# Video cropping using salience maps: a case study on a sidewalk dataset

Suayder M. Costa*, Rafael J. P. Damaceno*, Roberto M. Cesar Jr.*
*Institute of Mathematics and Statistics
University of São Paulo,
São Paulo, SP, Brazil,
Email: suayder@ime.usp.br, rafael.damaceno@ime.usp.br, rmcesar@usp.br

***Abstract*—Video cropping aims trim video frames to highlight a subject area. This paper introduces a new framework for automated video cropping tailored to sidewalk footage, which is particularly useful in applications such sidewalk navigability and urban planning. By developing a method for video salience annotation using simple mouse input, the introduced framework provides a simple and flexible approach for video cropping. This application is crucial in scenarios where accurately focusing on pedestrian areas is necessary to enhance analysis and decision-making processes. The experimental results obtained from real data in the wild shows that the method is robust to a large variety of sidewalk conditions in different Brazilian cities.**

## I. Introduction

Video cropping is used to trim down video frames to emphasize an area of interest in the footage. This process can be employed, for instance, to remove parts of a video, change the video orientation to accommodate the content on different screens (a process known as video retargeting [1]), focus attention on a specific object in the scene, or used to reduce network bandwidth load and processing time.

This task can be seen as an extension of image cropping, traditionally used to automatically determine the optimal image region (referred to as a window) through methods such as visual quality assessment or salience detection [2]. A common challenge in this task is determining the best window in the image, which can vary subjectively or be guided by object detection, semantic segmentation, and other techniques. When dealing with video content, the temporal component is also important and potentially adds complexity to the task.

A broader concept that supports these tasks is saliency prediction, which involves predicting human gaze fixation when perceiving dynamic scenes. Many cropping methods are based on this concept, initially analyzing each frame individually when it comes to videos. More recent research has enhanced these solutions through the use of 3D convolutions, which, however, introduces greater complexity and computational cost to the models [3].

Despite the existence of many models that perform video cropping, including those based on salience methods, there is a lack of solutions applicable to videos recorded in motion. This scenario hinders image stabilization and the filming present more variable information [4]. Another challenge is incorporating this technique into a smartphone for post-editing videos, given the hardware limitations of these devices.

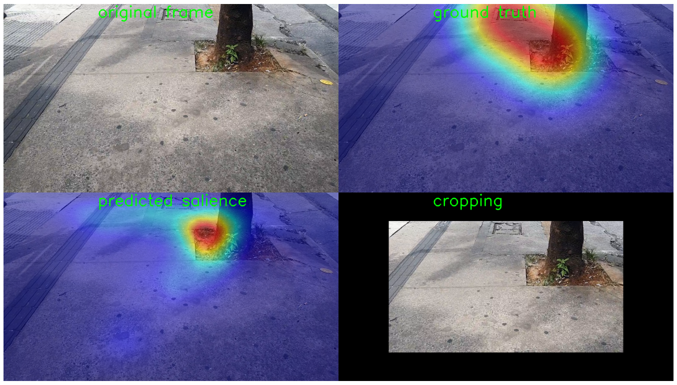


Fig. 1. An example focused on sidewalk pavement cracks using our framework. The top left shows the original frame, the top right displays the salience map resulting from our annotation process. At bottom left presents a salience map resulting from the salience model prediction, and the bottom right depicts a cropping window proposed by the salience prediction model.

This work aims to develop a framework for automated video cropping in the domain of sidewalk footage. Our proposal leverages existing saliency models that have been fine-tuned on a dataset focused on pedestrian paths around hospitals in three Brazilian cities, and can be extended to other countries.

Figure 1 demonstrates an example of what our framework accomplishes in a specific scenario, focusing, for instance, on sidewalk assessment. At the top left is the original frame of a video. At the top right is a saliency map generated through our annotation process, which was then used to fine-tune a salience prediction model. At the bottom left is an example of prediction using a fine-tuned model, and the image on the right represents a cropped version using the model's output to determine the optimal frame window.

