicted probability distribution and the ground truth

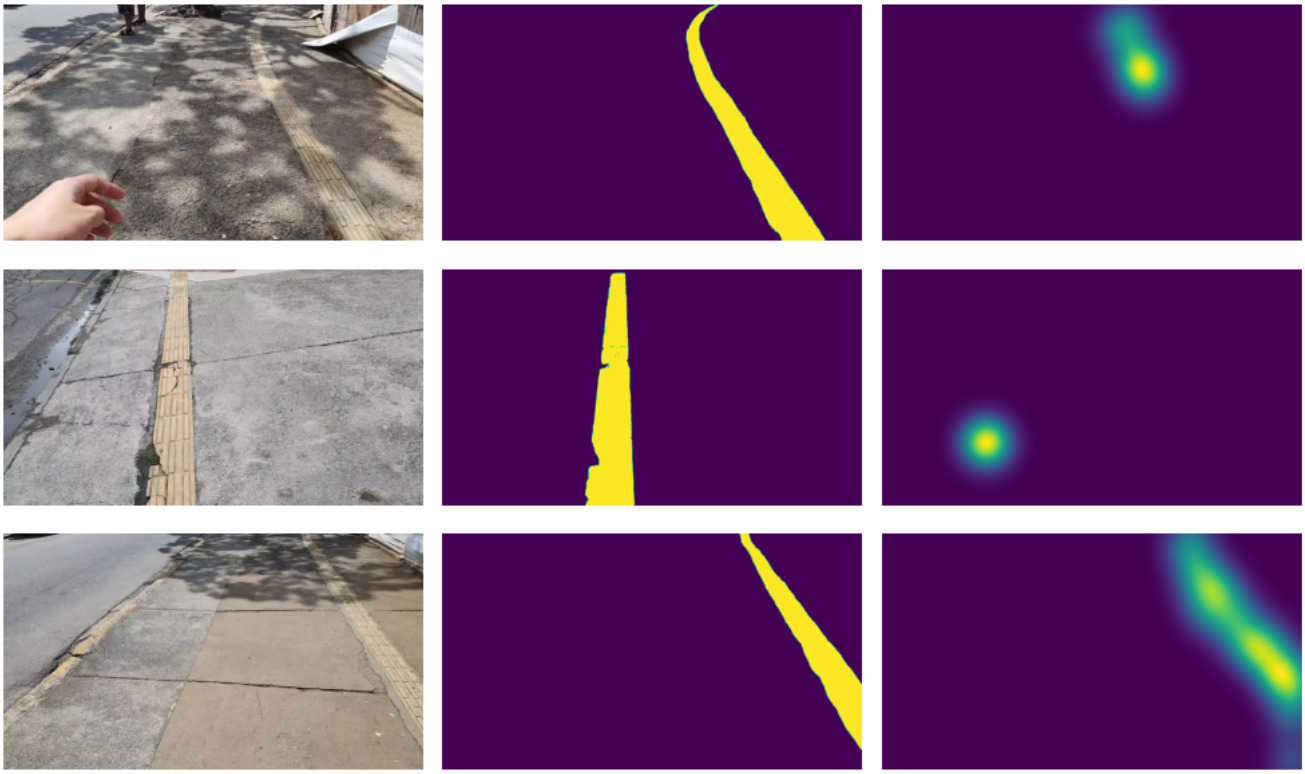


Figure 4. For each row, three images are presented: the original image, the segmented tactile paving mask generated by SAM2, and the corresponding attention map. In the first row, the attention maps highlight curves. In the second and third rows, there are no curves in the scene, but ground irregularities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

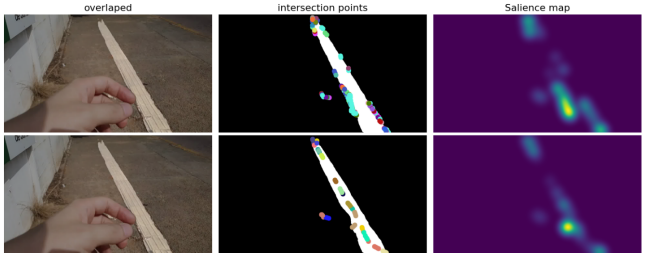


Figure 5. Comparison of two label generation approaches applied to the same frame. The first row uses the segmentation mask to determine the attention points, while the second row uses the skeleton instead.

distribution; and d) AUC, which measures the ability of the model to distinguish between relevant and irrelevant attention regions. These metrics are widely used in salience prediction studies [Zhang *et al*., 2022; Jain *et al*., 2021; Bellitto *et al*., 2021].

4.1 Dataset annotation assessment

We initially evaluate the annotation reliability and inter-annotator agreement. As shown in Table 1, two annotated datasets are considered in this paper, the human attention dataset and the tactile paving dataset. The latter has been annotated by a single participant since it only required a single clicked seed for SAM 2 segmentation. Therefore, we analyze the human attention dataset annotations. In order to evaluate annotator agreement, we evaluate the dispersion of the clicked positions by different participants in each frame. Figure 6 shows the results of this assessment carried out on the annotated dataset. Figure 6(a) shows the clicked positions by different users on some sample frames. In order to quantify the dispersion, the scatter measure [2] of each annotated frame has been defined as

$$S = \frac{1}{N} \sum d_E(x_i, \mu) \tag{4}$$

where x_i denotes the coordinates of the i-th click in a given annotated frame, N is the number of annotated clicks in the frame, μ is the average point of x_i, i.e. the average value of x_i, and d_E is the Euclidean distance. The histogram of S of all frames in the dataset is shown in Figure 6(b). As can be seen, the histogram is concentrated around $S = 160$, thus showing good inter-annotator agreement in our dataset.

