VAX: Using Existing Video and Audio-based Activity Recognition Models to Bootstrap Privacy-Sensitive Sensors

PRASOON PATIDAR, Carnegie Mellon University, USA
MAYANK GOEL, Carnegie Mellon University, USA
YUVRAJ AGARWAL, Carnegie Mellon University, USA

The use of audio and video modalities for Human Activity Recognition (HAR) is common, given the richness of the data and the availability of pre-trained ML models using a large corpus of labeled training data. However, audio and video sensors also lead to significant consumer privacy concerns. Researchers have thus explored alternate modalities that are less privacy-invasive such as mmWave doppler radars, IMUs, motion sensors. However, the key limitation of these approaches is that most of them do not readily generalize across environments and require significant in-situ training data. Recent work has proposed cross-modality transfer learning approaches to alleviate the lack of trained labeled data with some success. In this paper, we generalize this concept to create a novel system called VAX (Video/Audio to 'X'), where training labels acquired from existing Video/Audio ML models are used to train ML models for a wide range of 'X' privacy-sensitive sensors. Notably, in VAX, once the ML models for the privacy-sensitive sensors are trained, with little to no user involvement, the Audio/Video sensors can be removed altogether to protect the user's privacy better. We built and deployed VAX in ten participants' homes while they performed 17 common activities of daily living. Our evaluation results show that after training, VAX can use its onboard camera and microphone to detect approximately 15 out of 17 activities with an average accuracy of 90%. For these activities that can be detected using a camera and a microphone, VAX trains a per-home model for the privacy-preserving sensors. These models (average accuracy = 84%) require no in-situ user input. In addition, when VAX is augmented with just one labeled instance for the activities not detected by the VAX A/V pipeline (~2 out of 17), it can detect all 17 activities with an average accuracy of 84%. Our results show that VAX is significantly better than a baseline supervised-learning approach of using one labeled instance per activity in each home (average accuracy of 79%) since VAX reduces the user burden of providing activity labels by 8x (~2 labels *vs.* 17 labels).

CCS Concepts: • **Human-centered computing** → **Ambient intelligence**; • **Computer systems organization** → *Sensors and actuators.*

Additional Key Words and Phrases: ubiquitous sensing, privacy first design, human activity recognition

ACM Reference Format:
Prasoon Patidar, Mayank Goel, and Yuvraj Agarwal. 2023. VAX: Using Existing Video and Audio-based Activity Recognition Models to Bootstrap Privacy-Sensitive Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 117 (September 2023), 24 pages. https://doi.org/10.1145/3610907

1 INTRODUCTION

Human Activity Recognition (HAR) within home environments enables compelling applications around Active and Assisted Living (AAL), healthcare monitoring, security and surveillance, and tele-immersion applications [85]. As a result, over the last few decades, researchers built more accurate and practical HAR systems driven by

Authors' addresses: Prasoon Patidar, prasoonpatidar@cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; Mayank Goel, mayank@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; Yuvraj Agarwal, yuvraj@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA.

2474-9567/2023/9-ART117
https://doi.org/10.1145/3610907

inexpensive sensors [14, 56] and rapid advances in ML algorithms [84]. Given the advances in computer vision and audio processing and the presence of several large annotated datasets (*e.g.*, [19], [40], [2]), researchers have built ML models that can perform audio-based [29, 54], and video-based activity and body pose estimation [66, 94]. Notably, these pre-built ML models for Audio/Video (A/V) data are meant to be generalizable and work reasonably well in new outside-the-lab settings without requiring any in-situ training data. Consumers, however, are increasingly finding IoT sensors, particularly cameras and microphones in their personal spaces invasive [1, 16, 32, 33, 48]. The problem is further compounded, given numerous high-profile incidents of compromised IoT devices [6, 7, 39, 87].

Apart from microphones and cameras, researchers have also explored other sensing modalities, such as IMUs (accelerometers, gyroscopes), low-fidelity thermal arrays, high-fidelity thermal image sensors, WiFi signals, Doppler RADARs, and a slate of environmental sensors such as temperature, humidity, and PIR for movement [9, 10, 55, 56, 58, 91]. The key advantage of these approaches over video and audio-based HAR is that they are more privacy sensitive, and the data captured is often not personally identifiable, alleviating many privacy concerns. However, these approaches still suffer from (a) requiring in-situ training data for building supervised learning models which comes with a significant user burden; and (b) not enough training data to create generalized models that work in the real world without training. Most recently, researchers have explored domain adaptation [20], *i.e.*, converting training data from a richer (but more invasive) sensing modality to a more privacy-sensitive modality. For example, Ahuja *et al.* created synthetic Doppler RADAR data from videos captured to train a model for a set of motion-related activities [5]. Such approaches are still limited to a small set of activities (*e.g.*, activities with large body motion or activities that generate descriptive sound or motion signals). To build an approach that generalizes to a wider range of activities, we often have to rely on multimodal sensing but generating synthetic training data for all these modalities is not yet possible.

In this paper, instead of generating synthetic signals, we propose VAX, a hybrid approach that uses output labels from off-the-shelf models to perform *in-situ* training of specific sensors. VAX allows to quickly learn in-situ ML models for a variety of 'X' privacy-sensitive sensors (*e.g.*, Lidar, mmWave radar, low fidelity thermal cameras) using training labels obtained from models built on existing Audio and Video (A/V) datasets. This approach can go beyond the set of sensors explored in this paper and can generalize to a wide range of sensors without relying on mathematical or empirical modeling across different signals. We envision when starting to use VAX, users would augment it with a camera and microphone for a small duration (on the order of days). The camera and microphone, during this period, will use existing audio/video-based ML models to generate training labels for activities performed in the user's environment and train the rest of the 'X' sensors on VAX. Once the privacy-sensitive sensors are trained, VAX can inform the user to remove the microphone and camera altogether. The design of VAX is based on the following key insights: (1) current generation pre-built A/V models are reasonably good at identifying a variety of HAR activity patterns [29, 30]; (2) these A/V models are usually generalizable and work across environments; (3) combining multiple privacy sensitive sensors allows us to leverage the sensor data for each modality to build accurate per-home models for a diverse set of activities; (4) a limited set of activities, where the A/V models are inaccurate, need a small amount of in-situ training data to improve accuracy further.

In designing and implementing VAX[1], we had to solve three major technical challenges. (1) while there exist a number of off-the-shelf A/V-based ML models, they are trained with a diverse (and non-exhaustive) set of labels and are not uniformly accurate across all activities [30, 62], which makes the selection of appropriate pre-built A/V models and combining their outputs non-trivial. (2) overall accuracy of a wide range of A/V models on native datasets is less than 75-80% [69], and as a result, incorrect labels from sub-optimal A/V models will lead to inaccuracies in training the models on the privacy-sensitive sensors, and (3) data collected from different privacy

[1] www.github.com/synergylabs/vax

sensitive modalities is heterogeneous, thus, how to train user level activity recognition models for any set of 'X' privacy sensitive modalities.

To address these challenges, VAX bootstraps a set of off-the-shelf A/V-based ML models with some labeled data from a set of *starter* homes and then uses this ensemble of models to predict activities in new homes. We call this part of VAX the *A/V pipeline*. We then propose an unsupervised learning approach that utilizes unlabelled data from 'X' sensors to increase our activity detection rate and reduce the impact of erroneous labels from the A/V pipeline. In addition, we propose an approach to train activity recognition models across privacy-preserving modalities using the labels provided by the A/V pipeline. We collected data for 17 activities in 10 participant's homes and show that VAX can detect activities with an accuracy of 74% across all activities, and 84% across activities detected by our A/V pipeline (15 out of 17). Given that existing A/V models cannot yet accurately detect all activities across all homes, we propose that the user provides one labeled example for each "undetected" activity (approximately 2 activities per home). With such minimal user input, VAX's accuracy improves to approximately 84% (as compared to a baseline case of 79% where the user provides an input for each of the 17 activities and does not benefit from the bootstrapping provided by A/V pipeline). Note, the baseline also incurs 8× user burden for providing training labels (~2 labels *vs.* 17 labels).

