ActiSight: Wearer Foreground Extraction using a Practical RGB-Thermal Wearable

Rawan Alharbi, Sougata Sen, Ada Ng, Nabil Alshurafa, Josiah Hester
Northwestern University
Evanston, IL

***Abstract*—Wearable cameras provide an informative view of wearer activities, context, and interactions. Video obtained from wearable cameras is useful for life-logging, human activity recognition, visual confirmation, and other tasks widely utilized in mobile computing today. Extracting foreground information related to the wearer is the fundamental operation underlying these tasks; separating irrelevant background pixels from pixels of interest. However, current wearer foreground extraction methods that depend on image data alone are slow, energy-*in*efficient, and even inaccurate in some cases, making many tasks—like activity recognition —challenging to implement in the absence of significant computational resources. We built ActiSight to fill this gap; a wearable RGB-Thermal video camera that uses thermal information to make wearer segmentation practical for body-worn video. Using ActiSight, we collected a total of 59 hours of video from 6 participants, capturing a wide variety of activities in a natural setting. We show that wearer foreground extracted with ActiSight achieves a high dice similarity score compared to state-of-the-art, while significantly lowering execution time and energy cost.**

***Index Terms*—Wearable cameras, in wild, thermal**

I. Introduction

Wearable cameras are emerging as an invaluable tool for general understanding and recording of fine-grained human activity, interactions, context, and behavior. While initially put to use in the 1990s as a tool for augmenting human perception, body-worn cameras have since been used in a variety of applications and on a host of different users – for example, an individual consumer who wants to remember social interactions, recall life events, track physical activity and other forms of life logging [1], [2], blind individuals who use cameras to augment perception as tools for navigation [3], [4], a behavioral researcher who wants to understand complex behaviors of people in natural settings [5], and finally, as a tool to aid researchers in assigning activity or interaction labels to data obtained from other wearable sensor data. These labels, in turn, aid in training and validating activity detection models for automated eating detection, which use non-visual signals and machine learning to detect eating episodes [6]–[8]. Many other categories of users and applications exist. It is anticipated that in the future, wearable cameras will become more widespread owing to convenience, usefulness, lowered costs, and increased processing power due to the rapid advances in CMOS imagers and CPUs. This could lead to reams of video data in need of automated processing.

This vision of the future is fraught with practical issues due to the significant computational and energy resources required for performing typical image capturing and processing tasks. The passive nature of wearable cameras results in collecting relevant information (i.e., foreground related to the wearer) and irrelevant information (i.e., background containing other people and objects). Separating *wearer pixels*, pixels related to the wearer's head and active hand, from background pixels allows us to solve critical challenges facing wearable cameras, such as enhancing privacy by obfuscating the background [9] and reducing the cost of both manual [10]–[12] and automated processing [13]–[15] by focusing on processing of wearer information. The active hand is critical for enabling further interaction research. Therefore, extracting wearer pixels from the frame while discarding the rest of the background is the most critical and fundamental step in the processing pipeline of wearable cameras.

Deep learning RGB-based segmentation models have shown great potential in extracting wearer pixels from wearable camera data. However, such models are resource and time-intensive, especially when deployed real-time on-device, making applying them at the start of the pipeline impractical. To speed up data processing, other sensing modalities can be added to the camera and a lighter processing pipeline can initially process the data and used intensive approaches only when needed. Our key observation is that the wearer activity, like hand and head movements, can be extracted simply and speedily by augmenting a wearable camera with a low-resolution thermal imager directed at the wearer, significantly reducing the energy and time, and human effort to extract wearer pixels in wearable camera data.

In this paper, we present ActiSight (shown in Fig. 1), a practical, all-day battery-lifetime wearable camera platform (hardware and software) that uses thermal imaging as a complementary data stream to extract pixels related to the wearer in a frame, reducing the captured information and simplifying processing tasks. This hardware platform is coupled with an energy-efficient pipeline for speeding up wearer extraction tasks (i.e., pixels related to the wearer's head and hands), a fundamental task of modern and future wearable video capture.

ActiSight captures medium resolution RGB images and ultra-low resolution (8×8) thermal images (Fig. 1B). When the wearers hands are in the field of view, the thermal sensor provides utility by guiding the capture of wearer pixels in the RGB image. However, given the complexity of human behavior, the certainty of the thermal imager in providing

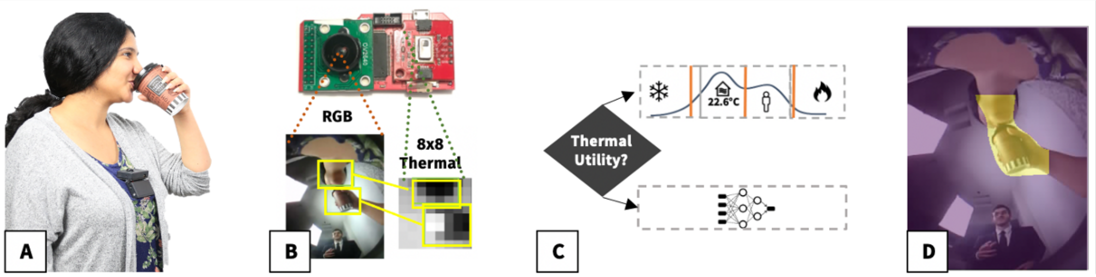
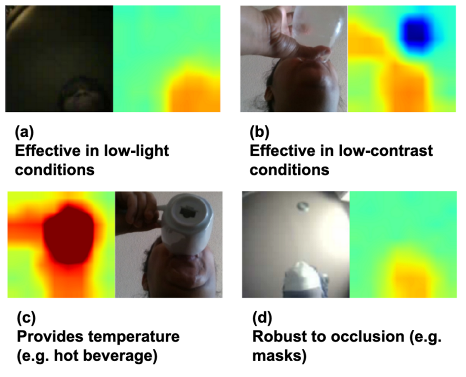


Fig. 1: (A) ActiSight with camera facing upwards towards the wearer (B) ActiSight uses a dual-imaging sensing stream: *(right)* very low resolution (8x8 pixel) and low power thermal infrared (to extract thermographic images) and *(left)* visible light/RGB (to extract photograph images similar to the one found in existing wearable cameras). ActiSight utilizes information from both cameras to determine pixels which contain the wearer (foreground). (C) A utility function with the thermal image as input decides between a low energy pipeline and a RGB-based NN segmentation network to extract foreground (D)

utility is not always guaranteed. To increase speed of the wearer processing pipeline, ActiSight takes advantage of the heat signature provided by the thermal imager, to extract wearer pixels when the thermal signature is deemed to be certain in providing utility for the task (Fig. 1C and D). In cases where the thermal sensor does not provide utility for wearer segmentation, ActiSight will default to using an RGB-based wearer extraction approach trained using neural networks trained using an efficient MobileNet network as the backbone (Fig. 1C - RGB-based approach). The wearer extraction pipeline is tested in multiple use cases ranging from hand detection and pose estimation to background obfuscation to filter non-wearer data. ActiSight provides a starting point for numerous exciting wearable camera-based applications. Our work lays the foundation for future research to more accurately identify the wearer and objects the wearer is interacting an important task, a critical task within human interaction research.