4.2 Saliency-based on human attention

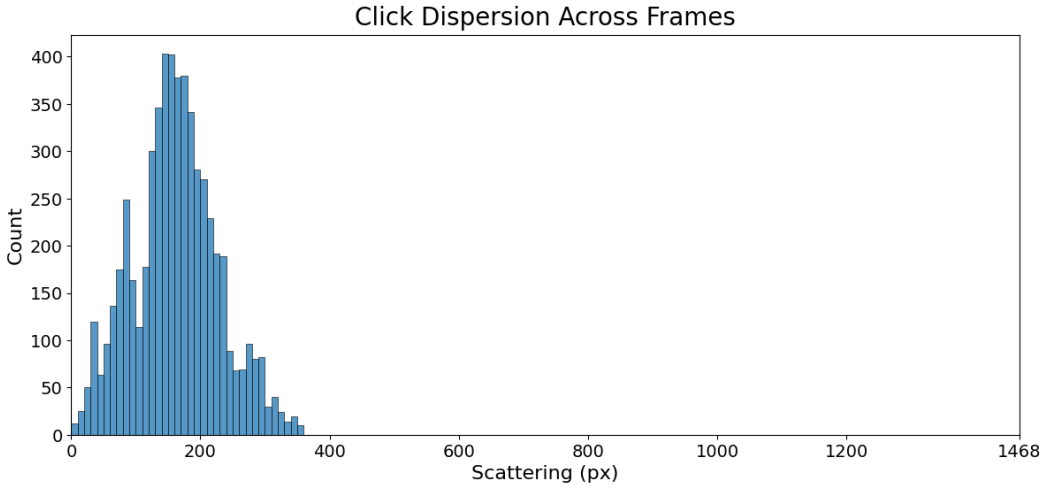
Table 2 presents the quantitative results obtained from testing two model variants: (1) fine-tuned version by initializing the model with weights pre-trained on DHF1K[3] [Wang *et al*., 2018] and (2) model trained from scratch using our dataset.

In general, fine-tuning the models in DHF1K benefit the results, except for STSANet, which suffered a higher drop in performance, with CC falling to 0.27, SIM to 0.26 and KLD increasing to 2.87. TMFI-Net with fine-tuning achieves the best overall performance, obtaining the highest CC (0.584) and SIM (0.532) scores, along with a competitive AUC (0.538). ViNet also improves with fine-tuning, showing gains in CC

[2] Inspired by the intraclass scatter matrix of Equation 8.14 of da Fontoura Costa and Jr. [2018].

[3] DHF1K is a video saliency dataset totaling 1,000 videos with diverse content and lengths, 700 of which are annotated with eye-tracking data from 17 observers.

(a) Dataset annotation click positions



(b) Histogram of annotation dispersion

Figure 6. Results show high coherence among the annotations provided by the different users. It is worth noting that the histogram of S (Equation 4) concentrates in the lower part of possible distances between pixels in the image, whose diameter is ≈ 1400.

Table 2. Results for saliency-based on human attention. In bold, the best results for each training approach, while underline is the best overall.

Model	Pre-trained	CC ↑	SIM ↑	KLD ↓	AUC ↑
ViNet	No	0.37	0.40	1.41	0.36
TMFI-Net	No	0.349	0.395	1.47	**0.582**
STSANet	No	**0.39**	**0.45**	**1.01**	0.44
ViNet	Yes	0.47	0.43	**1.14**	0.51
TMFI-Net	Yes	**0.584**	**0.532**	1.60	**0.538**
STSANet	Yes	0.27	0.26	2.87	0.48

(+0.10) and AUC (+0.15) compared to training from scratch, while maintaining a relatively low KLD (1.14). These results highlight TMFI-Net as the most robust model, benefiting from both training strategies, while ViNet does not stand out among the models and STSANet is more sensitive to fine-tuning.

Figures 7 and 8 provide some qualitative samples with and without fine-tuning, respectively. The first row in these figures present the ground-truth maps followed by the predictions in the subsequent rows. It is possible to see that the salient regions in the ground-truth correspond to important urban structures in sidewalks such as poles and curb ramps. In the case of the results shown in Figure 7, the model predictions also identify salient objects but, since no semantics is included in the model training, such salient regions may correspond to other objects such as pedestrians and motorcycles. This limitation suggests a possible direction for future work. In general, fine-tuning achieves better convergence and alignment to the ground truth, with more precise and concentrated attention maps.

Consider now the results produced by the method without fine-tuning, shown in Figure 8. In contrast with those in Figure 7, results without fine-tuning show more dispersed

attention maps, which explains the lower performance presented for the corresponding models. A possible hypothesis is that the model initialization used for pre-training allows a more focused and selective prediction response.

4.3 Saliency-based on tactile paving

This section presents the models' performance concerning the two segmentation-based strategies discussed in Section 3.3: positioning the attention point at the edge or centering it on the tactile paving. Table 3 summarizes the performance metrics for each configuration. Figures 9 and 10 provide visual representations of the model inferences for each approach.

Table 3. Results for attention maps at the border and the center of the tactile paving. In bold, the best results for each training approach, while underline is the best overall.

Name	Click	CC ↑	SIM ↑	KLD ↓	AUC ↑
TMFI-Net	Border	0.452	0.347	9.45	**<u>0.914</u>**
ViNet	Border	0.280	**0.43**	**4.65**	0.601
STSANet	Border	**<u>0.723</u>**	0.1018	7.9365	0.5343
TMFI-Net	Center	**0.579**	**<u>0.472</u>**	7.582	**0.907**
ViNet	Center	0.125	0.145	9.03	0.684
STSANet	Center	0.392	0.2053	**<u>2.83</u>**	0.73

In the border configuration, STSANet achieves the highest correlation with the ground truth (CC = 0.723), indicating that its predictions closely follow the spatial distribution of salient regions. However, its low SIM score (0.1018) and AUC (0.5343) suggest weak structural similarity and poor discriminative ability in ranking salient versus non-salient areas. The last row of Figure 9 highlights this behavior, where attention maps form a near-circular shape that aligns with the ground truth, but fails to accurately distinguish attention regions.