In summary, we make the following contributions:

- We present VAX, an end-to-end framework for training activity recognition models for a set of 'X' privacy-sensitive sensing modalities using labels from off-the-shelf A/V models.
- We present a novel method to bootstrap and combine these off-the-shelf A/V models with the unlabelled data from 'X' modalities to detect activities of daily living in a new home.
- We evaluated VAX with 10 participants performing 17 activities in three different locations (Kitchen, Living Room, and Bathroom) in their homes. We show that VAX can detect activities with an average accuracy of 84% using labels generated by VAX's A/V pipeline and only one input for activities which are not present in our A/V labels (~2 labels *vs.* 17 labels).
- We show that without any user input, VAX performs considerably better (74% *vs.* 38%) when compared with baseline approach (i.e., directly training models for activity recognition on privacy-sensitive sensors). We also show that with user input, VAX has comparable performance (84% *vs.* 79%) with significantly less labeling effort (2 labels/home *vs.* 17 labels/home) from users.

2 RELATED WORK

We have organized the related work into three categories. First, we review prior work on Human Activity Recognition (HAR) using audio and visual data as well as other privacy-sensitive modalities. Next, we discuss domain adaptation techniques across heterogeneous sensing modalities to address the scarcity of labeled data in the target domain. Finally, we discuss self-supervised learning (SSL) approaches in HAR, which focus on learning informative representations from unlabelled data, and augmenting it with a small amount of labeled data to build accurate ML models.

2.1 Human Activity Recognition (HAR) using Different Sensing Modalities

HAR is a well-researched problem, with researchers proposing different sensing modalities and various ML methods to detect activities. With rapid advances in deep learning and the availability of large-scale video datasets with 100s of annotated labels [19, 42, 92], the use of video data for HAR has become very popular (See [77] for a comprehensive survey). Notably, platforms such as OpenMMLab have emerged for video understanding (i.e MMAction2 [69], MMPose [70], MMDetection [23], etc.) which leverage pre-trained, and open-source, state-of-the-art action recognition models using video data. Similarly, acoustic (sound-based) activity recognition has also been explored for generalized HAR across various settings [29, 54, 71]. The approaches use microphone data to detect activities with unique sound signatures. Often these audio-based approaches convert sound signals into a 2-D image using Fast Fourier Transforms (FFTs) and then generate a feature representation using pre-trained deep CNN-based architectures. Large-scale datasets like AudioSet [37] and YouTube-8M [2] consist of labels for hundreds of diverse activity classes and are used to train generic audio-based models. While using audio and video data is indeed promising for HAR, their widespread adaptability is limited due to significant privacy and security concerns [16, 79]. In addition, the real-world accuracy for these models individually for common HAR tasks is still not that high [69].

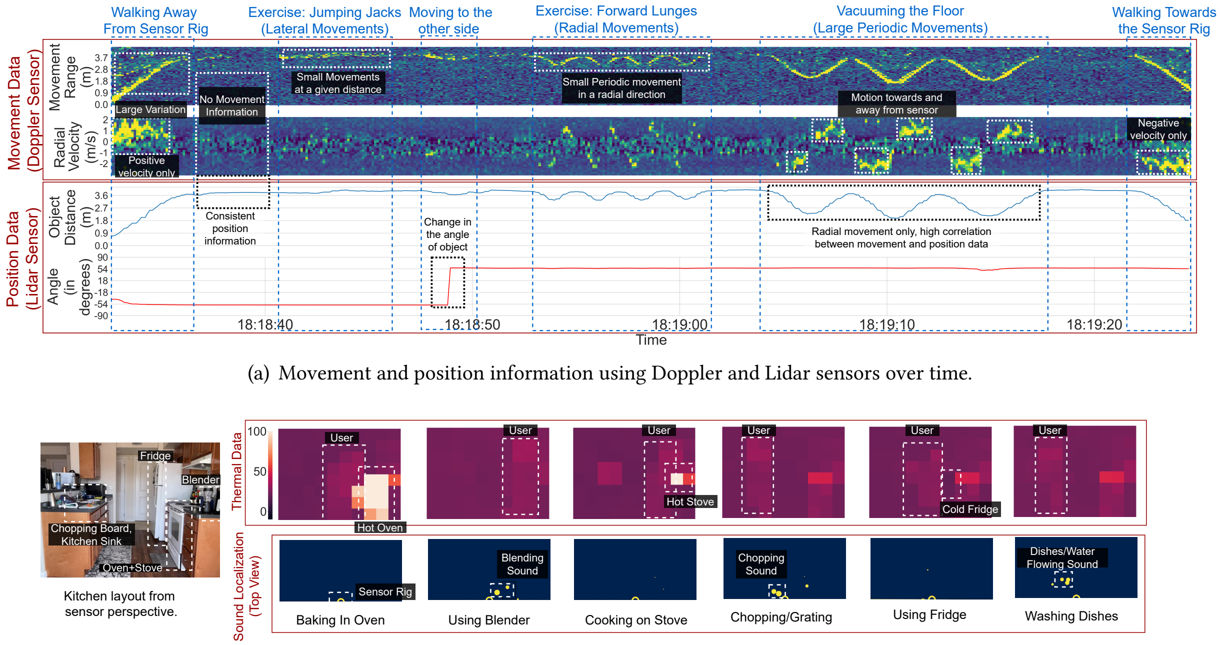
Researchers have also proposed other sensing modalities such as mmWave Doppler Radar, IMUs, Lidars, thermal arrays, pressure, humidity, temperature *etc.* for HAR [5, 12, 46, 53]. While these sensing modalities may seem less capable than video and audio based HAR, they have a distinct advantage in terms of reducing privacy concerns since the data sensed is generally not personally identifiable [83]. Building upon these approaches, multi-modal sensors and sensor fusion has also been proposed for more accurate HAR [4, 68, 73, 75, 78]. Some works also explore generative and adversarial models to generate synthetic data to reduce privacy burden by hiding user identity while supporting a particular class of applications [64, 102]. Most of these supervised learning-based approaches rely on significant in-situ labeled data to provide high-accuracy HAR. In addition, they are tested in controlled settings only and do not generalize well across different environments [18].

2.2 Domain Adaption Methods for Cross-Modality Learning

To reduce the labeling burden, researchers have explored domain adaption to transform raw data from one domain (source) to another (target) domain [21]. Relevant to our work, some of these domain adaptation techniques use video or audio data as the source domain [5, 8, 15, 53, 82, 97]. Vid2Dopler [5], for example, created a synthetic training data set for Doppler sensors using videos of exercises as examples. IMU2Doppler [12] explored domain adaptation for privacy-sensitive modalities to learn models for the Doppler sensor data using the IMU data as a source domain. While these cross-modality approaches have inspired our work, they are limited to specific kinds of activities, as the target sensing modality can only sense motion. An approach that can detect a wider swath of activities would most probably require multimodal sensing (*e.g.*, [56]), but with a wider set of sensors generating synthetic training data becomes harder. Thus, there is a need for an approach that can train a target sensing modality even if a direct physical relationship does not exist between the source and target sensing domains.

2.3 Self-Supervised Approaches to Reduce Labeling Effort

Researchers have also explored self-supervised learning (SSL) methods to minimize reliance on labeled data (See [26] for a comprehensive survey on SSL methods for temporal and multimodal data). ColloSSL [47], COCOA [27], CPC [41], FedCLR [81] are recent techniques to employ SSL for the HAR domain. The two works closest to us in terms of their vision are ColloSSL [47], which focuses on SSL across a time-synchronous multi-device setting, and FedCLR [81], which uses a combination of active learning and label propagation to provide pseudo labels to large amounts of unlabeled data. These approaches utilize deep learning architectures for label propagation and representation learning in a multi-device setting. However, these approaches target single sensing modalities for SSL or assume a correlation across different sensing modalities.VAX used an alternative approach in contrast to self-supervised learning, where unlabeled data is processed to obtain useful representations to get pseudo labels. In VAX, we deploy rich (but privacy-invasive) sensing modalities like video and audio sensing with other privacy-sensitive modalities for a brief amount of time and learn (pseudo) labels from global models built on these (audio and video) sensors. Our approach generates labels independent of what privacy-sensitive sensing is being used. Thus, we do not need to make assumptions about or restrict the selection of privacy-sensitive modalities.



