Person Re-ID Testbed with Multi-Modal Sensors

Guangliang Zhao*
GE Research
Controls & Optimization
Niskayuna, New York, US
glzhao@ge.com

Guy Ben-yosef*
GE Research
Digital Research
Niskayuna, New York, US
guy.ben-yosef@ge.com

Jianwei Qiu*
GE Research
Digital Research
Niskayuna, New York, US
jianwei.qiu@ge.com

Yang Zhao
GE Research
Digital Research
Niskayuna, New York, US
yang.zhao@ge.com

Prabhu Janakaraj
GE Research
Controls & Optimization
Niskayuna, New York, US
Prabhu.Janakaraj@ge.com

Sriram Boppana
GE Research
Advanced Technology
Niskayuna, New York, US
sriram.boppana@ge.com

Austars R Schnore
GE Research
Controls & Optimization
Niskayuna, New York, US
schnore@ge.com

ABSTRACT

Person Re-ID is a challenging problem and is gaining more attention due to demands in security, intelligent system and other applications. Most person Re-ID works are vision-based, such as image, video, or broadly speaking, face recognition-based techniques. Recently, several multi-modal person Re-ID datasets were released, including RGB+IR, RGB+text, RGB+WiFi, which shows the potential of the multi-modal sensor-based person Re-ID approach. However, there are several common issues in public datasets, such as short time duration, lack of appearance change, and limited activities, resulting in un-robust models. For example, vision-based Re-ID models are sensitive to appearance change. In this work, a person Re-ID testbed with multi-modal sensors is created, allowing the collection of sensing modalities including RGB, IR, depth, WiFi, radar, and audio. This novel dataset will cover normal daily office activities with large time span over multi-seasons. Initial analytic results are obtained for evaluating different person Re-ID models, based on small datasets collected in this testbed.

CCS CONCEPTS

• **Information systems** → *Extraction, transformation and loading.*

ACM Reference Format:
Guangliang Zhao*, Guy Ben-yosef*, Jianwei Qiu*, Yang Zhao, Prabhu Janakaraj, Sriram Boppana, and Austars R Schnore. 2021. Person Re-ID Testbed with Multi-Modal Sensors . In *The 19th ACM Conference on Embedded Networked Sensor Systems (SenSys '21), November 15–17, 2021, Coimbra, Portugal.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3485730.3493366

erson Re-ID, Multi-Modal, WiFi, Computer Vision, Face Recognition, Deep Learning, Neural Network

1 INTRODUCTION

Person re-identification (Re-ID) has drawn more attention in recent years, especially, there are more recent studies regarding multi-modal Re-ID using more than one sensor modalities [2, 6, 7]. Re-ID has applications in many use cases including monitoring and surveillance, badge-less entry, retail autonomous checkout, etc. The problem setup includes a pre-acquired gallery of IDs, and a query of IDs to be matched with the gallery for person re-acquisition. Re-ID methods are typically evaluated using the top-k accuracy, that is, accuracy is defined as when the same ID is found in the top k candidates from gallery (e.g., k=1, 5, or 10).

Among different sensor modalities, vision-based models are the most versatile and most widely studied for person Re-ID. Facial feature is one of the most important visual identification features. Face recognition is currently a popular research topic with promising results [3, 4, 16]. However, face recognition usually requires high resolution face image, which may not be always feasible. Therefore, person Re-ID typically focus on the whole-body appearance. There are dozens of publicly available vision datasets for Re-ID [8]. These datasets significantly advanced the development of the Re-ID technology. The datasets were typically collected using surveillance cameras at distance from different locations with different viewing angles. The data collection time span is usually short, and the same person's appearance under different cameras is typically not changed. The pixel size of a person is usually small. There are image-based Re-ID methods which uses non-consecutive frames to extract visual features, and video-based Re-ID method that use consecutive frames to obtain spatial-temporal cues for feature matching [26]. Recent studies also show that whole-body Re-ID models focusing on appearance features will fail if appearance changes [6].