In contrast, TMFI-Net exhibits the highest AUC (0.914), demonstrating better ability in identifying important regions, though its high KDL (9.45) indicates a significant mismatch in probability distribution. The second row of Figure 9 shows the inferences more focused on the tactile. ViNet performs best in terms of structural similarity (SIM = 0.43) and distributional alignment (lowest KDL = 4.65), suggesting that it generates the most globally consistent attention maps. However, its lower CC (0.280) and moderate AUC (0.601) imply that its attention maps do not correlate as strongly with the actual spatial locations of salient regions; it can be observed in the third row of the Figure 9, where most of the time ViNet is wrong in the place of the attention map.

On the other hand, when the attention point is centered on the tactile, the behavior of the results changes. This configuration reduces the standard deviation of the attention maps, leading them to be more concentrated, as depicted in Figure 10. As expected, the generated attention points cluster around the center, enhancing the overall consistency and effectiveness of the model predictions.

Specifically, when the attention point is at the center of the tactile paving, TMFI-Net continues to demonstrate strong discriminative ability with the highest AUC (0.907) and an improved CC (0.579), suggesting a better correlation with salient regions. Additionally, its SIM score (0.472) is the highest among the models. Also, its KDL value (7.582) has been reduced. STSANet exhibits a lower correlation coefficient (CC = 0.392) but a reduction in divergence (KDL = 2.83), indicating more stable attention patterns. Additionally, its AUC improves to 0.73, suggesting better alignment with real salient regions. This behavior is evident in the last row of Figure 10, where the attention is more concentrated on the tactile and the distribution shows a lower standard deviation. ViNet, on the other hand, performs worse in this configuration, with the lowest CC (0.125) and SIM (0.145).

Overall, the results show TMFI-Net is the most stable method, maintaining relatively high performance across both configurations. ViNet experiences a drop in performance, with no major visual changes except for the attention maps being less spread out. STSANet experiences a trade-off, with lower CC but better KDL and AUC, indicating more stable and concentrated attention patterns. In general, centering attention points refines focus and tends to favor the overall performance across the models.

4.4 Cropping

After applying the process described in Section 3.4 to reframe videos, we conducted a qualitative analysis using both strategies (human-based attention and tactile-based attention) to better showcase the model's capability to detect and crop interesting objects in scenes as well as its limitations. Figures 11 and 12 demonstrate the cropping and reframing operations applied to a video using human-based (Section 3.2) and tactile-based (Section 3.3) saliency maps, respectively. The predicted saliency maps are generated using TMFI-Net, selected as the best trained model by the criteria of best overall metrics and its consistent results across different methods.

The human-based attention method captures a broad range of objects in the predicted saliency maps. This is expected, given that the annotation guidelines did not impose restrictions on which objects participants could select, allowing for a more diverse representation of attention. The first row of Figure 11 shows the predicted saliency map overlaid on the frames. The attention is directed to elements that may interfere with pedestrian walking, such as sidewalk boundaries (columns 2 and 4), poorly positioned posts (columns 1 and 3), other pedestrians on the sidewalk (column 3), or step on the lowered sidewalk (column 4).

When the saliency maps are specialized for tactile paving, the cropping process can effectively focus on regions that are most informative for tactile navigation. The second row of Figure 12 shows the final cropped frames, highlighting regions with issues in the tactile, such as irregular elevations at the borders or intersections (columns 1 and 4), broken tactile surfaces (column 3 and 4) or changes in direction (column 2).

Each saliency strategy offers distinct advantages. The human-based attention method tends to highlight a wide variety of elements—such as pedestrians, vehicles, surface discontinuities, and obstacles—reflecting the natural variability of human perception and judgment. While this results in a broader coverage of potentially relevant areas, it also introduces higher variability in what is considered salient. On the other hand, the tactile-based saliency maps demonstrate a more targeted focus on features for tactile navigation, such as