The main contributions of this paper are: a) introducing of a new framework for automated video cropping; b) development of a method video salience annotation based on mouse clicks; and c) experimentation with video cropping using real-world data, enabling analysis of sidewalk conditions.

The paper is organized as follows. After this introduction, II presents recent papers on salience models, and video cropping. Section III details the dataset used in this work and the employed pipeline. Section IV outlines the results obtained, both quantitatively and qualitatively. Finally, Section V concludes the paper and discuss potential future work.

## II. Related Works

This section presents studies regarding salience maps and video cropping, which are techniques employed in our work. We explain the rationale behind our decisions regarding the tasks and strategies used to build our study.

Several studies on human visual attention have utilized machine learning techniques for salience prediction. For instance, the introduced models named SalEMA and SalCLSTM, neural networks adapted to incorporate temporal information, demonstrating enhanced performance in handling video salience prediction [5]. Later, ViNet [6] was proposed as architecture for audio-visual and non audio-visual saliency prediction and got success in the task by using 3D fully convolutional archtecture design. These 3D architecture also is studied with hierarchical learning and domain adaptation to the same applications [3].

Another task related to visual human attention, although different, involves movie editing. In this regard, strong correlations between movie editing annotations and spectators gaze distributions is identified [7], which could potentially improve movie edition based on human visual attention. Similar, our study explores human visual attention using distinct approach based on mouse clicks to track and generate visual attention distributions, aimed at fine-tuning a specific dataset related to sidewalk footage.

For video cropping, a well-established it the model named SalCrop which is based on spatio-temporal salience [8]. 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 the video content. The authors also provide a large-scale video cropping dataset composed of 100 training sequences, and 50 validation and testing sequences.

Reframing is a common sub-task of video cropping. Recent frameworks leverages the temporal component of videos [9], [10]. One such solution is based on mechanism that detects jumping frames and smooths their importance, which arguably reduces the jitter of resized videos [9]. Similarly to other works [8], their method also includes an initial stage for scene detection to split the videos into short sequences, followed by a salience detection module.

Differently from these studies, our work is focused on a specific domain and is not limited to aspect ratio transformation, which is the main characteristic of target reframing tasks. Moreover, we leverage salience models to extract information from videos with the goal of analyzing sidewalks, and their pavement conditions as a preliminary case study.

## III. Methods

This section presents the dataset that we used in our proposed work, the pipeline adopted to select and annotate data, the train and evaluation of salience detection models, and a case study on video cropping applications. We also discuss the metrics employed in the evaluation of the models.

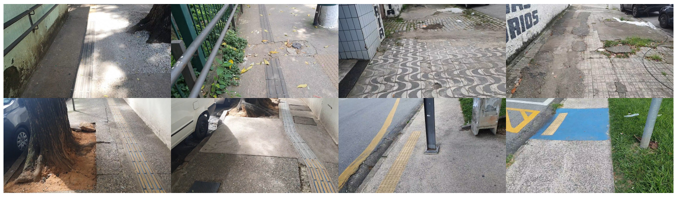


Fig. 2. Eight frames showing the ground of sidewalks near hospitals in three Brazilian cities.

### A. Dataset

The data used in this work is a subset of a dataset generated in our ongoing work, where we developed a new approach for multimodal data acquisition using smartphones [11]. This project is an initiative to facilitate the generation and analysis of multimodal datasets related to sidewalks. It involves collecting various types of data, such as video with audio and sensor data (accelerometer, magnetometer, among others) using smartphones mounted on chest supports worn by individuals. The most recent dataset focuses on videos recorded during walks through hospitals and transportation hubs.

The rationale behind our choice of this dataset relies on the video recordings captured while people are in motion, resulting in footage where frames vary as a reflection of their movement. A video cropping framework has the potential to extract the window of interest from each frame by isolating the subject area, thereby mitigating the effects of motion on the footage.