Radio frequency (RF) sensors are also used for person Re-ID in recent years, and various RF sensors are studied including software defined radio (SDR) [6], mmWave radar [2], and commodity WiFi [7]. SDR allows user to control the frequency band and is most versatile in capturing human features and have been used in recognizing human activities with promising results [1]. However, the in-house devices are expensive and difficult to setup for comparison with radar, WiFi and other sensing modalities. mmWave radars are low-cost commodity devices that have been used for pedestrian detection in automotive industry, and outperforms camera in certain scenarios like low light conditions [18]. The radar data granularity may be lower than SDR but higher than WiFi, but it also has a

[*] These authors contributed equally to this work

ACM DATA Workshop co-located with SenSys'21 and BuildSys'21, November 17, 2021, Coimbra, Portugal, 15.00

ACM ISBN 978-1-4503-9097-2/21/11…$15.00
https://doi.org/10.1145/3485730.3493366

smaller coverage area than WiFi. Recent studies have shown that radar is capable for person identification [2]. Commodity WiFi has been studied for person pose [20] and person identification [7]. When a person is performing activities under WiFi coverage, received signal strength indicator (RSSI) or channel state information (CSI) are captured from WiFi devices, and these dynamic changes could be extracted for person Re-ID. WiFi signal has larger coverage area, less granularity than SDR or radar, and is noisy. WiFi is also sensitive to the environment, including room layout, furniture location, nearby people activities, other WiFi interference, etc. Subtle environment changes may result in significant changes in RSSI and CSI. Many studies set strictly confined environment, and limit the person activity to certain types.

Recent studies have developed environment agnostic models that were trained with multi-modal sensors, which is applicable in different environments and is robust to certain environmental changes [7]. Inspired by these multi-modal person Re-ID studies, this work aims to create a person Re-ID testbed with multi-modal sensors, including RGB image, IR image, depth point cloud, WiFi CSI, radar, and audio data. The testbed will create a novel multi-modal dataset that tackles some common problems in the existing published datasets. 1. Large time span: the testbed will massively collects data 24/7 continuously without environmental constraint for several months spanning from summer to winter. 2. Unconstrained activities: with the testbed, people will conduct normal daily office activities as usual. 3. Appearance change: the same person will have various appearance including different clothing and accessories at different time. The data will be annotated by ID and tracklet instances. With the data collected, the performance of various sensing techniques can be compared under different conditions in the same environment, and the benefits of adding one modality to other modalities can be evaluated as well. The deployed testbed includes two sites with high traffic on a campus of around one thousand people, with a total of 5 Intel RealSense cameras which product RGB, IR and depth data simultaneously, 2 WiFi CSI subsystems with 1 and 3 WiFi links respectively, 2 radars, and 2 ReSpeaker mic arrays. To evaluate performance of different sensing techniques, different experiments with strictly confined environment were performed. This is an ongoing study, and initial analytical results are obtained with a small dataset annotated.

The major contributions of this paper are as follows. We propose a person Re-ID testbed with multi-modal sensors, including a novel dataset containing long recordings of multiple sensors. We also provide initial analytical results for model testing with a subset of the dataset. The rest of this paper is organized as follows. Section 2 describes the system architecture and the baseline multi-modal person Re-ID methods. Section 3 describes the testbed implementation layout, experiment details, and some initial results for testing different Re-ID models using small annotated dataset from the testbed. Section 4 concludes this study with some future works.

2 SYSTEMS AND METHODS

2.1 Overall system

The testbed is a system of systems, namely combined multiple sensing systems that are hosted by different computers, and these systems are loosely coupled. The vision system is responsible for person detection, and for triggering other systems. The vision detection results are also used as groundtruth for person detection and person localization by other systems. Denote the computer hosting vision system as PC1. Whenever the vision system detects a person under a camera, a flag on PC1 corresponding to this camera is marked "1". For each non-vision sensor, there is a process on PC1 monitoring all the flags frequently. If one of the flags is true, it remotely fetch the data from other computer which installs the sensor. The data is piped through ssh tunnel directly to PC1, and timestamped with PC1's system time. Though there might be very small delay when piping other sensor data, all the data modalities use the same system time in PC1 for each data point. This resolves the data synchronization problem.