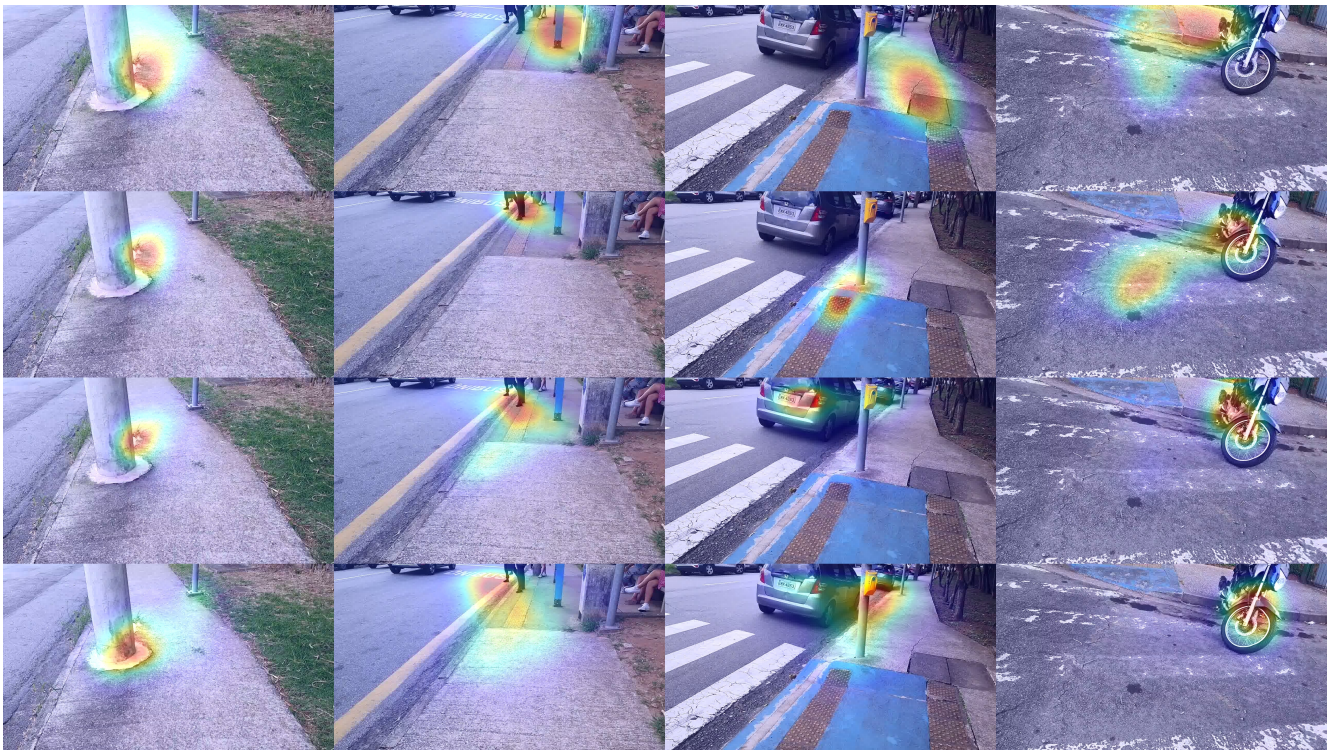


Figure 7. Predictions with the fine-tuned models for human-based attention. The first row is the ground truth, followed by TMFI-Net, ViNet, and STSANet.

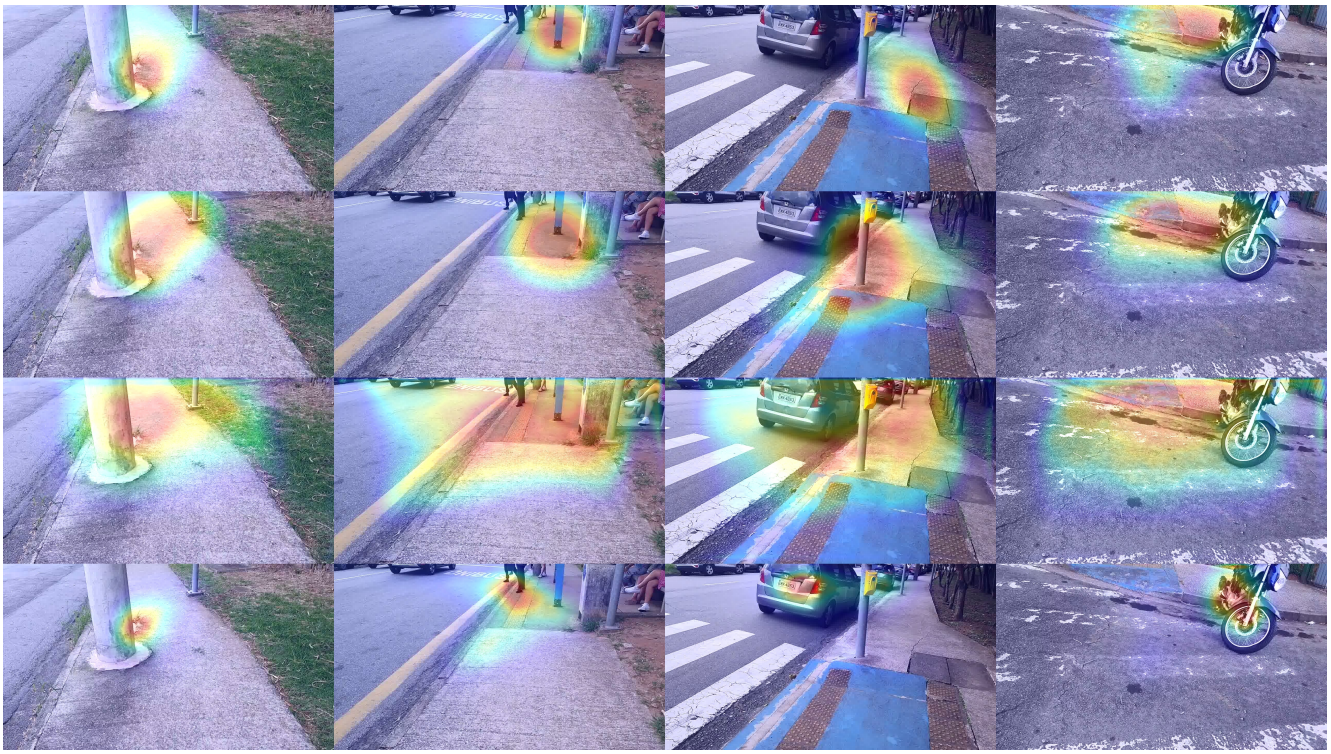


Figure 8. Predictions with the models trained from scratch for human-based attention. The first row is the ground truth, followed by TMFI-Net, ViNet, and STSANet.

continuity and condition of tactile paving. This focus allows for more precise cropping around structural issues. Overall, while human-based approach offers a generalist perspective for scene understanding, the tactile-based approach provides a domain-specialized saliency that is particularly advantageous for accessibility analysis and infrastructure assessment.

In order to quantify the cropping quality, we computed the Intersection over Union (IoU) between the bounding boxes generated using the ground-truth saliency maps and those produced from the predicted saliency maps (Section 3.4, Equation 3). In addition, we adapted the cropping framework proposed by Apostolidis and Mezaris [2021] by integrating our saliency model. The framework provides a more robust pipeline by incorporating a clustering step to filter out noisy saliency responses and select the best region and smoothing methods to do a better temporal transition of the areas. Our method

resulted in an IoU of 0.71 compared to an IoU of 0.72 for the Apostolidis and Mezaris [2021] framework, which represents a marginal improvement when considering the simplicity of our approach.