One of the goals of the ongoing project [11] is conduct studies related to the detection of pavement cracks, potholes, and any other obstacles. Because of this, the smartphone camera was focused on the ground at an angle that potentially favors the view of the pavements. Figure 2 displays two frames each from the videos JUNDIAI-HSV#BLOCK01, SANTOS-CHE#BLOCK01, SAOPAULO-HC#ROUTE02, and SAOPAULO-HUUSP#ROUTE02 to better illustrate the scene characteristics.

The subset used in our study contains seven video files filmed in the three Brazilian cities. As presented in Table I, there are a total of 65,000 frames of data, representing 2,165 seconds of video. The authors recorded the videos at 30 frames per second, with a resolution of 1280 by 720 pixels.

### B. Pipeline

This work adopts a pipeline composed of six stages: dataset selecting, dataset labeling, label processing, AI model experimentation, cropping application, and information analysis (see Figure 3). The following sections describe each of these stages.

*1) Data selecting:* It is important to mention that, due to the nature of human gait, the videos exhibit camera movement, resulting in a lack of stabilization. Therefore, we chose to use this dataset with the idea of generating crop areas that

TABLE I
TOTAL DURATION (S), AND TOTAL NUMBER OF FRAMES OF EACH VIDEO SAMPLE EXTRACTED FROM THE DATASET USED IN THIS WORK.

| ID | Duration (s) | Total frames |
|---:|---:|---:|
| JUNDIAI-HSV#BLOCK01 | 241.94 | 7,259 |
| SANTOS-CHE#BLOCK01 | 330.32 | 9,910 |
| SANTOS-HM#BLOCK01 | 321.43 | 9,644 |
| SAOPAULO-HC#ROUTE01 | 239.77 | 7,194 |
| SAOPAULO-HC#ROUTE02 | 689.88 | 20,680 |
| SAOPAULO-HUUSP#ROUTE01 | 190.71 | 5,722 |
| SAOPAULO-HUUSP#ROUTE02 | 151.08 | 4,533 |
| All | 2,165.14 | 64,942 |

could potentially stabilize the videos. Another goal was to use this content to support automatic video cropping for selecting frame windows focused on obstacles or pavement cracks. We opted to include videos from different cities in this study, depicting various types of pavement materials and potential obstacles.

*2) Data labeling:* We developed an application[1] to annotate the videos by clicking on points of interest while following the video walk-through. Two people annotated all seven video files, totaling 4,462 different annotated frames. We opted to perform one click per second, inspired in other studies related to salience detection [12], [13].

Every click generates a coordinate pair $(x, y)$ referring to the mouse position in the video frame at the time of clicking. This position indicates the location deemed significant by the annotator within the frame. This approach is informed by studies, which explore the relationship between mouse clicks and eye gaze in visual attention research [14]; and other ones proposed the capture of attention map based on clicks, arguing that discrete clicks enable a more explicit record of points of interest [13].

*3) Label processing:* To train a salience model, sparse points alone are not sufficient. We processed these points to generate salience maps as ground truth for every frame in the dataset.

To achieve this, we first applied 1D cubic interpolation to fill each $(x, y)$ point across the video frames. Then, to generate the final salience maps, we employ a Gaussian Mixture Model with a standard deviation of 120, as described in [15]. A sample of the frames and their annotations is shown in Figure 4.

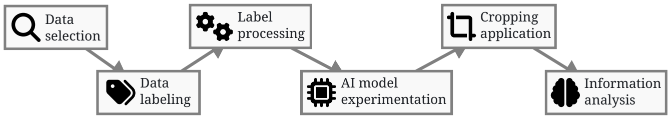


Fig. 3. Pipeline of this study composed of: dataset selecting, dataset labeling, label processing, AI model experimentation, cropping application, and information analysis.