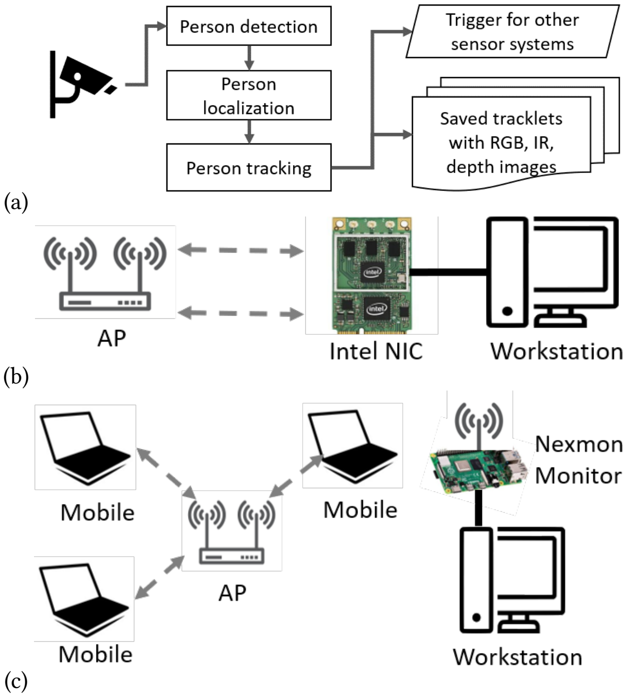

Figure 1: System Architectures for the data collection testbed: (a) vision system, (b) Intel 5300 WiFi CSI, (c) Nexmon WiFi CSI

2.1.1 *Computer vision system.* The vision system includes three analytics: person detection, person localization, and person tracking. Person detection module uses Yolo V5 [19] for localizing person bounding box on the image during inference. Person localization module uses both the provided camera intrinsic parameters and the estimated camera extrinsic parameters, and then estimates the detected person's location in real-world coordinate system. Person tracking module uses detection-based tracking model DeepSort [22]. Intel RealSense Lidar camera L515 is used as the multi-modal vision sensor. The camera streams include RGB, IR, and depth. The depth is obtained using IR laser, which could be interfered with ambient light or be absorbed by target materials, resulting a short working range. Different streams can be aligned and depth can be transferred to 3D point cloud with the given camera intrinsic parameters. The perspective-n-point method is used to estimate camera extrinsic. To do this, the 3D real-world coordinate system is defined, and four 3D points in the coordinate system and their corresponding 2D projections in the image are annotated.

For person localization, due to the short range of the depth in RealSense L515, only RGB image is used for localizing a person during inference. First, the plane of ground floor is estimated using RANSAC algorithm from the depth image. Then for a detected person, the center pixel coordinate of the feet is converted to 3D point corresponding to camera where the ray from camera origin meets the ground plane, assuming a person is always on the ground. Lastly, the 3D point corresponding to camera is converted to the real-world 3D coordinate, and the location of the person is obtained.

For each input iteration whenever there's a detected person, output includes the cropped RGB image, IR image, and depth image. The trigger is set for other sensing systems to collect data, until the person leaves the scene of camera. For test area that is covered by multiple cameras with overlapped views, the trigger can be set by any camera so that the data collection of other senses continues until the person leaves the scenes of all cameras. For sensing system hosted by different computer, SSH tunnel is used to redirect captured data to local file, so that all the sensor data will use the same system time. The architecture of vision system is shown in Figure 1 (a).

2.1.2 WiFi system. As the IEEE 802.11n standard and WiFi MIMO devices are widely used nowadays, WiFi channel state information (CSI)-based system becomes a promising solution for human sensing with low-cost commodity hardware. We have investigated two commodity WiFi hardware devices for developing our WiFi CSI-based Re-ID systems: Intel 5300 network interface card (NIC) with open-source IEEE 802.11n toolkit [11], and regular WiFi NIC with Nexmon CSI extractor [9].