(a) Movement and position information using Doppler and Lidar sensors over time.

(b) Thermal data (10x8) and Sound Localization information at a single timestamp (snapshot).

Fig. 1. Visualizing information across various activities from various privacy-sensitive sensors. (a) shows the movement range and radial velocity of moving objects using a Doppler sensor and object distance and angle using a Lidar sensor in the sensor plane across a sequence of activities happening in the living room. (b) shows snapshots of Thermal and Sound Localization at a given timestamp for various kitchen activities.

3 VAX HARDWARE DESIGN

In this section, we describe the design of our VAX hardware apparatus which includes both audio/video sensors as well as a selection of privacy-sensitive sensors. For the audio and video sensing modalities, we used a Yeti USB microphone and a Logitech USB web camera with a 720p resolution. In selecting the suite of privacy-sensitive sensors, we looked at prior works in HAR [43, 49, 51, 57, 84, 98] and considered aspects such as sensing range, a breath of signals captured, fidelity and sensitivity of the sensed signals. We describe below the different sensing modalities that are present on our VAX apparatus and our rationale for including them in our testing.

- **Movement sensing using Doppler radars:** mmWave Doppler radars are promising for HAR, especially now as they become more affordable [24, 34, 59, 67, 72, 106, 107]. These radar sensors emit a reference RF signal, and measure the reflected signal back from objects in the vicinity. Whenever there is movement, such as by an object or by a person doing some activities, the reflected signals are Doppler-shifted, which can be used to create a 1-D Doppler plot. With frequency modulation (FMCW), a 2-D plot of range *vs.* the Doppler plot can be produced to characterize the type of motion at different ranges (See Figure 1).
- **Position sensing using a Lidar:** Lidar (light detection and ranging) uses the Time of Flight (ToF) or parallax of a laser beam to perform range finding. Most Lidar sensors are designed to sense range across one axis. However, electromechanical (spinning on a single axis) Lidars can detect a range of objects in 360 degrees. Lidar is common in robotics to detect and avoid obstacles and on mobile embedded systems for

guidance. However, Lidar can be repurposed to detect activities with positional relevance (i.e., sitting on the sofa could imply relaxing or sitting at the dining table could imply having a meal). SurfaceSight [55], for example, utilizes Lidar to enable object detection and user interactions. Figure 1 shows an example of how Lidar signal can vary based on different human activities.
- **Thermal sensing (Infrared Sensors):** Thermal/infrared cameras are widely used for low light scenarios since they can detect IR (serves as a heat signature) from all kinds of objects. Various versions exist to support numerous applications such as in-home security, personal safety, energy efficiency, pest control, home maintenance, and automotive care. However, with sufficiently high resolution, thermal sensors can also raise privacy concerns even if the data is not personally identifiable [86]. In VAX, we used a low resolution (10x8) thermal sensor, which does not reveal any Personally Identifiable Information (PII), but still can provide signals on activities with a thermal signature of appliances such as a fridge or a stove (see Figure 1).
- **Sound localization (Micarrays):** Unlike full fidelity audio, being able to detect the direction from which sounds are coming (i.e. localize the sound) and their intensity (i.e. dB level) is much more privacy-preserving while being helpful to detect activities around appliances. This can be done using an array of microphones spatially separated on a Printed Circuit Board (PCB), and measuring the difference in arrival rate to infer the 3-D location of sound sources (see Figure 1). However, even the mere presence of a microphone, even if it is only inferring the location and the intensity of sound may still raise privacy concerns in case it is hacked to capture full-fidelity audio. We believe that in the future, sensors that are guaranteed to be limited to certain functionality in hardware are going to be possible [100] and hence we have included a commercial off-the-shelf sound localization sensor called the MicArray [105] on our current VAX rig.
- **Sensing ambient vibrations:** Inertial Measurement Units (IMUs), which includes an accelerometer, a gyroscope, and a magnetometer are often used for HAR on wearable devices [60, 71] or even in the ambient environment such as a wall socket [56]. We include an IMU on our VAX hardware to similarly sense ambient vibrations, coupled through the wall and other structural elements when performing different activities such as exercising or operating appliances such as a vacuum cleaner or washer/dryer. Other people have also looked at finer-grained activities of daily living through vibrational sensing [76].
- **Environmental sensors (Temperature, Humidity, Pressure, Light, Color, RSSI, Passive Infra-Red):** Some common human activities result in changes to the environment that can be picked up by ambient sensors such as temperature (e.g. using a space heater) or humidity (activities in the bathroom) or a light sensor (watching TV, turning on lights) [28, 56]. As a result, we utilize the Mites.io [13, 68] multi-modal sensor hardware which incorporates twelve different sensing dimensions. Notably, while the Mites device does provide highly featurized data from an on-board microphone, we do not use that data in VAX.

Figure 2 shows our current VAX hardware prototype. While all of our chosen privacy-sensitive sensors can be integrated onto a single board, we chose to build our prototype using off-the-shelf hardware modules and sensor evaluation boards. We limit ourselves to ambient sensors to reduce the user burden of using any smart (or sensor instrumented) devices or wearable sensors. However, utilizing additional sensing modalities can bring interesting opportunities and challenges which is beyond the scope of this paper. We used a Logitech USB Webcam (C910, 720p resolution) as well as an iPhone (as a backup data source) for both audio and video data. We use the AWR1642BOOST-ODS [44], a 77GHz mmWave evaluation board with an integrated DSP and an ARM Cortex-R4 Processor from Texas Instruments. This evaluation board incorporates a patch-antenna structure that provides a wide field of view (FOV) in both azimuth and elevation planes. The integrated microcontroller on this board is, however able to provide data at a relatively low frequency of 5Hz, which we overcome by adding a separate DCA1000 FPGA board [45]. This board connects directly to the ADCs of the radar board and extracts the samples of the doppler signals generated in a binary format. Samples collected with the DCA 1000 EVM board are sent via

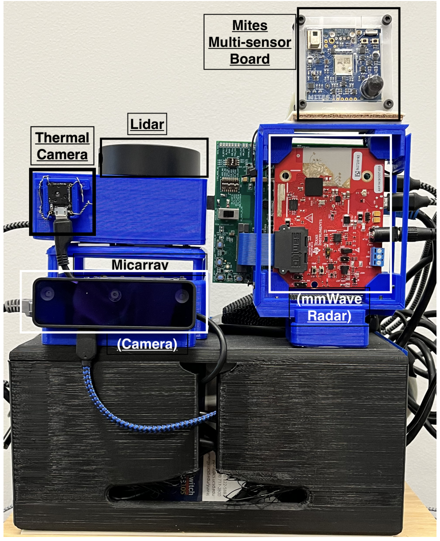


Fig. 2. VAX hardware apparatus. Doppler, Thermal, and the Mites are fixed vertically, and Lidar and Micarray are kept horizontally. The setup is mounted on a box with a combined USB and Ethernet switch enabling data transfer.