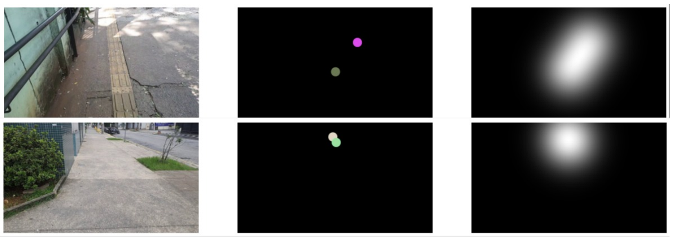


Fig. 4. Two samples of our data and their annotations. The first column shows the original frame, the second column shows the clicked positions by each annotator, and the third column presents the ground truth salience map.

*4) AI Model Exploration:* To automatically extract the salience maps from the video, we adopted a previously established method from the literature, specifically leveraging encoder-decoder convolutional architectures.

ViNet [6] was the chosen network due its simplicity, good reported metrics, and computational efficiency compared to other models available for the same task. ViNet is pre-trained on the DHF1K dataset [16], and has been tested on various public datasets, showing solid results.

Our experiments focused on two primary objectives with our dataset: assessing the performance of ViNet pre-trained with DHF1K and conducting fine-tuning using our annotated data.

*5) Cropping application:* Automatic cropping applications leverage salience maps as information of the most significant parts of the image, thereby enhancing the overall composition and focus of the image. The cropping itself is a linear cropping operation modeled as an optimization problem, where we want to maximize the attention inside a given desired bounding box shape.

Figure 5 shows the cropped frame in the third column, resulting from the described process. For each frame (first column), an attention map (second column) is generated to highlight the most important areas. Subsequently, cropping is applied, followed by resizing the frame to fit the initial dimensions.

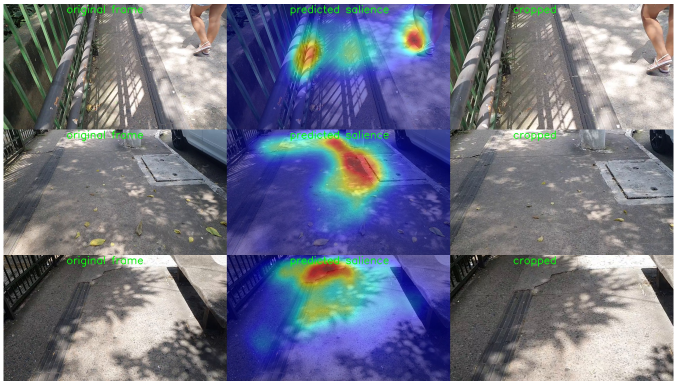


Fig. 5. Illustration of the results from our cropping application. Each row represents a frame from the selected video. The first column shown the original frame, the second column displays the predicted salience map, and the third column presents the final cropped image.

[1] Our video annotation tool is freely available, but it was omitted here for double-blind review purposes.

TABLE II
EVALUATION OF VINET AND FINE-TUNED VINET MODELS ACCORDING TO CC, SIM AND KLDIV METRICS FOR SALIENCE PREDICTION ON THE DATASET USED IN OUR WORK.

| Model | CC | SIM | KLDiv |
|---|---|---|---|
| Original ViNet | 0.39 | 0.42 | 1.05 |
| Fine-tuned ViNet | 0.46 | 0.45 | 1.42 |

### C. Information analysis

The results obtained in our work were analysed from two perspectives: quantitative and qualitative. The quantitative evaluation was conducted using three main metrics to compare the predictions with Gaussian ground truth: Similarity (SIM), which measures the extent to which the predicted and ground truth salience maps overlap; Linear Correlation Coefficient (CC), which assesses the linear relationship between the predicted and ground truth maps; and Kullback-Leibler Divergence (KLDiv), which quantifies the difference between the predicted probability distribution and the ground truth distribution. These metrics are commonly used in the salience prediction studies [3], [6], [8].

In the qualitative analysis, we manually observed the generated salience maps and resulting crop to detect important elements in the scenes. The focus was on identifying objects or pavement defects that can hinder the walkability on sidewalks.