(a)
Effective in low-light conditions

(b)
Effective in low-contrast conditions

(c)
Provides temperature (e.g. hot beverage)

(d)
Robust to occlusion (e.g. masks)

Fig. 2: Thermal data is a useful complement to RGB data because unlike light/color sensor used the cameras, thermal sensors can capture images even in (a) low illumination (e.g., the dark) and despite (b) low color contrast contrast (e.g., liquid in a glass), (c) provide extra information that can not be determined by RGB images (e.g., drinking something hot), and are (d) robust to occlusion (e.g., thermal detects the human regardless of the mask)

II. ActiSight Approach: Using Thermal Signatures to Extract Foreground

One of the primary motivations for using thermal as a second modality in ActiSight is that it provides information that augments and complements the RGB camera's data. Methods that exclusively use RGB-only cameras capture information that can be obtained from the visible light spectrum. This means that RGB images are greatly affected by scene illumination, occlusion, and object positioning. In contrast, thermal images are not affected by scene illumination, meaning that a thermal image obtained in the dark will look the same as a thermal image captured in a well-lit environment. Fig. 2 shows several other scenarios where thermal imaging can overcome shortcomings of RGB imaging, making thermal wearable cameras a useful complement to RGB wearable cameras.

In this section we will fist explain details about thermal infrared sensing, as it is often confused with night vision. We then compare the thermal approach with other approaches in terms of foreground information extracted and current draw.

A. Background on Thermal Infrared

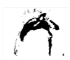
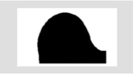
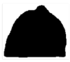
In this section, we will summarize the essential background on thermal infrared; for more detailed background, we refer the readers to Gade *et al.* 's survey paper [16]. All objects with a temperature above zero Kelvin (e.g., humans) emit thermal radiation. Thermal infrared sensing is not to be confused with infrared night vision, though they both rely on infrared radiation. Night vision systems capture near-IR (0.7–1.4 μm) that is not naturally emitted by objects but instead artificially projected towards a scene by the system. Thermal sensing, on the other hand, does not require external IR projection or illumination because it captures IR that is naturally emitted by objects in the sensors field of view (capturing mid- and long-wavelength IR spectrum, 3–14 μm). Humans emit a constant IR radiation (around 8μm), which primarily motivates the exploration of thermal sensing as a second modality in wearable cameras. One concern that might arise with thermal imaging is its power consumption. We next demonstrate how our thermal approach is not only capable of capturing images in various scenarios, but is also energy-efficient and provides reasonable power to information trade-off compared with existing approaches to extract the foreground.

B. Sensor Options and Exploratory Hardware Study

While the addition of thermal images provide useful complementary information to RGB images, it is important to consider the impact of adding a second sensing modality in the wearable.

As a proof of concept, we collected time-synchronized thermal (Grid-EYE), RGB (RealSense), and depth images (RealSense). Then we extracted foreground using multiple approaches, including: *thermal* segmentation using thresholds based on human temperature, *depth* segmentation using distance threshold (i.e., the distance of the wearer's head to the camera, human skin *color* segmentation, and *semantic* segmentation using DeepLabV3 [17]. Groundtruth foreground was obtained from five images manually by labeling each pixel that belonged to the wearer. We then compared each foreground obtained from each segmentation approach to the groundtruth foreground segment using the dice similarity score (i.e., pixel-wise f-score) [18]. The higher the score, the higher the similarity, and therefore the higher foreground information obtained. In TABLE I, we report the average current draw and the foreground similarity score retrieved from 1 minute of data, and we show a sample image. We measured the current draw required to process the data on a laptop with an Intel Core i9 processor and the current drawn by the system used to collect the data. This figure shows that an 8×8 thermal IR sensor array (Grid-EYE) achieves a good balance in the current draw to similarity score trade-off, thus striking a good balance between the system's energy requirements and foreground information extraction. This provided motivation for our exploration of using thermal sensing in ActiSight due to the aforementioned advantages, that make it an ideal second modality to an RGB wearable camera.

TABLE I: Thermal strikes a good balance in current draw to information trade-off. We report average foreground information retained (dice similarity score) and current draw

	Foreground Extraction Approach			
	Color	**thermal**	**Depth**	**Semantic**
Current Draw (A)	0.26	0.23	0.46	0.66
Similarity Score	0.42	0.78	0.87	1
Sample Image				
				

III. ActiSight Hardware

We developed the ActiSight prototype that comprises a custom printed circuit board centered around an STM32L496 ARM Cortex-M4 microcontroller (MCU), enclosed in a 3D printed case. This initial prototype serves as a reference design for ActiSight, and enables evaluation of the usefulness, wearability, and efficacy of the ActiSight platform. The full implementation details, schematics, code, etc., enabling construction of a ActiSight can be found at *https://github.com/anonymized.*

To enable all-day wear, we designed and implemented a variety of techniques, both at the software stack level and in the hardware design. These techniques allowed ActiSight to achieve low power operation and extended its battery lifetime. Chiefly, we have developed a prototype (shown in Fig. 1) that integrates the *minimum* number of hardware components necessary to accomplish video recording and foreground detection. To ensure all-day wear, we optimize for power by taking the following steps: (1) our prototype does not include any user interface components, or extraneous sensing like audio. This is in contrast to consumer devices like GoPro, which has 1 to 3 hour battery lifetime, but which sport a full color LCD display and allow user interactions; (2) we reduce power draw by reducing the video resolution to 320×240 in the default setting, as HD video is not often needed for recognizing a wide set of everyday human activities; and (3) to reduce the load on the SD card, which is the highest power consuming component next to the camera sensor, frames are compressed and batched in the microcontroller's internal memory space, and then stored to the micro SD card once that buffer is filled.