4.5 Computational performance

In order to assess the computational performance of our approach, we conducted experiments on an NVIDIA A5000 GPU with 24 GB VRAM and CUDA 12.4. The reported metrics represent average inference results for a 15-second video, providing an overview of each model's computational efficiency under identical hardware conditions. Table 4 highlights the performance trade-offs among the evaluated models. TMFI-Net achieves the highest throughput (16.57 FPS) but requires the most resources, with 358.4 GFlops and a peak VRAM usage of 8662.1 MB. ViNet offers a balanced profile—comparable speed (16.83 FPS) with the lowest computational cost (114.1 GFlops) and moderate memory use. STSANet is the most memory-efficient (3349.56 MB peak VRAM) but delivers the slowest frame rate (10.38 FPS).

Given these characteristics, ViNet appears best suited for real-time or edge deployment, balancing speed and efficiency. TMFI-Net may excel in high-performance server environments, while STSANet could be preferable for lightweight applications where memory constraints are critical. However, none are currently optimized for deployment on resource-constrained devices. Further improvements, such as model pruning, quantization, or knowledge distillation, could reduce memory footprint and computational load significantly [Im Choi and Tian, 2023; Ramanathan *et al*., 2020], enabling more efficient real-time applications across diverse hardware platforms.

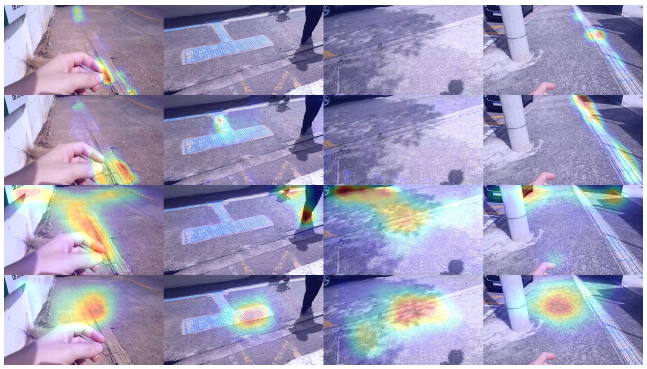


Figure 9. Predictions from models trained with attention focused on the side of the tactile paving. The first row is the ground truth, followed by TMFI-Net, ViNet, and STSANet.

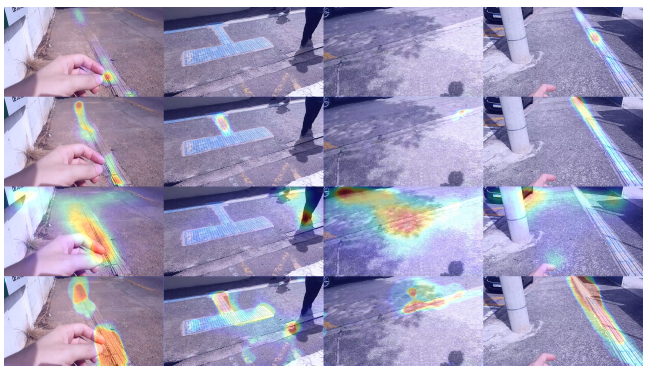


Figure 10. Predictions from models trained with attention maps focused on the center of the tactile paving. The first row is the ground truth, followed by TMFI-Net, ViNet, and STSANet.

Table 4. Performance comparison between ViNet, TMFI-Net, and STSANet for inference time.

Metric	ViNet	TMFI-Net	STSANet
model_load (s)	3.60	4.76	5.828
model_vram (MB)	120.47	240.45	619.53
frames_per_second	16.83	16.567	10.38
model_gpu (GFlops)	114.1	358.4	210.9
vram_peak (MB)	3255.8	8662.1	3349.5
vram_reserved (MB)	4510	8900	3406

5 Conclusion

In this study, we proposed two new strategies for salience prediction in urban context videos, addressing the challenge of identifying and labeling crucial features, such as cracks and surface defects. By leveraging both human visual attention and the SAM2 model, our methods generated rich labeled video data with minimal user interaction, which was then used to train saliency detectors. These detectors, in turn, contributed to the development of video cropping tools capable of improving the analysis of urban environments.

Our evaluation of saliency prediction models for human attention and tactile paving demonstrated that TMFI-Net consistently outperformed ViNet and STSANet in most metrics. For human attention, pre-training improved performance in most cases. In tactile paving detection, centering attention points on the segmentation masks yielded more stable predictions. We also experimented with using the saliency predicted by our models to crop the recorded sidewalk videos and keep only the most relevant segments. The cropping application effectively highlighted critical regions, focusing on relevant obstacles and tactile paving. This information could be used to provide curated information to aid mobility-impaired pedestrians to navigate more safely.

Research directions for future works include (1) refining weak-labeling strategies for tactile paving to expand the range of detectable elements, (2) developing hybrid models that combine human and tactile saliency for more robust environmental understanding—even when specialized in tactile cues, and (3) optimizing the framework for real-time deployment via lightweight and quantized models, enabling practical applications on smartphone devices for accessibility and infrastructure assessment.

Declarations

Authors' Contributions

SC, RJPD, HM, and RMCJ contributed to the conception of this study. SC performed the experiments and is the main contributor of this manuscript. SC, RJPD, HM, and RMCJ are the writers of the manuscript. All authors read and approved the final manuscript.