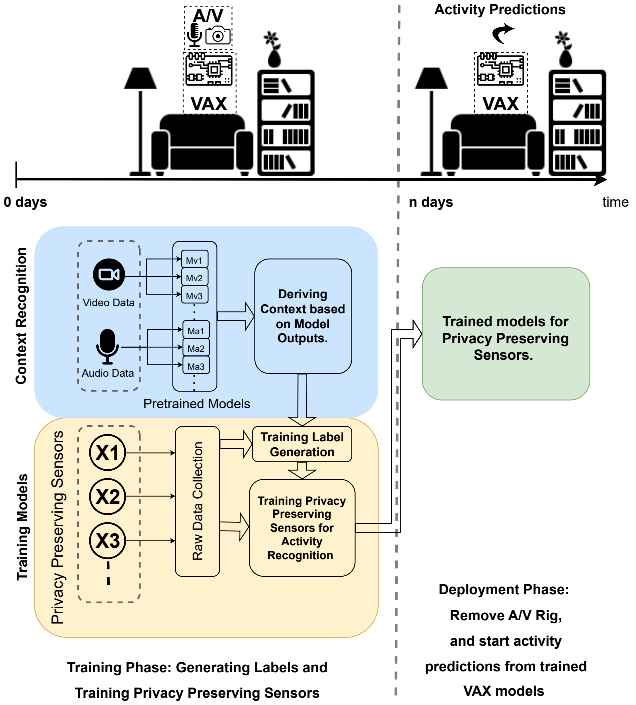

A/V
VAX
Activity Predictions
VAX
0 days
n days
time
Context Recognition
Video Data
Audio Data
Mv1
Mv2
Mv3
Ma1
Ma2
Ma3
Deriving Context based on Model Outputs.
Pretrained Models
Trained models for Privacy Preserving Sensors.
Training Models
Privacy Preserving Sensors
X1
X2
X3
Raw Data Collection
Training Label Generation
Training Privacy Preserving Sensors for Activity Recognition
Deployment Phase: Remove A/V Rig, and start activity predictions from trained VAX models
Training Phase: Generating Labels and Training Privacy Preserving Sensors

Fig. 3. High-level architecture of our VAX training and inference engine, highlighting the key components. VAX recognizes environmental context during the training phase using a set of off-the-shelf audio/video models and generates labels to train privacy-preserving sensors for recognizing activities. A/V sensors are removed and the trained VAX models for privacy-preserving sensors are deployed in the environment.

UDP to a small form factor PC (Intel NUC). This raw data is further processed to create a range-doppler heatmap over time with a sampling rate of 15 Hz. For thermal sensing, we use a FLIR Lepton 3.5 thermal camera with FLIR Systems 3.2a radiometry [3], which is interfaced with a Purethermal Mini Development board to enable communication via USB. This sensor captures high-fidelity thermal data (160x120 pixels) at 8Hz. We observed that the high spatial resolution of this sensor can visibly identify and differentiate between two users based on their physical features (i.e., height, weight, facial features, etc.). To alleviate such privacy concerns, we reduced the resolution of the thermal sensor by 16x to just 10x8 pixels. We attempted to use the Panasonic Grid-Eye sensor, which provides an 8x8 thermal resolution, but the data quality and range were not great so we used the FLIR sensor instead, and down-sampled it as a proxy of a very low-resolution thermal sensor. We use the Slamtec RPLIDAR A1M8 [50], a mechanical LiDAR to capture object distance in a single horizontal plane. It is a two-dimensional Lidar sensor with a maximum range of 6m. The horizontal scanning range is 0°-360°. The scanning frequency is 5.5 Hz when sampling 360 points on each revolution with an angular resolution of 1°. We also tried to use a 3D-Lidar to capture object distances; however, we chose not to use it due to the significant latency for it to provide stable readings and intermittent data during periods of high movement. For sound localization, we use a ReSpeaker 4-Mic Array with a Raspberry Pi [105], which is a quad-microphone expansion board for Raspberry Pi designed for AI and voice applications. We only capture output for sound localization and intensity at very low resolution (1Hz) from this sensor. Finally, we use a multi-modal sensor board, the Mites [13, 68], to sense vibrations (using its IMU) and a range of environmental data (i.e. temperature, humidity, pressure, light intensity, and color, movement based on PIR, low-resolution IR). Our VAX rig is wall-powered, and all the sensors are either connected over USB or Ethernet with a small form factor PC (Intel NUC, 8-core, 16 GB RAM) to collect and store processed data.

4 VAX MACHINE LEARNING AND SOFTWARE SETUP

In Figure 3, we show the high-level software architecture for VAX. We have separated the ML components into the training and deployment phases. In our vision, for the initial training phase, we deploy a camera with a microphone along with privacy-sensitive sensors in a user's home. Our Audio-Video (A/V) pipeline uses off-the-shelf pre-trained ML models to detect activities and generates training labels. Then, our privacy-sensitive pipeline uses these generated labels from the A/V pipeline to train ML models for privacy-sensitive sensors. Subsequently, during the deployment phase, the camera is physically removed and the A/V is not used anymore, and the trained models on the privacy-sensitive sensors are used for activity detection. We now go into the details for each of these phases.

4.1 Deployment and Data Collection

We deployed our VAX prototype and collected data for participants in their own homes. For each home, we deployed our prototype in three locations (kitchen, bathroom, and living room) to capture the diversity of activities. We started with a union set of 30+ activities from recently published activity recognition literature [54, 56, 71]. We filtered activities from these sets based on two criteria, (i) we removed activities that do not belong to the three location contexts we have considered, i.e., activities like drilling, using screwdrivers, etc., and (ii) we removed activities that do not have any movement or audio signature, like sitting, or sleeping. We intentionally kept a wide range of high-level activities (i.e., Cooking, Exercising, etc.) to allow for diverse movement and sound patterns. We finalized a list of 17 activities. For the kitchen, we collected samples for activities such as using the oven (Baking), using a blender/mixer (Blender), heating food in the microwave (Microwave), using a fridge (FridgeOpen), chopping and grating vegetables/fruits (Chopping+Grating), cooking a meal on the stove (CookingOnStove), washing dishes in the sink (WashingDishes). For the bathroom, we collected data on activities like using a hair brush (HairBrush), using an electric hair dryer (HairDryer), washing hands (HandWash), using an electric shaver (ShaverInUse), and flushing the toilet (ToiletFlushing). Finally, we collected samples for activities that can happen anywhere in a home, such as eating or drinking (Eating/Drinking), exercising (Exercising), coughing (Coughing), and operating a vacuum cleaner (Vacuum). We asked the participant to perform these activities in their living room. All activities are performed by a single participant in their own home. We collected data from 10 participants, and the sample size is based on a review of similar participant studies in HAR literature [12, 25]. Each study took around 4-5 hours, including the time to carry the VAX rig to different homes and to set it up at multiple locations in the home, which resulted in approximately 10 hours of data collection across all participants. Our study design was approved by our Institutional Review Board (IRB).

4.2 Activity Recognition Using Off-the-shelf A/V Models

In this section, we present a novel approach for discovering good label-activity matches from any set of off-the-shelf A/V models. A straightforward approach to collect activity labels from A/V modalities is to use the best-performing models for each or do a confidence vote across multiple A/V-based ML models. We shortlisted twenty popular off-the-shelf A/V models [11, 29, 31, 35, 36, 38, 40, 61, 62, 65, 71, 89, 90, 93, 95, 96, 99, 101, 103, 104, 108] based on open-source availability of the models, including their pre-trained weights, as well as their performance benchmark against publicly available datasets [19, 42, 92]. However, our empirical evaluation of these off-the-shelf A/V models, used as-is, on the data we collected for our chosen 17 activities of daily living did not provide accurate results. Some of the reasons for this inaccuracy include: (*i*) the models are trained to provide labels for activities that have overlapping signatures (*e.g.*, blenders in a kitchen or tools like a power drill). (*ii*) the models are trained for HAR in different settings and thus do not have a clear one-to-one mapping for common activities of daily living (*e.g.*, atomic actions like sitting can attribute to multiple human activities). Merely combining all the outputs of different models is also non-trivial since it needs to consider that different models may be better suited to certain activities vs. others. Based on our experiments with these twenty A/V models, we make the following observations:

- Skeleton-based action recognition models work consistently across different settings as their input (i.e., human pose) is environment agnostic.
- No single A/V model provides consistent labels across the 17 activities we tested on.
- The same skeleton-based action recognition model trained on different types of large-scale datasets can learn multiple activity signatures and perform well on different categories of activities.
- Many audio-based models perform consistently well to identify a similar set of activities (*e.g.*, using an electric shaver, blender, microwave), i.e., we do not need more than one audio model for our ensemble of A/V models.
- The capability of video models to detect activities and confidence depends on the vantage point (e.g., front or side view) and occlusions.
- The labels provided by existing models may not have a 1-to-1 correspondence with each other or even with our chosen activities, requiring reconciliation. For example, "Exercising" might be recognized by one video model and as "Doing Jumping Jacks" by another model.