A. Dual-image Sensing Stream

While most wearable cameras rely on a single sensor stream, the RGB, we leverage two sensing streams — RGB and Thermal IR. These two streams allow us to extract foreground activities.

1) ***RGB Image Stream****:* We used the Omnivision OV2640 camera Arducam OV2640, which outputs JPEG images over a DCMI interface. We configure the camera to operate at a resolution of 320×240, which we empirically found is sufficient to recognize our fine-grained activities. The video is captured at five frames per second (fps), a comfortable frame rate for humans to review the video without introducing significant jitter. Also, we believe five fps is a good trade-off point to save battery while also being able to capture fine-grained activities (e.g., the highest frequency activity is chewing, and its frequency is not greater than 2.5Hz [19]). We attached a 180° fish-eye lens to the camera to increase the field of view (FoV). A wide-angle (fish-eye) lens allows a broader field of view to allow for capturing activities from an egocentric position.

2) ***Thermal IR Stream for Foreground Sensing****:* We used a Panasonic GridEYE Infrared Array - AMG8833 [20](GridEYE), which comprises an 8×8 infrared thermopile array. Each of the 64 pixels can provide independent temperature readings. The Grid-EYE resembles a very low-powered, low-resolution thermal camera; which is a useful approximation to the high powered stereo depth cameras which easily interpret foreground, as well as detect human movement.The FoV of the Grid-EYE is 60°, which is narrow. Unlike the RGB camera, the Grid-EYE's FoV can not be modified. Therefore, we directed the Grid-EYE toward the wearer's face in order to capture the wearer's hands and head movements. This allows us to capture a wide range of hand-head activities that are of interest to the research community. Activities such as eating, drinking, smoking, and coughing, are examples of such hand and head related activities, which we also collected along with their confounding gestures.

B. Configurable Components, Attachment and Encapsulation

Since ActiSight contains a MicroSD card slot on board, it allows for flexible and expandable data storage. To enhance privacy, ActiSight encrypts all data on the fly using a stream cipher (salsa20 [21]). Although in our prototype we use a 180° fish-eye lens, ActiSight allows usage of any lens that is compatible with an M12 mount. We experimented different mounting approaches including (1) a lanyard to secure the device around the neck, and (2) a magnetic back-plate that detaches and is placed between the wearer's shirt and the body, and magnets mounted on the back of the camera to secure the device on the shirt itself. We found that using both a lanyard and magnetic plate allows the camera to be stably placed on the body without adding discomfort to the wearer. It is also important to note that since we are attaching the camera on the clothing and around the neck, the camera will be affected by large movements such as the wearer bending down. However, this displacement is momentary and the camera will go back to the desired position after the wearer returns to a regular upright body posture. The PCB is enclosed in a different compartment than the battery which distributes the weight across the device and hinges allow for camera angle adjustment, depending on location and wearer body type.

C. Device power consumption

To show that our ActiSight prototype has all-day battery lifetime, we performed a measurement campaign of the energy-efficiency of the prototype when using a 1200mAh LiPo battery. We estimate the battery life based on the capacity of the battery measured in milliampere-hour (mAh). Ampere is an electrical unit used to measure the current flow towards the load. The battery life or capacity can be calculated from the input current rating of the battery and the load current of our prototype device, which we measured. We verified the data captured by examining the frame rate of the collected video, the size of the images, and comparing the timestamps of the video captured on the SD card. One person wore it continuously until the battery fully discharged. The total number of frames collected in the single use was 224,225 with an effective frame rate of 5.19 frames-per-second, and a total memory footprint of 1.675 GB. These results confirm that our prototype device provides all-day (12 hours being sufficient for most applications) battery lifetime in-wild when using a 1200mAh LiPo battery.

D. Calibration

Similar to any multi-modal or dual-imaging device, the sensing streams have to be calibrated once. In ActiSight the two sensing streams (RGB and thermal) have different FoV and resolutions, making the calibration process challenging. We used the traditional method for automated camera calibration using a checkerboard in our first calibration attempt. We built a custom checkerboard that can be detected by both thermal and RGB [16]. However, given the low-resolution (8×8) of the thermal camera, the checkerboard pattern is not captured by the thermal sensor (it appears as one big blob due to thermal crossover), making automatic calibration challenging. Therefore, we opted for a manual calibration approach to move and scale the thermal image until it lines up with the RGB frame. We create an interface that will help scale and move the thermal image over the RGB frame with 1-pixel resolution to obtain the transformation parameters (scale and registration position). We perform this manual calibration process on randomly sub-sampled pairs of thermal and RGB images obtained from ActiSight, and then we confirm the calibration by visualizing the output on different frames. This calibration process is only performed once per device.

IV. ActiSight Wearer Extraction Pipeline

To capitalize on the fact that thermal and RGB images provide complementary information, we designed and implemented a processing pipeline that can extract foreground pixels from both RGB and thermal images. However, to ensure practicality and efficiency, we use the thermal-based segmentation approach as our default approach, and only use the RGB-based approach when the thermal approach fails. As illustrated in Fig. 3, the pipeline checks if the thermal frame provides utility for foreground extraction (we provide details about performing the utility test in Section IV-A). If the thermal frame passes the utility test, i.e., the thermal frame is deemed useful (as depicted in Example 1 in Fig. 3), we perform thermal-based foreground segmentation using the approach described in Section IV-B. If, however, the thermal fails the utility test (as depicted in Example 2 in Fig. 3), then we perform the RGB-based segmentation – details provided in Section IV-C.

A. Thermal Utility Test

In some cases, thermal sensors fail to capture the human in the foreground (e.g., when the sensor is under direct sunlight or when there is minimal temperature contrast between objects and the environment). These failure cases can be identified by checking the pixel value distribution (i.e., whether the temperature value distribution is uniform) and the range of the pixel values (i.e., whether the range of value lies between 0 to 80 for the thermal sensor that we are using). Therefore, a thermal utility test is helpful in instructing ActiSight on the possibility of determining the presence of a human in the foreground, based on the captured thermal image. We test thermal utility by checking the variability of the pixels (i.e., standard deviation >1) and the range of the pixels (i.e., min >=0 and max =<80) of the frame. In Fig. 3, the thermal frame in Example 1 provides utility as it passes the thermal utility test (std=4.5, min=17.5, max=34.25) and visually it can be seen that the frame captures the wearer's head, hand and the object in hand. However, the thermal frame in Example 2 does not pass the utility test (std=109.5, min=-511.25, max=19.75) and as one can see visually, the wearer's head is not captured in this frame.