From the Intel system, we collect CSI measurements between a WiFi router access point (AP) and the Intel 5300 NIC mounted on a workstation. The Intel 5300 NIC reports CSI from 3x3 MIMO antennas for 30 OFDM subcarriers with a total bandwidth of 20 MHz [11]. For the Nexmon CSI system, we use a Raspberry Pi board with the Nexmon toolkit as the standalone monitor to collect the CSI data between WiFi AP and all connected IEEE 802.11n devices. The architectures of the Intel 5300 system and Nexmon system are shown in Figure 1 (b) and (c), respectively.

2.2 Re-id based on face recognition

A pipeline including face detection and face identification was used to evaluate the novel dataset and to provide baseline results. Since visual face recognition features are typically powerful for identification tasks, we hypothesized that face-based visual features would probably be able to perform good identifications in most cases. Our goal was then to rather track failure modes of state-of-the-art face models, keeping in mind that the success of non-visual sensory models on such face-failed cases would contribute to a multi-modal re-id system. Our face recognition pipeline is based on three stages: (i) A RetinaFace model [3] for face detection, (ii) face alignment stage on which face boxes are rotated and translated to canonical pose based on 5 detected face landmarks (provided by the RetinaFace model as well), (iii) a face embedding stage based on the MagFace model [16], in which each detected face box is represented by a 512-D feature vector by which similarity to other faces is computed.

2.3 Whole body Video based re-id framework

Whole body person Re-ID is a challenging computer vision task due to complex environments, such as changes in illumination, different camera viewpoints, low image resolution and potential person occlusion in images. Deep learning-based person whole body Re-ID has achieved very promising results on public benchmark datasets, such as MARS [28], iLIDS-VID [23] and DukeMTMC-VideoReID [21]. The whole body person Re-ID can be divided into two categories: image-based methods and video-based methods. Image-based Re-ID methods only extract the spatial features from a single image. In contrast, video based-Re-ID methods extract both spatial and temporal information from video sequence, which enables more robust and reliable person feature representation. However, person appearance misalignment is a common issue in video-based whole body person Re-ID due to imperfect person detection/tracking and potential motion blur during video acquisition. Currently, two methods are commonly used to mitigate this issue: graph-based temporal feature alignment [24] and 3D convolution- based temporal feature alignment [10]. Gu et al. [10] propose Appearance-Preserving 3D convolution (AP3D) to resolve the appearance misalignment in temporal domain, and it achieves state of the art result on MARS, iLIDS-VID, and DukeMTMC-VideoReID. In our video-based Re-ID pipeline, we leverage the existing AP3D network with a Resnet-50 [13] pretrained on ImageNet [17] as backbone. The revised Resnet-50 maintains similar network architecture with the original version except 2D residual blocks are replaced with AP3D residual blocks and total number of classes are changed to number of person ID in corresponding video Re-ID dataset. With this modification, all 2D convolution layers are turned into AP3D convolution layers, which allows the network to extract spatio-temporal from input video sequence.

2.4 Pose estimation based on WiFi

Human pose information can be used to infer the internal body dynamics, which carries out identity information based on gait [5, 14]. A multi-senatorial Re-ID system can therefore be benefited by pose estimation based on non-visual sensor, such as WiFi, and which provides identity also when the visual signal is missing or weak. Our pose estimation framework is using a set of synchronized pairs of CSI measurements and video frames. It includes a teacher-student DNN model, similar to [20, 27], in which during training a vision model is predicting a pose from the video frame, and compares it to the pose that is predicted by a WiFi model. Our vision pose model is based on a Mask-RCNN [12] model, which extracts 3D coordinates of 17 body joints. Our WiFi model is a Multi-Perceptron model which receives the CSI data as input, and outputs 3D coordinates of 17 body joints. The visual network is freezed during training, while the WiFi model is learned.