For building our A/V pipeline, we finalized 2 off-the-shelf models for activity recognition using human pose estimation, *POSEC3D* [31] and *STGCN* [103] trained on the *NTU60* [88] dataset with 60 activity classes. We further use 3 more copies of *POSEC3D*, each trained on different datasets, i.e , *NTU120* [63], *UCF101* [92] and *HMDB51* [52] with 120, 101 and 51 different activity classes. For the off-the-shelf model for audio-based HAR, we use YAMNet [29], a deep network architecture that predicts 521 audio event classes based on the AudioSet-Youtube [37] corpus.

4.2.1 Building an A/V model from a set of training homes: For our chosen set of off-the-shelf A/V models, we collect the top-k predictions for a set of activities across a set of homes (i.e., reference homes) that we use for training data. Using a shallow neural network classifier, we learn a non-linear mapping for each off-the-shelf A/V model from raw A/V labels to VAX activity labels. The intuition behind this approach is as follows: the top-k predictions from raw A/V labels might not have a static mapping to VAX activity labels, but they are weakly consistent across different home environments. This is true as audio models are agnostic to (home) environments, and video models are trained on large datasets like Youtube-8m [2], which contains observations for the same activity across a wide variety of home environments. We also observe that audio models show a more consistent set of top labels for single VAX activity across environments than video. Thus, we designed two ensembles comprising of different classifiers. The first ensemble (audio-only ensemble) consists of classifiers built from audio models, which works well on various activities with distinct audio signature (i.e., Blender, HairDryer, Vacuum, etc.). The second ensemble (audio-video ensemble) combines classifiers from both audio and video models to detect all the activities (see Figure 4). Combining the two (audio-only and audio-video) works better than using a single audio-video ensemble or two separate ensembles for audio-only and video-only off-the-shelf A/V models. These two ensembles are combined together to create a final VAX A/V model. To predict labels for activity instances in new homes, we run our set of top 5 (*YAMnet*, *STGCN*, *POSEC3D+NTU120*, *POSEC3D+UCF*, and

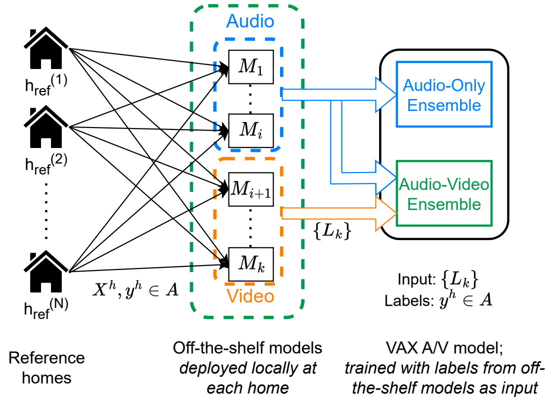


Fig. 4. Building A/V models from data collected from reference homes. A/V data (X^h) collected from reference homes are used as input from off-the-shelf models deployed locally, and outputs raw labels (L_k) from off-the-shelf models with VAX activity labels ($y^h \in A$) are used as input to train global VAX A/V Model.

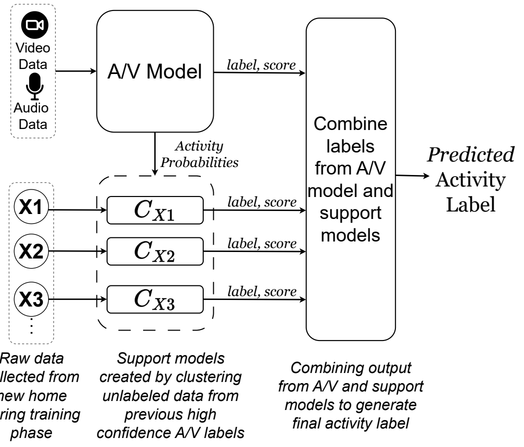

Video Data
Audio Data
A/V Model
label, score
Activity Probabilities
X1
X2
X3
C_{X1}
C_{X2}
C_{X3}
label, score
label, score
label, score
Combine labels from A/V model and support models
Predicted Activity Label
Raw data collected from new home during training phase
Support models created by clustering unlabeled data from previous high confidence A/V labels
Combining output from A/V and support models to generate final activity label

Fig. 5. Generating activity labels in new homes using A/V models and support models from privacy-sensitive sensors.

POSEC3D+HMDB51) off-the-shelf A/V models and collect the top-k labels for all models. These models are then passed through the audio-only and the audio-video ensemble to generate activity predictions, and the activity with the highest confidence value is chosen. To reduce the error in the VAX A/V pipeline, we generate confidence thresholds for each ensemble, and an activity instance is marked undetected if predictions from both ensembles are below these confidence values.

4.2.2 Improving A/V labels with data from the privacy-preserving sensing modalities. One challenge with our A/V pipeline is to set a cutoff threshold that optimizes both accuracy and the detection rate (i.e., the fraction of activities with a confidence score higher than the A/V model cutoff thresholds) for activities. Optimizing for a high-accuracy model with a low detection rate could miss detection on all instances for a particular activity, thus leading to no training labels for the next phase for VAX. In contrast, optimizing for high detection rates at the cost of accuracy would lead to more noisy generated labels from the A/V pipeline, affecting the accuracy of the privacy-sensitive sensor pipeline. To alleviate the issue of scarcity of good labels, we opportunistically utilize information from raw data collected from privacy-sensitive 'X' sensing modalities to increase the detection rate for activities. For example, (i) activities like FridgeOpen and Baking (or using an oven) will have clear separability in thermal signatures (see Figure 1), (ii) activities like Walking and Drinking/Eating will have a clear separability in their Doppler signatures. We can find similar pairs (or sets) of activities that provide a clean separability of activity signatures for a given sensing modality. We can leverage this information to increase the confidence score for an instance of an activity that did not receive any label from the A/V pipeline but *looks* very similar to another labeled instance. However, this would not work for activities that do get classified even once with high confidence from the A/V pipeline. For each 'X' privacy-sensitive modality, we identify activities, which are separable in that specific modalities feature space based on cluster membership of A/V predictions, i.e., if two or more high-confidence activity labels from A/V fall in the same cluster. Finally, we use these clustering models to enhance confidence scores for highly separable activities.

Figure 6 and 7 shows a 2-D embedding of clusters created using data from the Lidar and Thermal sensors, respectively. The different colored dark dots represent points clustered together into different clusters without

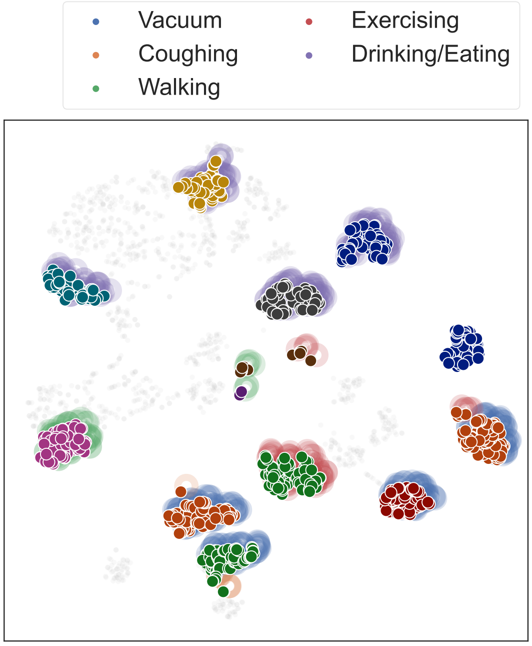


Fig. 6. tSNE embedding (2-D) for supporting clusters using Lidar data to improve A/V pipeline labels.