## IV. RESULTS

This section presents the results we obtained regarding the metrics (quantitative results) and the manual analysis (qualitative results) of the generated salience maps and the extracted croppings.

### A. Quantitative results

Table IV-A present the quantitative results obtained from testing both the models pretrained on DHF1K and the version fine-tuned on the dataset used in our work.

These results indicate that the fine-tuned model performs better on the sidewalk dataset used in this work compared to the pretrained model, as expected. Although we opted to use the pretrained model on the DHF1K dataset, we encountered issues with the KLDiv coefficient, which caused predictions to be too dispersed, resulting in unstable cropping. The right side image in Figure 6 exemplifies a prediction with fine-tuned model.

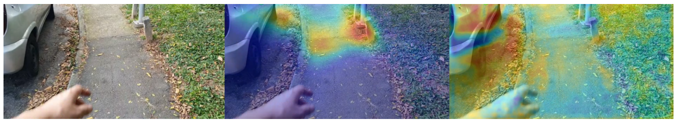


Fig. 6. Predictions with the different models. The left side image is the original frame, the middle is the salience map predicted with ViNet, and the right side image is the salience map predicted by ViNet fine-tuned.

### B. Qualitative results

This section presents a qualitative analysis conducted on five example frames to better showcase the models' capabilities in detecting interesting objects in scenes. The first example is Figure 7a, which highlights the salience map focusing on a crack on the ground. This behavior was consistently observed across other examples, demonstrating the model's capability to detect surface irregularities that could be potential hazards.

In Figure 7b we show a salience map focusing on tactile pavement within the frame. Although the model successfully highlights the tactile pavement, it fails to detect the pavement crack at the bottom. This oversight goes in contrast with the Figure 7a that highlight most the crack. It suggests that the model is proficient at identifying prominent features, as also can be observed in other examples.

In other situations, such as presented in Figure 7c, the model's capability to detect multiple regions of interest simultaneously can be helpful in analyzing complex scenes with various significant features.

Figure 7d highlights a region where potential obstacles are present. This behavior indicates the model's effectiveness in identifying areas that might affect the individual's trajectory.

Finally, Figure 7e illustrates a salience map focusing on a curb ramp and partially on a crosswalk. The model's ability to highlight these essential features underscores its potential usefulness in assisting navigation and enhancing the mobility of individuals.

The observed results suggest that the cropping framework has the potential to serve as a tool for assessing sidewalk conditions. The frames exemplified illustrate scenarios involving one or more elements or objects that can positively or negatively affect walkability, detected by the framework.