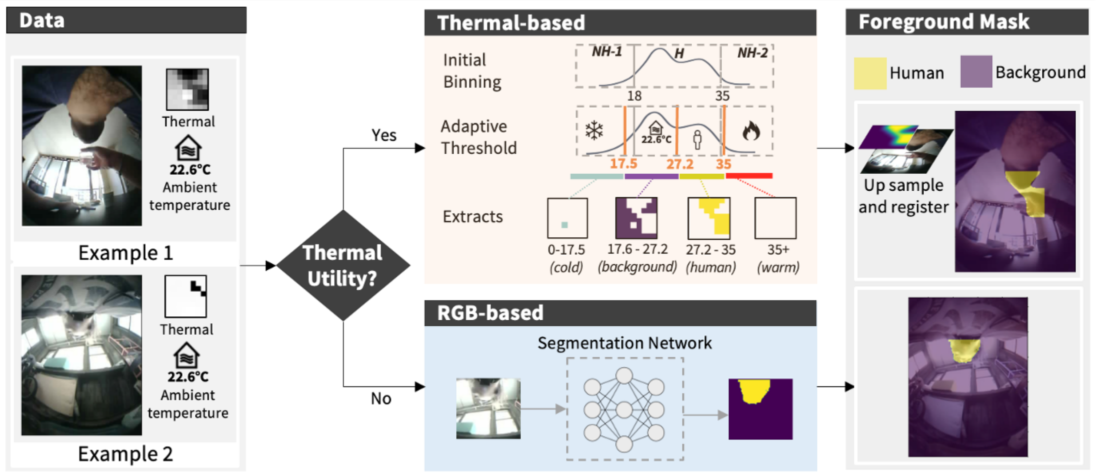


Fig. 3: ActiSight Wearer Extraction Pipeline: The thermal-based segmentation approach is the default approach for image segmentation. However, when thermal fails to provide enough utility for segmentation due to sensing limitation, we use an RGB-based segmentation network to extract wearer pixels.

B. Thermal-based Foreground Extraction

We extract the foreground using the thermal-based approach only if the thermal frame provides utility for segmentation, as described in Section IV-A. In order to extract humans and objects in the foreground pixels, we segment the thermal frames by applying a hierarchy of adaptive thresholds for each 64 pixel thermal frame $t_j \in T$. The adaptive threshold pipeline comprises the following steps: Initial Binning, Adaptive Threshold, Human Pixels Extraction, and Foreground Mask Creation.

*1) **Initial Binning**:* First, we assign the i^{th} pixel $t_j(i)$ to either a bin signifying human pixels (H) or to one of two bins, NH_1 for cold objects and NH_2 for warmer objects, based on the pixel intensity. Empirically, based on data from 13 participants, we identified that $t_j(i) \in [18 \,..\, 35]$ encompasses the range of human body temperature. Fig. 3 illustrates the initial binning output, creating three regions: NH_1, H, and NH_2. Pixels with an intensity value below 18 °C (cooler than the human body temperature range) are binned into NH_1, while pixels higher than 35 °C (warmer than the human body temperature range) are binned into NH_2.

*2) **Adaptive Threshold**:* To distinguish between background and wearer foreground, we apply Otsu's adaptive thresholding method [22], to each of the three regions (NH_1, H, and NH_2), which separates background from foreground pixels by maximizing inter-class variance. This step creates four bins that hold pixels signifying cold, background, human, and warm objects. Otsu's adaptive threshold identified a split in the bin signifying human-related pixels at a temperature of 27.2 °C. After applying Otsu's adaptive thresholding, to determine which bin signifies the background temperature bin (from 18 to 27.2 °C, or 27.2 to 35 °C), we estimate the temperature of the ambient environment (as a representation of the temperature of the background pixels). We estimate ambient temperature by calculating a moving average (across 10 frames, empirically set) of the median temperature value of each frame. For example, the estimate of the ambient temperature in the j^{th} frame shown in Fig. 3 is 22.6 °C, the median temperature value across pixels in the entire frame, calculated by averaging across 10 frames. This estimate works under the assumption that the majority of the pixels represent the background, and while this is not always true for every frame, when averaging across 10 frames, in most scenarios (where the camera is not occluded), we are able to obtain a reasonable enough estimate of the background to determine which temperature bin range belongs to the background and which to the human or object in the foreground. Because 22.6 °C in Fig. 3 falls in the 18 to 27.2 °C bin, we assign that bin to represent the temperature of the background objects in the RGB image.

*3) **Human Pixels Extraction**:* Once we identify the background pixel range across the 3 bins, we conjoin the bins along the Otsu threshold boundary to define a fourth bin that represents the temperature ranges for the background pixels. Fig. 3 shows how the Otsu threshold for the bin representing cold objects was set at 17.5 °C and 22.6 °C for the bin representing the human, and as a result the cold bin range is adjusted to 0 - 17.5 °C, and the background bin range is defined to be between 17.5 and 27.2 °C, and the human bin range remains from 27.2 to 35 °C. The warm bin in this case was empty as there are no warm objects in the FoV. If the participant in the picture was drinking a hot drink instead of cold one, we would have seen more pixels in this range. Finally, we extract a binary mask for the human, the object (warm/hot or cold pixels near the human hand), and background.

*4) **Foreground Mask Creation**:* After extracting the human and object pixels from the thermal frame, we overlay it with the corresponding RGB frame to identify the corresponding pixels that signify the human and object. We map the 8×8 Thermal sensor array frame to the 320×240 RGB frame using a series of linear transformation functions. We first up-sample the IR sensor array using a Gaussian kernel. This up-sampling process resizes the IR frame to 110×110 (the FoV of the thermal is smaller than the RGB). Then, we register or map the IR frame onto a fixed location in the RGB frame obtained from the calibration process explained in Section III-D. In our case, we did not observe the need to remove the lens distortion in the RGB image as the distortion was more severe at the edges of the RGB frame, rather than at the center, where the thermal and RGB FoV overlap.