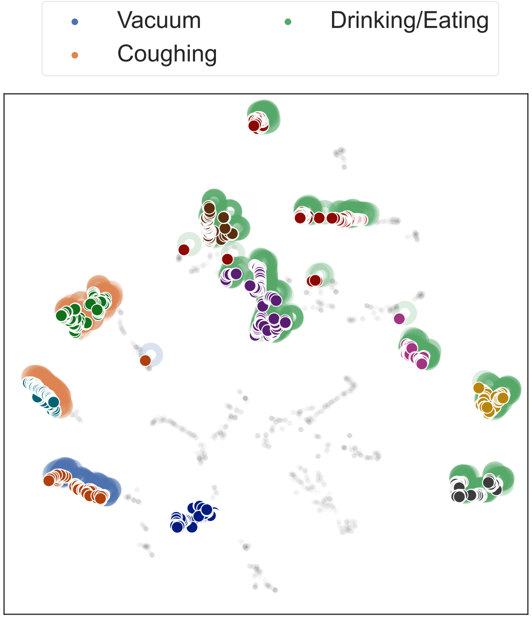


Fig. 7. tSNE embedding (2-D) for supporting clusters using Thermal data to improve A/V pipeline labels.

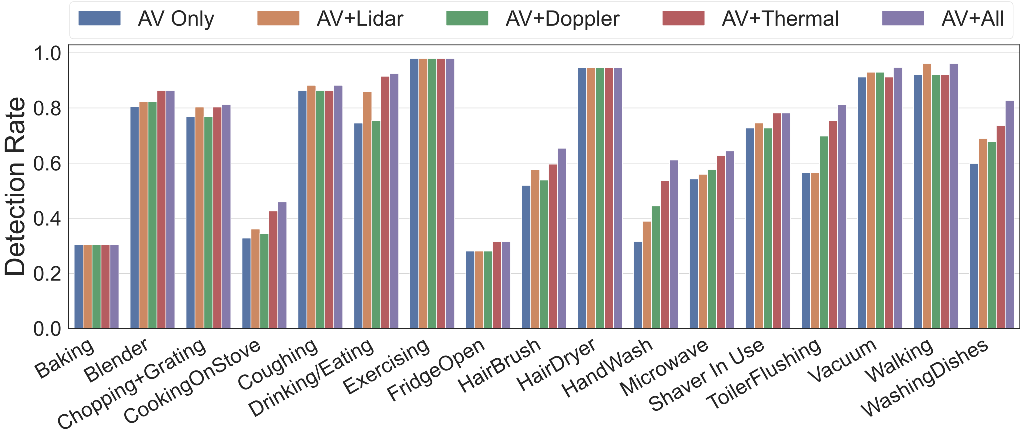


Fig. 8. Increasing the detection rate of activity labels using support models from Lidar, Thermal, Doppler, and when using all three (Lidar+Thermal+Doppler) sensors.

any activity information, highlighted areas that correspond to dense regions for data from different activities, and grey points represent non-clustered points (or outliers). We observe that some activities like Walking (green shade region), Drinking/Eating (purple shaded regions) are clearly separable with Lidar data, whereas activities like Coughing (red shaded region) and Vacuum (blue shaded region) are closer to each other, and even confusing among the single cluster. We can see similar patterns in thermal data (i.e., good separation for Coughing and Vacuum and Drinking/Eating). This figure illustrates that Lidar data can help the VAX A/V pipeline to enhance the detection rate for low-confidence predictions for these activities. Figure 8 summarizes how support models for particular privacy-preserving sensors can help increase detection rates for different activities. Activities like Drinking/Eating and WashingDishes see a high detection rate by combining all the sensors (i.e., "AV+All"). Some activities like Exercising and HairDryer show no improvement by using any of the privacy-sensitive sensors. One reason for this is that none of the prominent clusters for these activities have instances that are marked 'Undetected' by the A/V pipeline.

4.3 Training Models for the Privacy Sensitive Sensors

For training models for privacy-sensitive 'X' sensors, the first challenge is making sure that there is training data for all activity classes.

4.3.1 Getting Labels for Undetected Activities using a Human-in-the-loop. At the end of the training phase, users can disconnect the A/V sensors and not use the A/V pipeline. At this point, the VAX pipeline will start to predict activities based on the models trained on the privacy-sensitive sensors only. However, a subset of activities might not be detected with high confidence using the A/V pipeline. These activities might also vary from one home to another. To improve the end-to-end accuracy of the VAX pipeline, we ask for limited human input for only these undetected activities in this phase. We solicit this input by asking users to perform *a single instance* of those undetected activity classes. As we show in Section 5, using only a single training instance for an average of 2 undetected activities (out of a total of 17 activities) for each home VAX can give an accuracy of 84%, which is better than a baseline approach where users provide one label for each of the 17 activities (accuracy of 79%), and with significantly less user burden.

4.3.2 Featurization of raw data from privacy-sensitive modalities: We train an ensemble of sensor-level models instead of a monolithic model that combines all sensing modalities. There is an opportunity cost of losing on added advantage of fusing multiple sensors; however, we prioritize reliability in long-term deployment. Such an approach would also allow for utilizing existing models (*e.g.,* Vid2Doppler for Doppler RADAR [5]). Another benefit of using an ensemble of sensor-level models is the ease of featurization as different modalities have different kinds of featurization techniques, which are well-studied in prior literature [5, 55, 56]. For Thermal data, we use a rolling window for 5 contiguous images (8x10 pixels each) and use the maximum temperature value in ° Celsius for each pixel, respectively. For the Doppler data, we use a rolling window for 20 contiguous Range-Doppler (256x32) heatmaps and break this down into two parts, (i) range-bins (i.e., distance from the sensor) of primary motion activity over time, (ii) average and variation (standard deviation) for different velocity bins over time, and concatenate the two to create final feature vector. For the Lidar data, we first remove the boundary of the room from the data, and then mark the distance of the closest object for all angles (0-360), and then take the maximum value over a rolling window of 10 contiguous samples. For the Micarray, we use a similar featurization technique as Lidar, with distance from a sound source weighted by the amplitude of the sound. As Micarray data is sampled at 1 Hz, we do not aggregate across multiple samples over time. For the other sensors (i.e., PIR, IMU, Environmental Sensors), we directly use featurized data (over 100ms) from multi-modal sensor board, mites [68].

4.3.3 Handling class imbalances in training labels: We use a minority oversampling approach called SMOTE [22] for data augmentation across activities, which removes a large number of instances. SMOTE (Synthetic Minority Oversampling Technique) works by selecting examples that are close to the feature space, drawing a line between the examples in the feature space, and drawing a new sample at a point along that line. In general, data augmentation leads to lower generalizable accuracy. However, we have a key advantage in our setting. Higher anomaly scores for labeling instances for a given activity lead to higher class imbalance but also suggest easier separability due to denser data distribution for normal samples. Thus, the classes that are clearly separable in a sensing modality benefit from outlier removal without suffering from errors due to sampling imbalance.