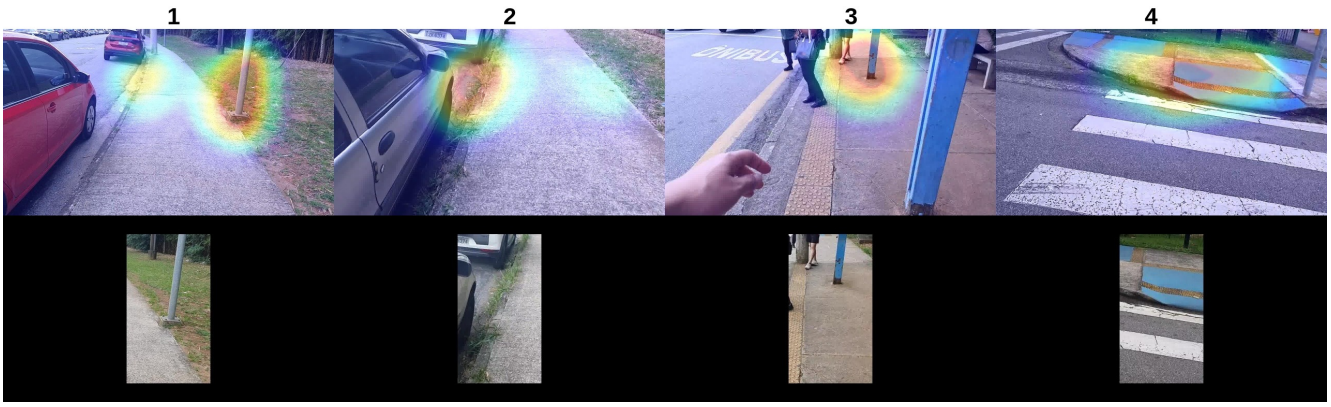


Figure 11. Reframing and cropping for the best-trained model (TMFI-Net) in human-based saliency. The first row presents the original frame with overlaid predicted attention map, while the second is the frame after the crop.

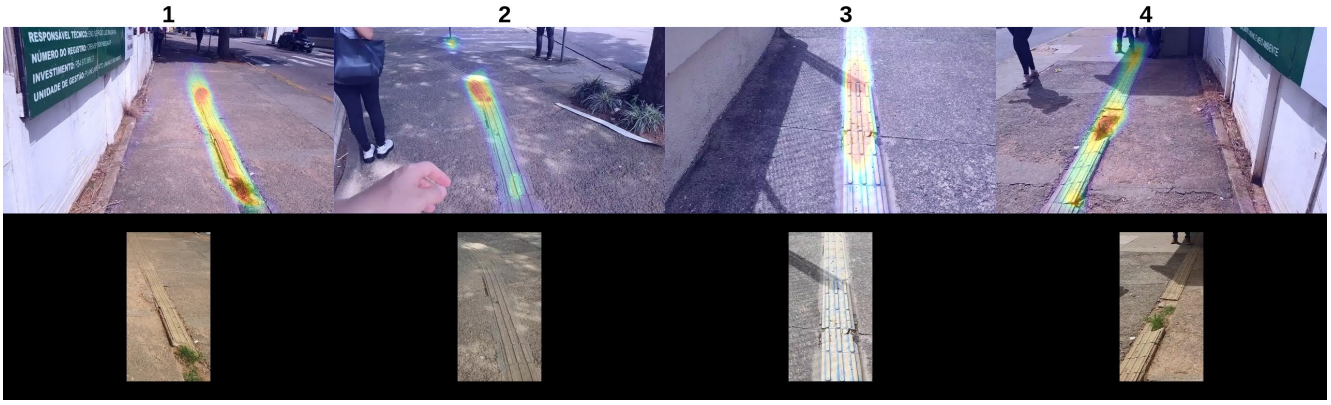


Figure 12. Reframing and cropping for the best-trained model (TMFI-Net) in tactile-based saliency. The first row overlays the predicted attention map on the original image, while the second presents the cropped frame.

Competing interests

The authors declare that they have no competing interests.

Funding

The research was funded by FAPESP (grants #2022/15304-4, #24/21006-1, #25/13940-9, #25/02274-8), CNPq, INCT Sim-AI, CAPES, FINEP and MCTI PPI-SOFTEX (TIC 13 DOU 01245.010222/2022-44). HM is supported by the Beijing Natural Science Foundation (IS23060), the Youth Teacher International Exchange & Growth Program (No. QNXM20250001), and the Fundamental Research Funds for the Central Universities (FRF-TP-22-048A1).

Availability of data and materials

The annotation tool generated during the current study is available in `https://github.com/suayder/VideoClickCapture`, and the experimental code is in `https://github.com/suayder/JBCS-saliency-methods`. The datasets analyzed during the current study will be made available upon request. A video demo of the proposed approach is available in the project GitHub.

References

Abreu, D. R. d. O. M., Novaes, E. S., Oliveira, R. R. d., Mathias, T. A. d. F., and Marcon, S. S. (2018). Internação e mortalidade por quedas em idosos no brasil: análise de tendência. *Ciência & Saúde Coletiva*, 23(4):1131–1141. DOI: 10.1590/1413-81232018234.09962016.

Apostolidis, K. and Mezaris, V. (2021). A fast smart-cropping method and dataset for video retargeting. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2618–2622. DOI: 10.1109/ICIP42928.2021.9506390.

Baba, T. (2021). Vidvip: Dataset for object detection during sidewalk travel. *Journal of Robotics and Mechatronics*, 33(5):1135–1143. DOI: 10.20965/jrm.2021.p1135.

Bellitto, G., Proietto Salanitri, F., Palazzo, S., Rundo, F., Giordano, D., and Spampinato, C. (2021). Hierarchical domain-adapted feature learning for video saliency prediction. *International Journal of Computer Vision*, 129(12):3216–3232. DOI: 10.1007/s11263-021-01519-y.