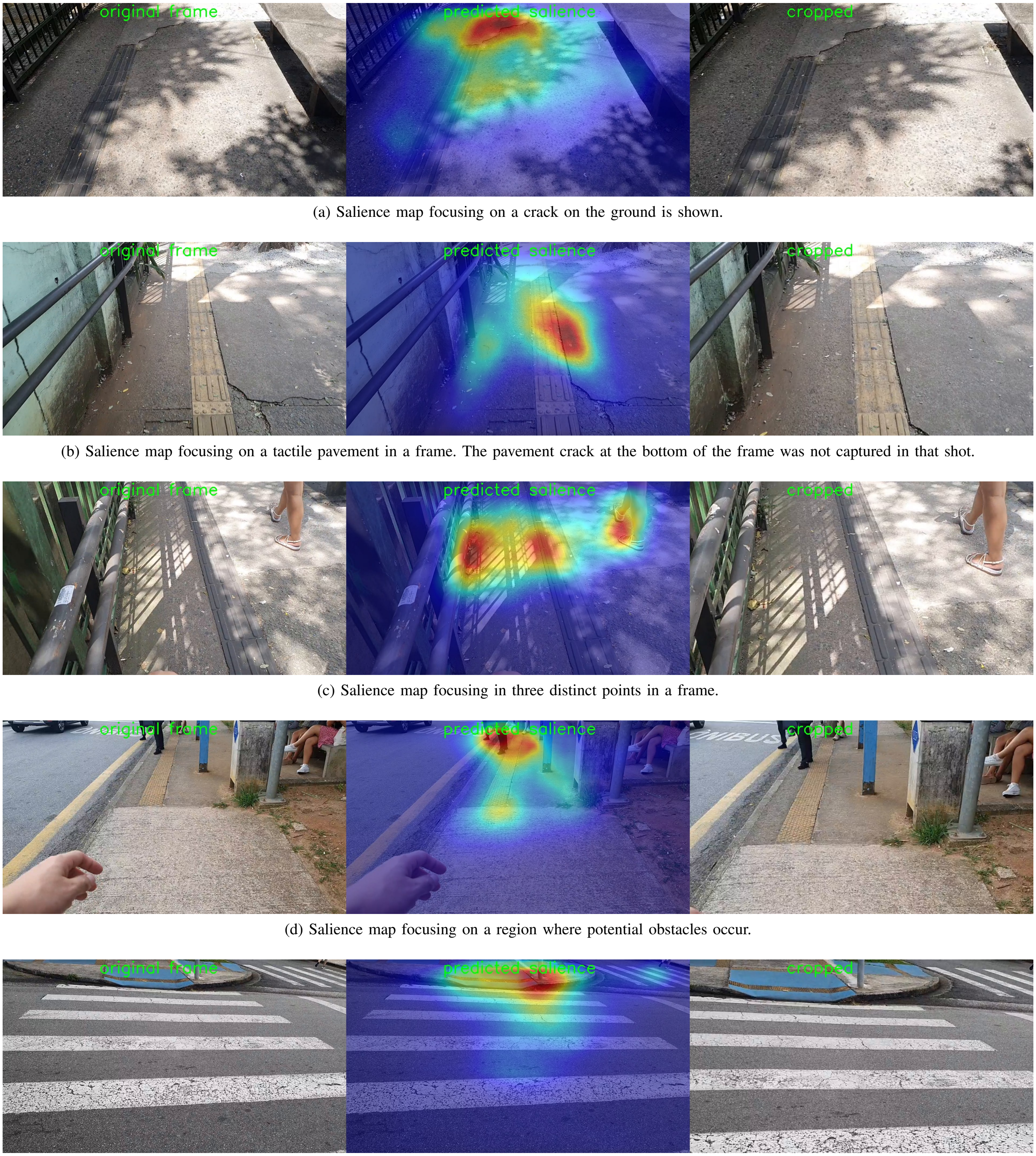
As a cropping application, our solution can condense large videos and high-resolution content into shorter versions with lower resolution, as needed. The cropping framework, along with the video salience annotation tool, can support policy decision-making solutions.

## V. CONCLUSION

In this work, we proposed a framework for automated video cropping in the domain of sidewalk footage. We initially used a well-know salience model, which was later fine-tuned for our specific domain. We evaluate the results of both the pretrained and fine-tuned models with respect to CC, SIM and KLDiv metrics. Moreover, we qualitatively analyzed the video outputs while using the salience model results for video cropping.

Our findings suggest that salience detection is a promising technique to study subjects such sidewalk conditions or walkability. The application of video cropping allow us to focus and direct attention to the most relevant content of interest in the videos, enhancing the effectiveness of visual analysis in sidewalk footage.

However, one limitation of our study is the number of annotators per frame, which can result in significant variation in attention points. As future work, we plan to include more annotators to increase the amount of labeled data. Additionally,



(a) Salience map focusing on a crack on the ground is shown.

(b) Salience map focusing on a tactile pavement in a frame. The pavement crack at the bottom of the frame was not captured in that shot.

(c) Salience map focusing in three distinct points in a frame.

(d) Salience map focusing on a region where potential obstacles occur.

(e) Salience map focusing on a curb ramp and partially on a crosswalk.