4.3.4 Reducing noise in training labels: To reduce noisy training labels, We use a confidence learning approach built on top of another approach proposed by Northcutt *et al.* [74], which addresses the problem of identifying noisy data in a general setting where no annotation information is available except the observed noisy labels. We start by calculating out-of-sample probabilities $P_{ik}^{(x)}$ for all data by training naive classifiers $\theta^{(x)}$ over featurized data across all 'X' sensors, respectively. For each sensor x, we calculate a confidence joint matrix, $C_{ij}^{(x)}$, to partition and count labels across multiple classes. Diagonal entries of $C_{ij}^{(x)}$ count correct labels, and non-diagonals capture asymmetric label error counts (As an example, $C_{i=3,j=1}^{(x)} = 10$ is read, as per sensor x, "Ten examples are labeled 3 but should be labeled 1."). This is further normalized to $Q_{ij}^{(x)}$ to make sure each sample row for $C_{ij}^{(x)}$ sums to the marginal probability of activity from $\theta^{(x)}$, and probability across all activities sums to 1. Finally, we classify samples with the lowest confidence calculate activity level threshold ($\delta_i^{(x)}$, i.e., any sample predicted as activity label i would be classified as noise if $P_{ii}^{(x)} < \delta_i^{(x)}$) for label noise characterization, and remove samples which are classified as noise across the majority of sensors. One critical hyperparameter for this approach is the selection of naive classifier $\theta^{(x)}$. We selected an ensemble of KNN and SVM with Gaussian kernel, which are shown to be the most robust classification methods for training with imperfect labels in prior literature [17], and are empirically shown to work best on our training data. Finally, we train the same classifier on clean data to provide the final trained model for each sensor, and for final activity predictions, we return the activity label for the highest prediction probability aggregated across all sensors.

4.4 Putting it together into an end-to-end VAX pipeline

Our entire pipeline for VAX is written in Python with over 10k lines of code and consists of three components, (i) training A/V model from reference homes, (ii) generating activity labels using trained A/V model and support models from 'X' sensors and (iii) training 'X' sensors with labels from A/V model. For training across all components, we use module implementation for a variety of machine learning methods (i.e., KNN, SVM, OPTICS, and SMOTE) from scikit-learn and imblearn libraries. Our noise reduction algorithm is built on top of Cleanlab [74], a generic package for identifying and reducing noise in datasets. We open-sourced our entire VAX system [80], which can be used out of the box with label predictions from any off-the-shelf A/V and can be extended to add more variety of 'X' sensors with no implementation changes.

5 EVALUATION AND RESULTS

We evaluated VAX to answer the following questions.

(1) How well does our A/V pipeline detect activities with a model trained on different sets of reference homes and predicts activity labels along with support models in a new home?
(2) How well does the VAX pipeline do in two scenarios?
  (a) Using activity labels detected from the A/V pipeline with no user input.
  (b) Using activity labels detected from the A/V pipeline along with one sample from user input on activities that are not detected by the A/V pipeline.
(3) How does VAX compare with a baseline approach in terms of accuracy and user burden?

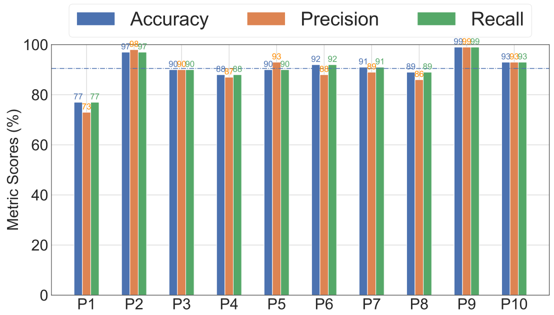


Fig. 9. VAX A/V pipeline: Accuracy, Precision, and Recall for activity recognition for all participants, on detected activities (75% of all instances). In general, our A/V pipeline performs well across all homes with an average Accuracy of 90.5%, Precision of 89.5%, and Recall of 90.5%. The A/V model for all participants is trained in a leave-one-home-out fashion.

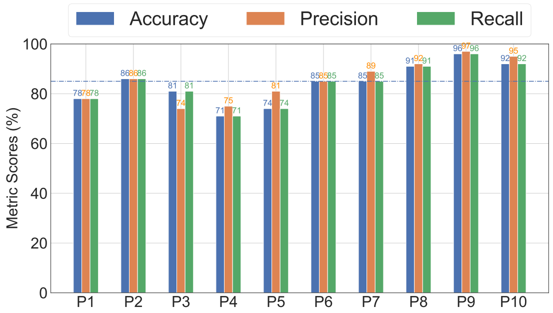


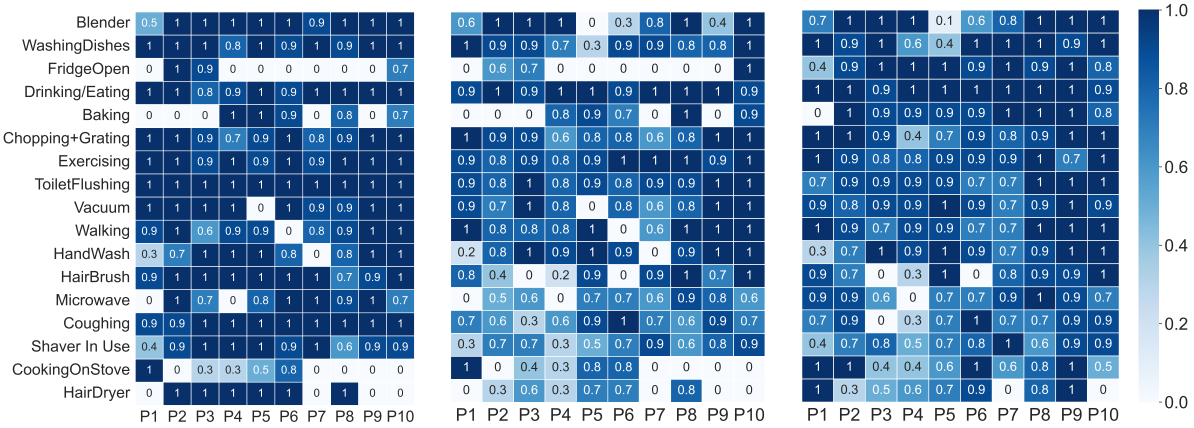
Fig. 10. Privacy sensitive pipeline: Accuracy, Precision, and Recall for activity recognition (across 17 activities) for all participants. We observe an average Accuracy of 85%, Precision of 85%, and Recall of 85% and that these remain consistent. The variation in Accuracy, Precision, and Recall across users is due to different amounts of noise (incorrect predictions with high confidence) in A/V labels.

5.1 Performance of A/V pipeline

Figure 9 shows the performance of our A/V pipeline (trained in leave one home out fashion) across participants for activities detected with our A/V pipeline (i.e., confidence score higher than cutoff values, 60% for audio-only ensemble and 40% for audio-video ensemble). We observed that 75% of the activities follow these criteria and thus are labeled using the A/V pipeline. Across these detected activities, our A/V pipeline shows high accuracy, precision, and recall. The information captured from the A/V modality is also consistent across multiple homes, *i.e.,* an A/V model trained on a set of reference homes can capture activity patterns in new homes with support from privacy-sensitive sensors. Figure 12 shows the confusion matrix across activities predicted from the A/V pipeline and ground truth. We observe that A/V is good at identifying most activities with an observable sound signature, such as WashingDishes, Blender, *etc.* Our pipeline also does well for activities like HairBrush, or Exercising, which have unique motion signatures compared to other activities in the set. Our A/V pipeline has the highest confusion among activities like Baking and FridgeOpen. While both these activities involve large motion, they involve opening and closing doors, and the video-based ML models often confuse such activities. Apart from these two activities, any activity involving very little motion and insufficient sound, such as CookingOnStove, was hard to detect using audio or video-based models.

In Figure 14, we evaluate the impact of the number of reference homes on the accuracy and detection rate of activities from the A/V Model without any support from 'X' sensors. We observe that as the number of reference homes increases, the count of detected activities increases monotonically. We also observe that even with a low detection rate, the A/V pipeline has good accuracy, which shows that the A/V pipeline is reliable in detecting activities even with a smaller count of reference homes. This result means that when the labels generated by the A/V models are used to train other sensors, not all user actions will be labeled, and training might be slow. More importantly, given the precision of the A/V models is high, the labels will be accurate.