C. RGB-based Foreground Extraction

When the thermal frame does not provide utility for segmentation (i.e., does not pass the utility test described in Section IV-A), we use the RGB frame to extract foreground (see Example 2 in Fig. 3). Specifically, we use Feature Pyramid Network (FPN) segmentation network [23] with a MobileNetV2 [24] backbone (pre-trained on ImageNet [25]) and fine-tuned using our dataset to predict the foreground. It is possible to replace FPN with other segmentation models such as U-net [26], DeepLabv3+ [17], and DDRNet-23-slim [27]. However, we chose FPN because it performed the best on our data set. Similarly, one can use other backbone networks; we chose MobileNet as the backbone because it runs efficiently on resource constrained devices. We modified the output layer to produce two labels – background and wearer. We train it on our dataset using the following parameters: Dice loss function [28], Adam optimization method [29], 0.0001 learning rate and 40 epochs. To obtain a generalizable segmentation model, we used a leave-one-participant-out (LOPO) evaluation method (i.e., training on all participants except the test participant). At the prediction time, we crop the region in which RGB overlaps with thermal, creating an image size of 110×110, which we feed to the segmentation network. It is possible to feed the whole RGB frame to the segmentation network; however, we assess our pipeline on the overlapped region to investigate the complementary aspect of thermal and RGB modality in foreground extraction.

V. Data Collection

Using ActiSight in natural settings allowed us to capture a variety of natural activities and contexts. We recruited six participants using Craigslist and Research Match to wear ActiSight while performing their everyday activities. All participants were instructed on how to wear and operate ActiSight. Participants were requested to wear the camera and collect at least 8 hours of data, while they performed their everyday activities. We instructed participants to remove the camera under any circumstance where they or any one around them were uncomfortable with video recording. At the end of the study, participants were asked to review their footage and delete any segments they did not want to share. None of the participants deleted any segments. This study was approved by the University's Institutional Review Board and participants were compensated for their time. TABLE II summarizes the participant demographics. In total, 59 hours of footage (over 1 million frames) was captured using the ActiSight system, of which 39 hours was used in the analysis. Footage without useful information (e.g., when the camera was worn inside a jacket) was excluded from evaluation.

TABLE II: In-wild participants' demographic data and data collection details

P	Gender	Race	Age	Hours collected
P1	F	White	40	10
P2	M	White	58	9
P3	F	Black	62	12
P4	F	Black	25	9
P5	M	White	39	9
P6	F	Black	33	10
Total				59 hours

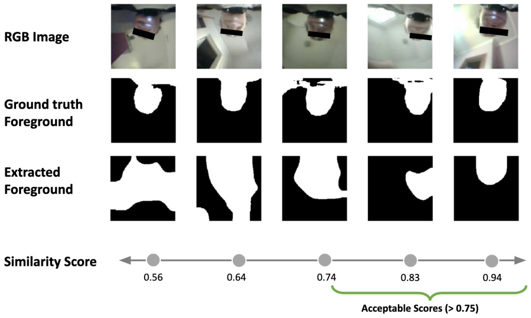


Fig. 4: The Dice similarity score for various images. The score is higher when the two images are similar (highest = 1).

VI. Evaluation Method

A. Foreground Groundtruth Segmentation

In order to assess our approach, we need to obtain the groundtruth of foreground segmentation. Manually labeling wearer pixels is infeasible in our case as we have more than 1 million frames in total. Therefore we used an existing deep learning semantic segmentation network to extract pixel-level image segmentation from the RGB frames to determine which pixels signify a human or "people" in the frame, producing a soft labeled dataset. To convert the soft labels to groundtruth, we performed a visual inspection of each segmented frame manually and excluded frames that had poor quality groundtruth segmentation (i.e., the majority of the wearer pixels were not detected). If part of the wearer pixels were missing, we corrected the groundtruth segmentation by combining fragmented segments and removing pixels that do not involve the human wearer (i.e., bystanders). We experimented with multiple human segmentation models and finally used DeepLabV3+ [17] with Xception network [30], pre-trained on the COCO dataset [31] as it most accurately predicted human pixels in our dataset (DeepLabV3+ achieves an mIoU of 87.8% on COCO dataset). We ran the model on our dataset, which contained more than 1 million frames in total. This extraction process took 3 days using a machine with four GeForce RTX 2080 Ti GPUs. Finally, by manually labeling the 59 hours of data, we extracted the 120K frames, which contained accurate groundtruth (i.e., we confirmed the correctness of the soft groundtruth by manually reviewing each frame, turning it to an actual groundtruth). Two people performed this labeling process over three weeks requiring approximately 144 hours of labor. One person acted as the main reviewer and the second person confirmed the review of the first person. No one had access to the thermal data while labeling to avoid bias.

B. Evaluation metrics

We used two separate quantitative metrics to evaluate ActiSight extracted foreground against the groundtruth foreground: (1) a pixel-level comparison using Dice Similarity Coefficient Score [32], and (2) resource consumption (i.e., processing time in frames per second and power consumption in Watts). We also report (3) qualitative results on the whole dataset to identify cases where our thermal performed better than our RGB approach and vice versa.

*1) **Pixel-level comparison**:* We calculated the Dice Similarity Score for each extracted foreground for every frame in our dataset. The Dice similarity score, also known as the f1-score, is used to evaluate similarities between our detected foreground and the groundtruth foreground. The Dice similarity score value ranges from 0 (low similarity) to 1 (high similarity). Fig. 4 shows a sample of extracted foregrounds and their corresponding similarity score.

*2) **Resource consumption**:* Efficiency is essential when it comes to the real-time processing of data collected continuously and passively (such as that from wearable cameras) because these data capturing methods often result in large quantities of data that needed to be processed. Thus, low processing time means we can process more number of frames per second and extract meaningful information such as foreground with low latency (time delays). In order to compare the efficiency of our approach, we first randomly selected a subset of images from our dataset (1095 images). We then calculate the average frame rate per second (fps) when we run ActiSight foreground extraction pipeline and when we run the groundtruth extraction method. For control purposes, we performed this test on the same computer with an Intel Core i9 processor (available on most laptops). All code is written and provided to the community using Python 3.7.

*3) **Qualitative Performance Analysis**:* Although pixel comparison metrics using Dice similarity provide us with an understanding of how our segmentation compared against groundtruth, the understanding is limited as we obtained groundtruth for a subset of the images. Moreover, the groundtruth was extracted from RGB, which makes the evaluation biased to RGB (for example, it is hard to obtain groundtruth segments from RGB images captured in low illumination). Therefore, we analyze the output of the thermal- and RGB-based foreground extraction by first running each foreground extraction method on the whole dataset (all frames). We then qualitatively investigate the cases in which thermal-based segmentation performed better than RGB-based segmentation. Since we do not have groundtruth for all of the frames, we instead overlap the foreground produced from the thermal- and the RGB-based approach and identify the cases where the overlap was low (IoU<0.3). We then go over the frames that have a low IoU to assess the reason for this discrepancy in the foreground extraction (i.e., lack of illumination for RGB, direct sunlight for thermal, etc.). This analysis will highlight cases from the real-word, showcasing the importance of both RGB and thermal sensing in extracting foreground without biasing the groundtruth to any sensing modality.