Fig. 7. Five example frames to better showcase the models' capabilities in detecting interesting objects in scenes.

we intend to conduct new training procedures, and adapt and deploy the fine-tuned model on smartphones. The goal is to evaluate the models for post-recording video editing, potentially facilitating on-site analysis.

## Acknowledgment

The authors would like to thank FAPESP (grants #2015/22308-2, #2022/12204-9, #2022/15304-4), CNPq, CAPES, FINEP and MCTI PPI-SOFTEX (TIC 13 DOU 01245.010222/2022-44) for their financial support for this research.

## References

[1] K. Apostolidis and V. Mezaris, “A fast smart-cropping method and dataset for video retargeting,” in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 2618–2622.

[2] Y.-L. Chen, T.-W. Huang, K.-H. Chang, Y.-C. Tsai, H.-T. Chen, and B.-Y. Chen, “Quantitative analysis of automatic image cropping algorithms:a dataset and comparative study,” in *WACV 2017*, 2017.

[3] G. Bellitto, F. Proietto Salanitri, S. Palazzo, F. Rundo, D. Giordano, and C. Spampinato, “Hierarchical domain-adapted feature learning for video saliency prediction,” *International Journal of Computer Vision*, vol. 129, no. 12, pp. 3216–3232, Dec 2021. [Online]. Available: https://doi.org/10.1007/s11263-021-01519-y

[4] Y. Wang, Q. Huang, C. Jiang, J. Liu, M. Shang, and Z. Miao, “Video stabilization: A comprehensive survey,” *Neurocomput.*, vol. 516, no. C, p. 205–230, jan 2023. [Online]. Available: https://doi.org/10.1016/j.neucom.2022.10.008

[5] P. Linardos, E. Mohedano, J. J. Nieto, N. E. O’Connor, X. Giró-i-Nieto, and K. McGuinness, “Simple vs complex temporal recurrences for video saliency prediction,” in *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*. BMVA Press, 2019, p. 182. [Online]. Available: https://bmvc2019.org/wp-content/uploads/papers/0952-paper.pdf

[6] 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,” in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 3520–3527.

[7] A. Bruckert, M. Christie, and O. Le Meur, “Where to look at the movies: Analyzing visual attention to understand movie editing,” *Behavior Research Methods*, vol. 55, no. 6, pp. 2940–2959, Sep 2023. [Online]. Available: https://doi.org/10.3758/s13428-022-01949-7

[8] K. Zhang, Y. Shang, S. Li, S. Liu, and Z. Chen, “Salcrop: Spatio-temporal saliency based video cropping,” in *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, 2022, pp. 1–1.

[9] Z. Tang, C. Lv, and Y. Tang, “Adaptive cropping with interframe relative displacement constraint for video retargeting,” *Signal Processing: Image Communication*, vol. 104, p. 116666, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0923596522000236

[10] H. Imani and M. B. Islam, “Spatio-temporal consistent non-homogeneous extreme video retargeting,” in *2024 IEEE International Conference on Consumer Electronics (ICCE)*, 2024, pp. 1–6.

[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] M. Jiang, S. Huang, J. Duan, and Q. Zhao, “Salicon: Saliency in context,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1072–1080.

[13] N. W. Kim, Z. Bylinskii, M. A. Borkin, K. Z. Gajos, A. Oliva, F. Durand, and H. Pfister, “Bubbleview: an interface for crowdsourcing image importance maps and tracking visual attention,” *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 24, no. 5, pp. 1–40, 2017.

[14] M. C. Chen, J. R. Anderson, and M. H. Sohn, “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*, 2001, pp. 281–282.

[15] Y. Gitman, M. Erofeev, D. Vatolin, B. Andrey, and F. Alexey, “Semi-automatic visual-attention modeling and its application to video compression,” in *2014 IEEE international conference on image processing (ICIP)*. IEEE, 2014, pp. 1105–1109.

[16] W. Wang, J. Shen, J. Xie, M. Cheng, H. Ling, and A. Borji, “Revisiting video saliency prediction in the deep learning era,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.