2.5 Fusion Strategies

Multiples sensory processing modules require methods for fusing information for improving the Re-ID performance. In this work we are using three main fusion strategies: (i) Integration scores of quasi-identifiers coming from different modalities (e.g., visual face recognition, visual whole-body recognition, WiFi-Gait recognition, etc.). (ii) A student-teacher learning strategy, where a strong vision model provides supervised examples for a weaker non-vision model targeting the same task (e.g., the task of body pose estimation from WiFi or Radar). (iii) A tracking strategy, where a Re-ID task is successfully achieved by different modalities at different time segments, but continuously maintained (e.g., Radar and vision models enable continuous person localization, while a vision-only model cannot).

3 EXPERIMENTS AND RESULTS

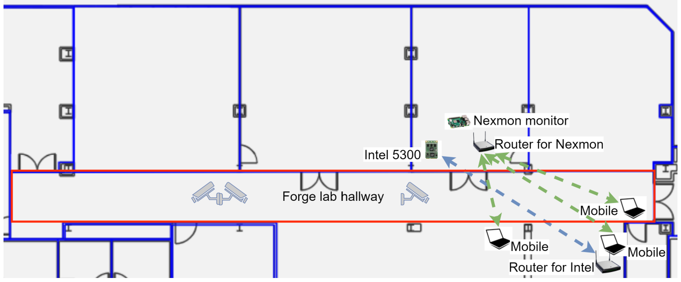

Figure 2: Experiment layout. Red rectangle is the Forge hallway. Blue dotted line is Intel WiFi link. Greed dotted lines are Nexmon WiFi links.

The experiment has a primarily setup in GE Research Forge lab and a secondary setup in main lobby. In Forge lab, 3 RealSense L515 cameras are mounted on the roof to cover the whole hallway with approximate length of 33 meters and width of 2 meters, where most of the traffic happens. Two WiFi sensing systems, Intel 5300 system with single WiFi link, and Nexmon system with 3 WiFi links, are both set at the east side of the hallway, mainly for comparison purpose. The layout of both WiFi systems can be found in Figure 2. The main lobby setup has 2 RealSense L515 cameras mounted above entrance gates.

Different types of experiments are conducted in the testbed. Some experiments are under confined environment, where the Forge lab is restricted to access temporarily to prevent unexpected interference, and certain activities in the hallway for different studies. Some experiments perform different working patterns, some experiments perform appearance changes. Most data are collected under unconstrained environment with normal daily office activities, like group chat, carrying bags or office equipment, security and janitor activities, and passing the hallway is the most frequent activity.

The vision system runs 24/7 to generate tracklets during the time persons appear in the camera field of view. Each tracklet is a person instance detected by vision pipeline with multiple frames from the vision tracking model, and other sensor modality data is also saved simultaneously. During annotation, all the instances are assigned to their corresponding IDs. WiFi, radar and audio data are fused with tracklet frames by their timestamps. On average, every month there are around 2500 tracklets saved for one camera. Note that one person's single pass under one camera may result in more than one tracklets.

Face re-id pipeline: Our experiments consisted of a RetinaFace model pretrained on the WiderFace dataset [25] and based on a ResNet50 backbone, and a MagFace model pretraind on the MS1MV2 dataset [4] and based on a MobileNetV1 backbone. Since the recordings of our data were taken during the covid19 pandemic time, a significant potion of the recorded faces were covered by masks. For this reason we have used pretrained models augmented with masked faces examples at training. Quantitative ID accuracy evaluation of an annotated subset of our data was shown high identification rate (91.8% of total number of frames). However, the multiple detection and identification failures occur, which will serve for bench-marking future models. Challenging factors for face identification are currently being explored quantitatively including: face image quality and resolution, masked faces, crowded scenes, occlusions, and extreme pose and illumination.