Bruckert, A., Christie, M., and Le Meur, O. (2023). Where to look at the movies: Analyzing visual attention to understand movie editing. *Behavior Research Methods*, 55(6):2940–2959. DOI: 10.3758/s13428-022-01949-7.

Chen, M. C., Anderson, J. R., and Sohn, M. H. (2001). What can a mouse cursor tell us more? correlation of eye/mouse movements on web browsing. In *CHI'01 extended abstracts on Human factors in computing systems*, pages 281–282. DOI: 10.1145/634067.634234.

Chen, W., Xie, Z., Yuan, P., Wang, R., Chen, H., and Xiao, B. (2023). A mobile intelligent guide system for visually impaired pedestrian. *Journal of Systems and Software*, 195:111546. DOI: 10.1016/j.jss.2022.111546.

Costa, S. M., Damaceno, R. J. P., and Jr., R. M. C. (2024).

Video cropping using salience maps: A case study on a sidewalk dataset. In *Extended Proceedings of the XXXVII Conference on Graphics, Patterns and Images (SIBGRAPI 2024) – Workshop on Works in Progress (WiP)*. Sociedade Brasileira de Computação (SBC). DOI: 10.5753/sibgrapi.est.2024.

da Fontoura Costa, L. and Jr., R. M. C. (2018). *Shape Classification and Analysis: Theory and Practice*. Taylor and Francis. Book.

Damaceno, R., Ferreira, L., Miranda, F., Hosseini, M., and Cesar Jr, R. (2024). Sideseeing: A multimodal dataset and collection of tools for sidewalk assessment. *arXiv preprint arXiv:2407.06464*. DOI: 10.48550/arXiv.2407.06464.

Ghilardi, M. C., Macedo, R. C., and Manssour, I. H. (2016). A new approach for automatic detection of tactile paving surfaces in sidewalks. *Procedia Computer Science*, 80:662–672. International Conference on Computational Science 2016, ICCS 2016, 6-8 June 2016, San Diego, California, USA. DOI: 10.1016/j.procs.2016.05.356.

Gitman, Y., Erofeev, M., Vatolin, D., Andrey, B., and Alexey, F. (2014). Semiautomatic visual-attention modeling and its application to video compression. In *2014 IEEE international conference on image processing (ICIP)*, pages 1105–1109. IEEE. DOI: 10.1109/ICIP.2014.7025220.

Hosseini, A., Kazerouni, A., Akhavan, S., Brudno, M., and Taati, B. (2024). Sum: Saliency unification through mamba for visual attention modeling. *arXiv preprint arXiv:2406.17815*. DOI: 10.48550/arXiv.2406.17815.

Im Choi, J. and Tian, Q. (2023). Visual-saliency-guided channel pruning for deep visual detectors in autonomous driving. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–6. IEEE. DOI: 10.1109/iv55152.2023.10186819.

Imani, H. and Islam, M. B. (2024). Spatio-temporal consistent non-homogeneous extreme video retargeting. In *2024 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–6. DOI: 10.1109/ICCE59016.2024.10444165.

Ito, Y., Premachandra, C., Sumathipala, S., Premachandra, H. W. H., and Sudantha, B. (2021). Tactile paving detection by dynamic thresholding based on HSV space analysis for developing a walking support system. *IEEE Access*, 9:20358–20367. DOI: 10.1109/ACCESS.2021.3055342.

Jain, S., Yarlagadda, P., Jyoti, S., Karthik, S., Subramanian, R., and Gandhi, V. (2021). Vinet: Pushing the limits of visual modality for audio-visual saliency prediction. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3520–3527. DOI: 10.1109/IROS51168.2021.9635989.

Jana, P., Bhaumik, S., and Mohanta, P. P. (2021). Unsupervised action localization crop in video retargeting for 3d convnets. In *TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON)*, pages 670–675. DOI: 10.1109/TENCON54134.2021.9707226.

Jiang, M., Huang, S., Duan, J., and Zhao, Q. (2015). SALICON: Saliency in context. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1072–1080. DOI: 10.1109/CVPR.2015.7298710.

Kim, N. W., Bylinskii, Z., Borkin, M. A., Gajos, K. Z., Oliva, A., Durand, F., and Pfister, H. (2017). Bubbleview: an interface for crowdsourcing image importance maps and tracking visual attention. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(5):1–40. DOI: 10.1145/3131275.

Le, T.-N.-H., Huang, H., Chen, Y.-R., and Lee, T.-Y. (2024). Retargeting video with an end-to-end framework. *IEEE Transactions on Visualization and Computer Graphics*, 30(9):6164–6176. DOI: 10.1109/TVCG.2023.3327825.

Lee, S., Ye, X., Nam, J. W., and Zhang, K. (2022). The association between tree canopy cover over streets and elderly pedestrian falls: A health disparity study in urban areas. *Social Science & Medicine*, 306:115169. DOI: 10.1016/j.socscimed.2022.115169.

Li, M., Lang, X., Gong, R., Zhou, J., Yang, X., and Sang, N. (2024). Tpsegmentdiff: An enhanced diffusion model for tactile paving image segmentation. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops*, MMAsia '24 Workshops. Association for Computing Machinery. DOI: 10.1145/3700410.3702130.

Linardos, P., Mohedano, E., Nieto, J. J., O'Connor, N. E., Giró-i-Nieto, X., and McGuinness, K. (2019). Simple vs complex temporal recurrences for video saliency prediction. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 182. BMVA Press. DOI: 10.48550/arXiv.1907.01869.

Lyudvichenko, V. and Vatolin, D. (2019). Predicting video saliency using crowdsourced mouse-tracking data. *arXiv preprint arXiv:1907.00480*. DOI: 10.30987/graphicon-2019-2-127-130.