TABLE III: Foreground extraction results using ActiSight Thermal-RGB approach

	Thermal-RGB				
	F1	**Pos**		**Neg**	
		P	R	P	R
P1	0.93	0.77	0.91	0.98	0.94
P2	0.86	0.81	0.81	0.90	0.89
P3	0.81	0.78	0.91	0.86	0.70
P4	0.80	0.75	0.88	0.88	0.75
P5	0.88	0.83	0.93	0.93	0.82
P6	0.94	0.93	0.86	0.95	0.97
Average	0.87	0.81	0.88	0.92	0.84

VII. Results

We next describe ActiSight's performance in extracting foreground. We also present ActiSight's performance evaluation, and a qualitative performance analysis.

A. Foreground Extraction Results

Evaluating ActiSight pipeline using in-wild data allows us to understand its performance in uncontrolled natural settings. To show the effectiveness of ActiSight in utilizing both the thermal- and RGB-based foreground extraction approaches, we first present the foreground extraction results independently (i.e., RGB-Only and Thermal-Only). Then we show the results after using the thermal utility function and our Thermal-RGB approach.

Fig. 6 shows the median similarity score of foreground extraction per hour for each participant using ActiSight Thermal-only (Fig. 6a) and RGB-only (Fig. 6b). While Fig. 6c highlights the differences in results achieved from (Fig. 6a) and (Fig. 6b), showing that for certain hours, we can see that one of the modalities performs better than the other in certain hours.

Fig. 7 shows the results of ActiSight when both thermal and RGB-based approaches are used together, showing the improvement that ActiSight combined approach achieves over the RGB-alone and the thermal-alone approaches. The thermal

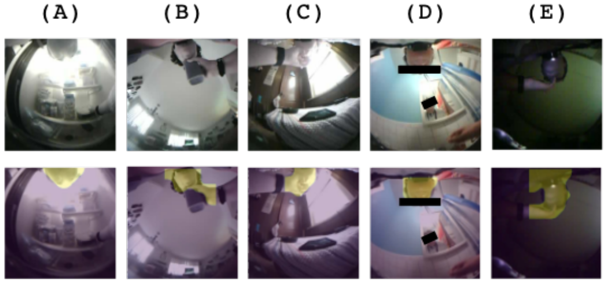
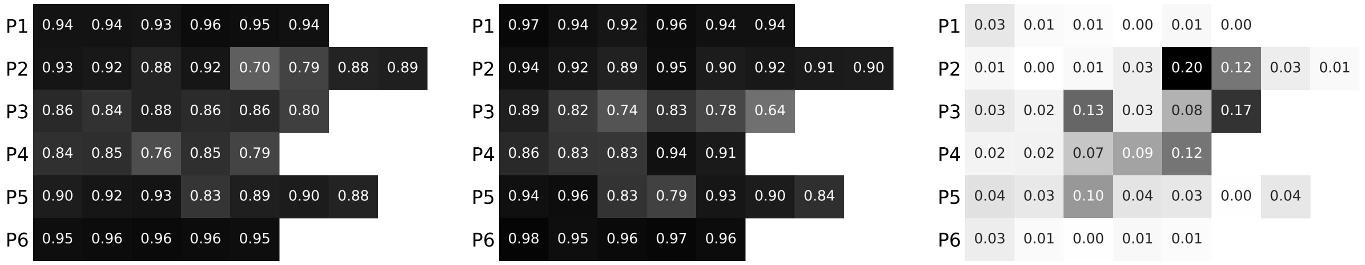


Fig. 5: Sample of foregrounds obtained from in-wild data. (A) searching for food in the fridge, (B) drinking cold beverage, (C) smoking, (D) brushing teeth, (E) eating in front of a computer.



(a) Thermal-Only

(b) RGB-Only

(c) Difference between (a) and (b)

Fig. 6: Median similarity score per hour for each participant using ActiSight (a) thermal-based approach and (b) RGB-based approach. (c) shows the difference between results obtained from (a) and (b), highlighting the complementary nature of thermal and RGB.

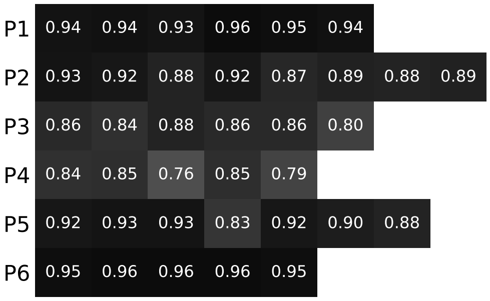


Fig. 7: ActiSight Thermal-RGB showing improvement over Thermal- and RGB-only approaches.

utility test enables ActiSight to decide whether to perform ActiSight RGB-based approach or ActiSight Thermal-based approach. In our dataset, 4% of the frames were processed using the RGB-based approach as the thermal did not provide enough utility, while 96% of the frames were processed using our thermal-based approach. TABLE III shows the overall results of ActiSight that attains a mean weighted F1-score of 0.87. Fig. 5 presents some sample foreground extraction cases under a variety of contexts.

B. Foreground extraction processing latency and power consumption

ActiSight's Thermal-based foreground extraction approach processes images on the CPU at 111 fps. This translates to requiring 32.4 minutes to extract the foreground from 12 hours of data. ActiSight's RGB-based approach on the other hand processes data at 21 fps on the CPU, requiring 156 minutes to process the same amount of data, demonstrating that RGB-based approach is 5× slower than the Thermal-based approach. Calculating the processing time for ActiSight's combined Thermal-RGB approach for foreground extraction will depend on the context. So we calculate the best and the worst-case processing time based on our dataset. The best case is when the thermal frame always provides utility, resulting in a processing time similar to the Thermal-only approach. The worst-case scenario in our dataset was for P2, where the thermal utility was low, 12% of the time. In the worst case the data will be processed at 100.2 (21 × 0.12 + 111 × 0.88) fps on CPU, requiring 47.23 (156 × 0.12 + 32.4 × 0.88) minutes to process the data on the CPU. Although the Thermal-RGB approach increases the processing time by 15 minutes, it also increases the performance of the foreground extraction, as shown in Section VII-A. In terms of power consumption, power consumption when running ActiSight on the CPU is 0.8 Watts.