Whole body re-id pipeline: Due to the limited video data annotation in our current testbed, we leverage the existing public datasets as our training dataset. To increase the robustness of the extracted spatiotemporal features from query video sequence, we create a combined dataset by combining three public video Re-ID datasets: MARS, iLIDS-VID and DukeMTMC-VideoReID. This combined dataset has total number of 1477 unique person ID. The AP3D with Resnet-50 model is trained by using this combined dataset. With public benchmark datasets, AP3D achieves about 90.1% on MARS and 96.3% on DukeMTMC-VideoReID for the top1 accuracy [10]. To evaluate the performance of the model trained on the combined dataset, we collect an evaluation dataset that has 18 unique persons with very different appearance as shown in Figure 3. There is a total of 113 person tracklets, and we further split this dataset into gallery and query set. With the gallery set, we pre-extract the spatiotemporal feature vector with size of 2048 for each person and store them in a dictionary format for inferencing. During the inference or test mode, same size feature vector is extracted from each video sequence in the query set, and then feature comparison against the pre-extracted gallery features is performed using the cosine similarity distance metric. At the end, distance-based ranking is done by sorting the distance from the smallest to largest. The top 1 ranking is our final predicted person ID as shown in Figure 3. With total of 47 query video sequence, the whole body Re-ID model trained with our combined public dataset achieves 70.2% accuracy for top1.

WiFi pose estimation pipeline: WiFi-pose experiments included a Mask-RCNN based model trained on CoCo [15], based on a ResNet50 backbone for detecting human body box together with body keypoitns [12]. The MultiPerceptron model including 7 linear layers was used, followed by ReLu non-linearity. Input includes flatten CSI measurements from 5 consecutive RF measurements (total of 1350 input elements). For preliminary training we have used 50 video segments synched with CSI data from two individuals presenting different pose configurations at different visual conditions. Sample predictions are shown in Fig. 4.

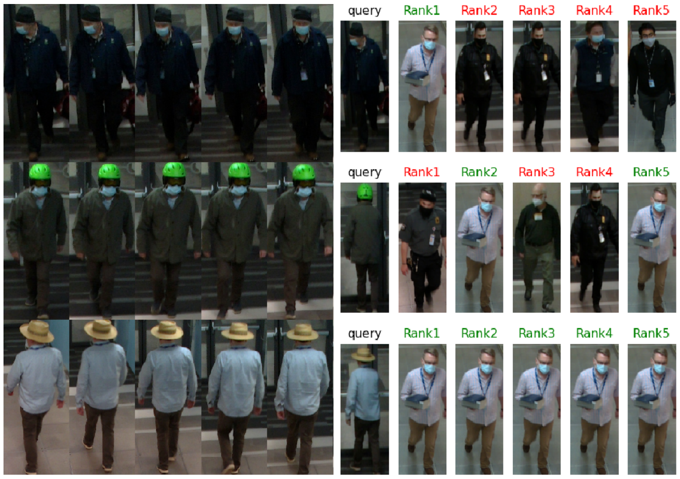

Figure 3: Video Re-ID result for human subjects with different appearance. Correct ReID results are highlighted in green, and detection errors are highlighted in red.

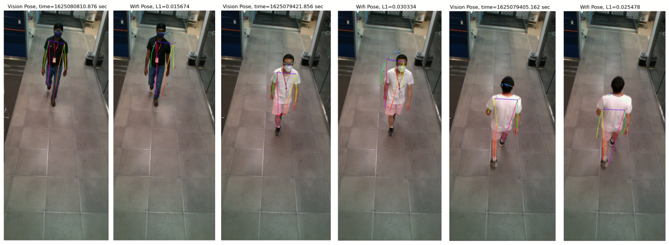
Vision Pose, time=1625080810.876 sec
Wifi Pose, L1=0.015674
Vision Pose, time=1625079421.856 sec
Wifi Pose, L1=0.030334
Vision Pose, time=1625079405.162 sec
Wifi Pose, L1=0.025478

Figure 4: WiFi-pose predictions (right) compared with vision-based pose predictions (left).