Miranda, F., Hosseini, M., Lage, M., Doraiswamy, H., Dove, G., and Silva, C. T. (2020). Urban Mosaic: Visual exploration of streetscapes using large-scale image data. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–15, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3313831.3376399.

Niu, L. and Bao, H. (2024). Fast tactile paving segmentation model based on reparameterized structure. In *Proceedings of the 2024 International Conference on Generative Artificial Intelligence and Information Security*, GAIIS '24, page 24–28, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3665348.3665354.

Ota, K., Kotani, N., Sugikawa, S., and Muraki, Y. (2024). Vertical video cropping considering multiple subjects. In *2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)*, pages 94–95. DOI: 10.1109/GCCE62371.2024.10760449.

Park, K., Oh, Y., Ham, S., Joo, K., Kim, H., Kum, H., and Kweon, I. S. (2020). Sideguide:a large-scale sidewalk dataset for guiding impaired people. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10022–10029. DOI: 10.1109/IROS45743.2020.9340734.

Quang Minh Khiem, N., Ravindra, G., Carlier, A., and Ooi, W. T. (2010). Supporting zoomable video streams with dynamic region-of-interest cropping. In *Proceedings of the first annual ACM SIGMM conference on Multimedia systems*, pages 259–270. DOI: 10.1145/1730836.1730868.

Ramanathan, V., Dwivedi, P., Katabathuni, B., Chakraborty,

A., and Thakur, C. S. (2020). Quicksal: A small and sparse visual saliency model for efficient inference in resource constrained hardware. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1678–1688. DOI: 10.1109/wacv45572.2020.9093354.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., R$^{\prime}$adle, R., Rolland, C., Gustafson, L., *et al*. (2024). Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*. DOI: 10.48550/arXiv.2408.00714.

Saha, M., Saugstad, M., Maddali, H. T., Zeng, A., Holland, R., Bower, S., Dash, A., Chen, S., Li, A., Hara, K., and Froehlich, J. (2019). Project sidewalk: A web-based crowdsourcing tool for collecting sidewalk accessibility data at scale. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3290605.3300292.

Shi, W., Goodchild, M. F., Batty, M., Kwan, M.-P., Zhang, A., *et al*. (2021). *Urban informatics*. Springer. DOI: 10.1007/978-981-15-8983-6.

Takano, T., Nakane, T., Yu, J., and Zhang, C. (2024). Tactile paving detection and tracking using tenji10k dataset. *IEEJ Transactions on Electrical and Electronic Engineering*, 19(10):1661–1672. DOI: 10.1002/tee.24123.

Tang, W., Liu, D.-e., Zhao, X., Chen, Z., and Zhao, C. (2023). A dataset for the recognition of obstacles on blind sidewalk. *Universal Access in the Information Society*, 22(1):69–82. DOI: 10.1007/s10209-021-00837-9.

Tang, Z., Lv, C., and Tang, Y. (2022). Adaptive cropping with interframe relative displacement constraint for video retargeting. *Signal Processing: Image Communication*, 104:116666. DOI: 10.1016/j.image.2022.116666.

Taylor, L. E., Mirdanies, M., and Saputra, R. P. (2016). Optimized object tracking technique using kalman filter. *Journal of Mechatronics, Electrical Power, and Vehicular Technology*, 7(2):57–66. DOI: 10.14203/j.mev.2016.v7.57-66.

Theodosiou, Z., Partaourides, H., Panayi, S., Kitsis, A., and Lanitis, A. (2020). Detection and recognition of barriers in egocentric images for safe urban sidewalks. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics*, pages 530–543. Springer. DOI: 10.1007/978-3-030-94893-1_25.

Wang, W., Shen, J., Guo, F., Cheng, M.-M., and Borji, A. (2018). Revisiting video saliency: A large-scale benchmark and a new model. In *The IEEE Conference on Computer Vision and Pattern Recognition*. DOI: 10.1109/cvpr.2018.00514.

Wang, Z., Liu, Z., Li, G., Wang, Y., Zhang, T., Xu, L., and Wang, J. (2023). Spatio-temporal self-attention network for video saliency prediction. *IEEE Transactions on Multimedia*, 25:1161–1174. DOI: 10.1109/TMM.2021.3139743.

Xia, H., Yao, C., Tan, Y., and Song, S. (2023). A dataset for the visually impaired walk on the road. *Displays*, 79:102486. DOI: 10.1016/j.displa.2023.102486.

Yussif, A.-M., Zayed, T., Taiwo, R., and Fares, A. (2024). Promoting sustainable urban mobility via automated sidewalk defect detection. *Sustainable Development*, 32(5):5861–5881. DOI: 10.1002/sd.2999.

Zhang, K., Shang, Y., Li, S., Liu, S., and Chen, Z. (2022). Salcrop: Spatio-temporal saliency based video cropping. In *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. DOI: 10.1109/VCIP56404.2022.10008849.

Zhang, L., Zhang, J., Lin, Z., Lu, H., and He, Y. (2019). Capsal: Leveraging captioning to boost semantics for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6024–6033. DOI: 10.1109/CVPR.2019.00618.

Zhou, X., Wu, S., Shi, R., Zheng, B., Wang, S., Yin, H., Zhang, J., and Yan, C. (2023). Transformer-based multi-scale feature integration network for video saliency prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7696–7707. DOI: 10.1109/TCSVT.2023.3278410.

Zhu, R., Shi, L., Song, Y., and Cai, Z. (2023). Integrating gaze and mouse via joint cross-attention fusion net for students' activity recognition in e-learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 7(3). DOI: 10.1145/3610876.

Zünd, D. and Bettencourt, L. M. A. (2021). *Street View Imaging for Automated Assessments of Urban Infrastructure and Services*, pages 29–40. Springer Singapore, Singapore. DOI: 10.1007/978-981-15-8983-6_4.