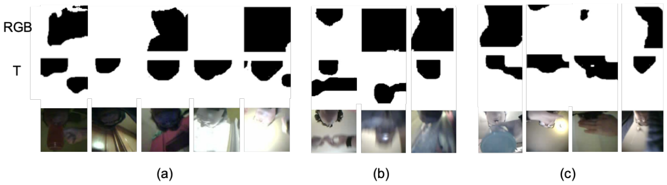


Fig. 8: Cases where the thermal-RGB (T) approach outperforms the RGB-only approach (RGB). (a) shows images captured under different illumination settings, (b) shows thermal not being affected by motion blur, and (c) shows that the thermal data is robust to occlusion caused by objects (e.g., cups, clothing) or body parts (i.e., hand covering the face).

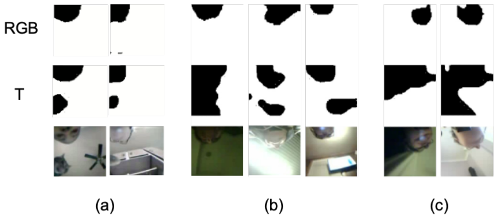


Fig. 9: Cases where the thermal-RGB (T) approach performs worse than the RGB-only approach (RGB). In some cases, the wrong object is foreground: (a) a cat and a spoon with warm food and (b) TV and heat emitting lamps. We have also observed failure cases when (c) the participant moves between rooms with different temperatures.

C. Foreground extraction qualitative performance analysis

We next discuss the qualitative results of the ActiSight Thermal-RGB approach and the RGB-only approach on the whole dataset. Fig 8 presents an example of cases where the Thermal-RGB approach outperforms the RGB-only approach, while Fig 9 shows examples in which the RGB approach outperforms the Thermal-only approach.

1) Successful cases: Fig 8(a) shows images captured under different illumination settings, showing that our Thermal-RGB approach is capable of extracting foreground regardless of the environment illumination condition. This is an improvement from the RGB-only approach, which often fails when the illumination is too low or too high. This happens because the contrast between the face and background during such conditions is low, making it hard to differentiate and extract the foreground in these images.

Thermal modality has several other advantages. For example, it is not affected by motion blur, as seen in Fig. 8(b). This is because thermal is not dependant on the visible light spectrum. Motion blur can affect the appearance of the image, making it hard to detect the wearer using RGB frames since some of the features are not apparent.

Fig. 8(c) shows that the thermal data is robust to occlusion caused by objects (e.g., cups, clothing) or body parts (i.e., hand covering the face). On the other hand, the RGB-based approach sometimes confuses objects to be part of the face or misses detecting the face when it is occluded. Although this can be improved by training the model with more data, since thermal relies on the temperature, it provides a more efficient object detection approach.

2) Failure cases: Fig. 9 shows cases when the Thermal-based or Thermal-RGB approach identifies the wrong object as the foreground. For example, in Fig. 9(a), a cat and a spoon with warm food were detected as humans. While in Fig. 9(b), TV and heat emitting lamps confounded the foreground extraction method. This confusion is due to the low resolution of the thermal sensor and the observations that these objects emit temperature close to the human body temperature range. This confusion can be mitigated by using a small classification network to classify the RGB regions of the foreground to detect if it belongs to humans.

Fig. 9(c) presents a scenario where the participant moved from one location to another location where the temperature was different from the initial location. The ActiSight approach takes the median temperature to identify the background pixels by referencing historical values (previous ten frames). This may introduce latency when the environment changes, causing an error in determining the background segment for a few frames. In the future, we can shorten this historical window length or use an external ambient temperature sensor and calibrate it with the thermal sensor to determine the temperature value of the background.

VIII. Related Work

When using wearable cameras to passively collect data in the wild, extracting relevant wearer foreground information is vital. Foreground extraction methods for data captured from wearable cameras aims to extract parts of the image relevant to the task at hand (i.e., the active hand interacting with objects). These extracted segments then undergo some computationally intensive processing, and therefore, more accurate foreground extraction can limit unnecessary and costly processing. The most common foreground extraction methods used in wearable camera research are based on one or more of the following: frame selection, region selection, or pixel selection.

Instead of extracting a rectangular region around the object of interest, pixel-level segmentation methods, the most fine-grained technique, aim to extract pixels related to the foreground. Foreground pixel extraction (i.e., pixel segmentation) is considered the most challenging task among the previously mentioned ones as the classification occurs at a per-pixel level. However, it is one of the most informative ones as it gives more details that can be used to infer human activity [14], the object in hand [33], and gesture recognition [34], [35]. CNN based semantic segmentation models (e.g., DeepLabv3+ [17], U-Net [26] and several others [36]) can extract human foreground using RGB images. While such approaches work well in certain cases, they fail when faced with images that are not present in the training set distributions, requiring the need for further training and fine-tuning. Researchers are investigating a multi-modal or dual-sensing modality segmentation approach to overcome the limitations of RGB-only (i.e., single modality) image segmentation techniques. For example, researchers have utilized depth information obtained from RGB-D cameras to extract hand segments [35], [37]. *ActiSight foreground extraction is based on pixels, however, unlike previous work, we utilize an efficient thermal-RGB approach to extract foreground. In particular, we primarily rely on the thermal modality to obtain foreground pixels and use the RGB modality for foreground extraction only when the certainty of the thermal imager is low in providing detection of the wearer pixels.*

IX. Conclusion

This paper introduces ActiSight, a practical wearable camera that enables energy-efficient and fast extraction of foreground pixels related to the wearer. In our approach, we augment the wearable camera's data with a second sensing stream, thermal, that aids in foreground extraction. Since there is no practical thermal-RGB wearable camera available, we built one that allowed for further validation of the capability of a low-resolution thermal camera for wearer extraction. Moreover, we developed a foreground extraction pipeline that utilizes thermal information to extract the foreground related to the wearer. Using ActiSight, we collected in-wild data and compared the foreground segmentation obtained from ActiSight with groundtruth achieving an acceptable dice similarity score of 0.82 for the in-wild data. This result is promising, especially when we consider the low-energy required to extract the foreground. By providing ActiSight to the community, we hope to enable efficient processing of wearer foreground extraction, which is an important step in applications such as human activity, gesture recognition, and hand pose estimation.