4 CONCLUSION

This work developed a multi-modal person Re-ID testbed by a loosely coupled system of different sensing subsystems, and created a novel multi-modal dataset that tackles several common problems in existing published dataset, such as short time span, no or limited appearance change. Initial analytic results for testing different Re-ID models under different conditions are obtained using small datasets collected from the testbed. The data collection is ongoing and data annotation is work in progress. Dataset is planned to be released in future.

Future work will need to investigate how to protect individual privacy and sensitive data for a real world application. The current proposed framework parses all the raw data to a central server without masking sensitive information or anonymizing individual identities. Future work may investigate deploying the Re-ID algorithms to an edge device network where feature extraction is performed by the multimodal sensors near the source of the data. The extracted features are then sent over the network to a central server to be classified for Re-ID. This keeps the relevant information while anonymizing sensitive data.

ACKNOWLEDGMENT

This work is supported by DARPA under contract number N6600118C4006.

REFERENCES

[1] Fadel Adib and Dina Katabi. 2013. See through walls with WiFi!. In *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*. 75–86.

[2] Yuwei Cheng and Yimin Liu. 2021. Person Reidentification Based on Automotive Radar Point Clouds. *IEEE Transactions on Geoscience and Remote Sensing* (2021).

[3] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5203–5212.

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4690–4699.

[5] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. 2020. Gaitpart: Temporal part-based model for gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14225–14233.

[6] Lijie Fan, Tianhong Li, Rongyao Fang, Rumen Hristov, Yuan Yuan, and Dina Katabi. 2020. Learning longterm representations for person re-identification using radio signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10699–10709.

[7] Shiwei Fang, Tamzeed Islam, Sirajum Munir, and Shahriar Nirjon. 2020. EyeFi: Fast Human Identification Through Vision and WiFi-based Trajectory Matching. In *2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, 59–68.

[8] Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia Camps, Richard J Radke, et al. 2018. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *IEEE transactions on pattern analysis and machine intelligence* 41, 3 (2018), 523–536.

[9] Francesco Gringoli, Matthias Schulz, Jakob Link, and Matthias Hollick. 2019. Free your CSI: A channel state information extraction platform for modern Wi-Fi chipsets. In *Proceedings of the 13th International Workshop on Wireless Network Testbeds, Experimental Evaluation & Characterization*. 21–28.

[10] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. 2020. Appearance-preserving 3d convolution for video-based person re-identification. In *European Conference on Computer Vision*. Springer, 228–243.

[11] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. 2011. Tool release: Gathering 802.11 n traces with channel state information. *ACM SIGCOMM Computer Communication Review* 41, 1 (2011), 53–53.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[14] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, Shiqi Yu, and Mingwu Ren. 2020. End-to-end model-based gait recognition. In *Proceedings of the Asian conference on computer vision*.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[16] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. 2021. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14225–14234.

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115, 3 (2015), 211–252.

[18] Xian Shuai, Yulin Shen, Yi Tang, Shuyao Shi, Luping Ji, and Guoliang Xing. 2021. milliEye: A Lightweight mmWave Radar and Camera Fusion System for Robust Object Detection. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*. 145–157.

[19] Ultralytics. 2021. YOLOv5. https://github.com/ultralytics/yolov5/wiki.

[20] Fei Wang, Sanping Zhou, Stanislav Panev, Jinsong Han, and Dong Huang. 2019. Person-in-WiFi: Fine-grained person perception using WiFi. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5452–5461.

[21] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. 2014. Person re-identification by video ranking. In *European conference on computer vision*. Springer, 688–703.

[22] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*. IEEE, 3645–3649.

[23] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. 2018. Exploit the unknown gradually: One-shot video-based person re-identification

by stepwise learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5177–5186.

[24] Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, and Ling Shao. 2020. Learning multi-granular hypergraphs for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2899–2908.

[25] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. 2016. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5525–5533.

[26] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[27] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7356–7365.

[28] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. 2016. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*. Springer, 868–884.