References

[1] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, and K. Wood, "Sensecam: A retrospective memory aid," in *International Conference on Ubiquitous Computing*. Springer, 2006, pp. 177–193.

[2] T. Bipat, M. W. Bos, R. Vaish, and A. Monroy-Hernández, “Analyzing the use of camera glasses in the wild,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–8. [Online]. Available: https://doi.org/10.1145/3290605.3300651

[3] K. Lee, D. Sato, S. Asakawa, H. Kacorri, and C. Asakawa, “Pedestrian detection with wearable cameras for the blind: A two-way perspective,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–12. [Online]. Available: https://doi.org/10.1145/3313831.3376398

[4] S. Rodger, D. Jackson, J. Vines, J. McLaughlin, and P. Wright, “Journeycam: Exploring experiences of accessibility and mobility among powered wheelchair users through video and data,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019, p. 630.

[5] G. O’Loughlin, S. J. Cullen, A. McGoldrick, S. O’Connor, R. Blain, S. O’Malley, and G. D. Warrington, “Using a wearable camera to increase the accuracy of dietary analysis,” *American journal of preventive medicine*, vol. 44, no. 3, pp. 297–301, 2013.

[6] A. Bedri, R. Li, M. Haynes, R. P. Kosaraju, I. Grover, T. Prioleau, M. Y. Beh, M. Goel, T. Starner, and G. Abowd, “Earbit: using wearable sensors to detect eating episodes in unconstrained environments,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, p. 37, 2017.

[7] S. Bi, T. Wang, N. Tobias, J. Nordrum, S. Wang, G. Halvorsen, S. Sen, R. Peterson, K. Odame, K. Caine *et al.*, “Auracle: Detecting eating episodes with an ear-mounted sensor,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, p. 92, 2018.

[8] Y. Dong, J. Scisco, M. Wilson, E. Muth, and A. Hoover, “Detecting periods of eating during free-living by tracking wrist motion,” *IEEE journal of biomedical and health informatics*, vol. 18, no. 4, pp. 1253–1260, 2014.

[9] R. Alharbi, M. Tolba, L. C. Petito, J. Hester, and N. Alshurafa, “To mask or not to mask?: Balancing privacy with visual confirmation utility in activity-oriented wearable cameras,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, p. 72, 2019.

[10] J. Chen, S. J. Marshall, L. Wang, S. Godbole, A. Legge, A. Doherty, P. Kelly, M. Oliver, R. Patterson, C. Foster *et al.*, “Using the sensecam as an objective tool for evaluating eating patterns,” in *Proceedings of the 4th International SenseCam & Pervasive Imaging Conference*. ACM, 2013, pp. 34–41.

[11] K. H. Ng, V. Shipp, R. Mortier, S. Benford, M. Flintham, and T. Rodden, “Understanding food consumption lifecycles using wearable cameras,” *Personal and Ubiquitous Computing*, vol. 19, no. 7, pp. 1183–1195, 2015.

[12] S. Pizza, B. Brown, D. McMillan, and A. Lampinen, “Smartwatch in vivo,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 5456–5469.

[13] M. Ma, H. Fan, and K. M. Kitani, “Going deeper into first-person activity recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1894–1903.

[14] S. Bambach, S. Lee, D. J. Crandall, and C. Yu, “Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1949–1957.

[15] T.-H.-C. Nguyen, J.-C. Nebel, F. Florez-Revuelta *et al.*, “Recognition of activities of daily living with egocentric vision: A review,” *Sensors*, vol. 16, no. 1, p. 72, 2016.

[16] R. Gade and T. B. Moeslund, “Thermal cameras and applications: a survey,” *Machine vision and applications*, vol. 25, no. 1, pp. 245–262, 2014.

[17] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[18] L. R. Dice, “Measures of the amount of ecologic association between species,” *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[19] J. Po, J. Kieser, L. M. Gallo, A. Tésenyi, P. Herbison, and M. Farella, “Time-frequency analysis of chewing activity in the natural environment,” *Journal of dental research*, vol. 90, no. 10, pp. 1206–1210, 2011.

[20] *Infrared Array Sensor Grid-EYE datasheet*, Panasonic, 4 2017, accessed: 09/2020. [Online]. Available: https://industry.panasonic.eu/components/sensors/industrial-sensors/grid-eye/amg88xx-high-performance-type/amg8833-amg8833

[21] D. J. Bernstein, “New stream cipher designs,” M. Robshaw and O. Billet, Eds. Berlin, Heidelberg: Springer-Verlag, 2008, ch. The Salsa20 Family of Stream Ciphers, pp. 84–97. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-68351-3_8

[22] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[23] A. Kirillov, R. Girshick, K. He, and P. Dollár, “Panoptic feature pyramid networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6399–6408.

[24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 4510–4520.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in *CVPR09*, 2009.

[26] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[27] Y. Hong, H. Pan, W. Sun, Y. Jia *et al.*, “Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes,” *arXiv preprint arXiv:2101.06085*, 2021.

[28] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, “Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations,” in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2017, pp. 240–248.

[29] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[30] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” *arXiv preprint*, pp. 1610–02 357, 2017.

[31] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: common objects in context,” *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[32] T. A. Sorensen, “A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons,” *Biol. Skar.*, vol. 5, pp. 1–34, 1948.

[33] K. Lee, A. Shrivastava, and H. Kacorri, “Hand-priming in object localization for assistive egocentric vision,” in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 3422–3432.

[34] L. Chan, Y.-L. Chen, C.-H. Hsieh, R.-H. Liang, and B.-Y. Chen, “Cyclopsring: Enabling whole-hand and context-aware interactions through a fisheye ring,” in *Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST ’15. New York, NY, USA: Association for Computing Machinery, 2015, p. 549–556. [Online]. Available: https://doi.org/10.1145/2807442.2807450

[35] S. Sridhar, A. Markussen, A. Oulasvirta, C. Theobalt, and S. Boring, “Watchsense: On-and above-skin input sensing through a wearable depth sensor,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017, pp. 3891–3902.

[36] I. Ulku and E. Akagunduz, “A survey on deep learning-based architectures for semantic segmentation on 2d images,” *arXiv preprint arXiv:1912.10230*, 2019.

[37] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt, “Real-time hand tracking under occlusion from an egocentric rgb-d sensor,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 1284–1293.