Figure 11(a) shows the F1-Score for our A/V pipeline across activities performed by all ten participants. We observe that our A/V pipeline can classify some activities (like Blender, WashingDishes, Exercising, etc.) across almost all homes. We also observe that across all homes, we are missing true detection for at least one activity (*e.g.*, P3) and, at most, five activities (*e.g.*, P7), and around two activities on average across all homes.



(a) A/V pipeline.

(b) VAX pipeline (A/V Labels only).

(c) VAX pipeline (Single user input on undetected activities).

Fig. 11. Performance (F1-Score) for activity detection across participants for A/V pipeline, VAX pipeline with no input from the user (A/V labels only), and VAX pipeline with one input each for undetected classes in A/V. For the A/V pipeline, most activities are detected across all homes. Some activities like Baking, FridgeOpen, and CookingOnStove are not detected for most homes, as high confusion between these activities leads to low confidence for A/V models, thus going undetected. VAX pipeline (A/V labels) performs well for well-detected activities in A/V and does not detect activity not labeled with the A/V pipeline. VAX pipeline (single user input on undetected activities) performs well across all activities showing 'X' sensors capability for classification is limited only by the presence of ground truth labels.

5.2 Performance of VAX pipeline

Once we get labels from the A/V pipeline, we train our privacy-sensitive sensors with the 'X' pipeline in leave-one-instance-out cross-validation for a given participant. Figure 11(b) shows the performance of VAX when trained on labels provided by the A/V pipeline and no user input. We see no detection for activities that the A/V pipeline could not detect accurately with VAX. Thus, we anticipate that when deployed in people's homes, VAX will identify the undetected activities by looking at activities not present in A/V inferences and asking the user to provide one labeled instance for those undetected activities. Figure 11(c) shows F1-Score across all activities for all participants once we have one labeled instance (either from the A/V pipeline or a simulated human input). We observe that the VAX pipeline can reliably classify activities with only one sample for undetected classes. We also see that F1-score for activities included in the A/V pipeline improves. This happens due to the reduction in noise in overall training data with more ground truth labels.

Figure 10 shows an accuracy of 85%, precision of 85%, and recall of 85% for VAX with one label for each class (either automatically received from A/V models or simulated). This result highlights VAX's capability of fusion of 'X' sensors to classify various activities successfully. Figure 13 shows the confusion matrix for activity detection from privacy-sensitive sensors. We observe that detection across almost all activities is good, showing a clean differentiating signature in one or more 'X' sensors. We observe that 'X' sensors do not perform well for some activities such as HairBrush, HairDryer, and Shaver In Use. These are all bathroom activities performed at roughly the same location. Given Doppler RADAR is sensitive to small motions, with more training samples, the models should be able to detect these activities. However, confirmation of this hypothesis remains a future work. Figure 15 shows the F1-score for activity detection across different sensors. We observe that for most of

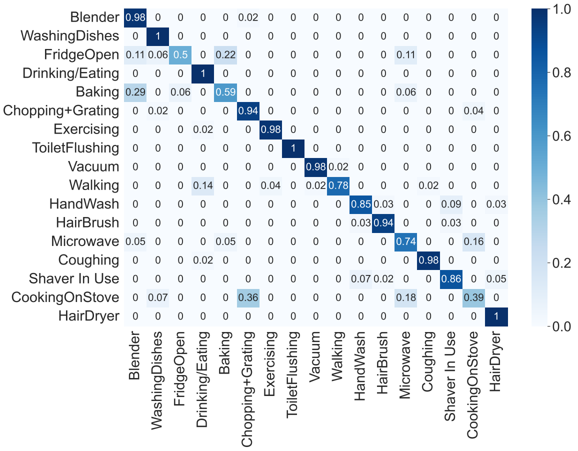


Fig. 12. A/V Pipeline: Activity Confusion Matrix averaged across all participants. A/V pipeline performs well for most activities that generate descriptive sounds like Washing Dishes, Blender, HandWash, Toilet Flushing *etc.* The majority of confusion happens due when there are similar movement (or pose) patterns across multiple activities like FridgeOpen classified as Baking (Opening and closing doors action), or Coughing confused with Drinking/Eating, or Chopping-Grating classified as CookingOnStove (hand motions in close spaces).

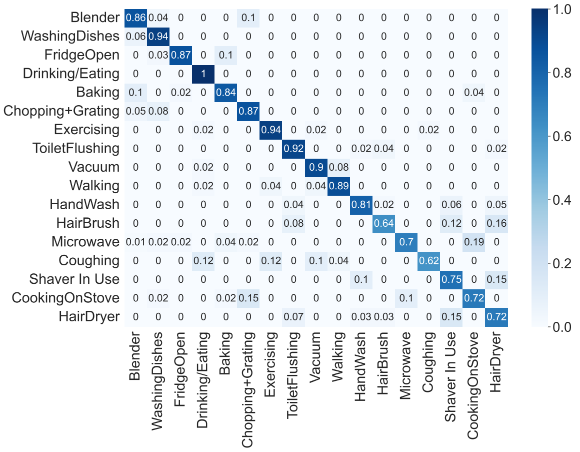


Fig. 13. Privacy sensitive pipeline: Activity Confusion Matrix averaged across all participants. Most activities are detected well due to a clean differentiating signature in one or more 'X' sensors. The worst performance is across activities like HairBrush, Coughing, HairDryer, etc., where human movement is limited, and their positional difference is captured by Thermal, Lidar, or Micarray sensors, thus relying completely upon small variations in hand/torso movement from Doppler sensor.

the activities, Thermal, Doppler, and Lidar sensors show a good F1-score. This points out that in the future, even one of these sensors might give sufficient accuracy across all activities and be more practical in terms of cost. Other sensors like Micarray show good F-1 scores across a subset of activities like HairBrush, Coughing, Blender, WashingDishes, Drinking/Eating, Vacuum etc., all of which have a distinctive sound signature. We also observe that other low-fidelity sensors, including environmental sensor or PIR motion, does not contribute to any of the activities and thus can be removed or replaced with other sensors in future works.

5.3 Comparing VAX with other Baseline Approaches

Figure 16 shows a head-to-head comparison of activity detection for the VAX pipeline with a few baseline approaches. We consider two different baseline approaches: (i) pre-trained models on the privacy-sensitive 'X' sensors using data from reference homes and run prediction on a new home in a leave-one-home-out fashion, i.e., X-Only (pre-trained) and (ii) trained a model on the in-home using data from privacy-sensitive 'X' sensors by asking users to perform the activities and collect labels, i.e., X-Only (no pre-training). We compare these approaches with VAX in two different settings.

a) No user input is provided (Figure 16(a)): There are no labels to train X-Only (no pre-training) model when user input is not provided. In comparison to X-Only (pre-trained), VAX performs considerably better across all activities (74% *vs.* 38%). Furthermore, when we compare activities that are detected with our A/V pipeline (15 out of 17 on average), the performance of VAX increases from 74% to 84%, whereas there is a

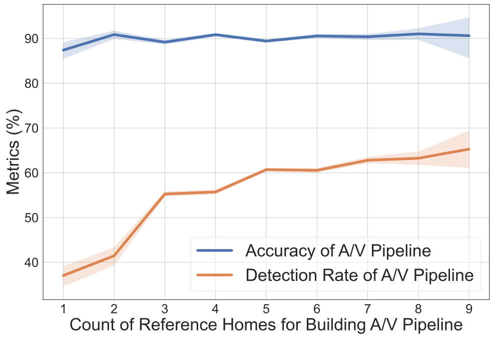


Fig. 14. A/V Pipeline: Accuracy and Detection Rate for activities with the count of reference homes. Activity detection accuracy (i.e., correct prediction out of detected activities) remains high across all homes, and Detection Rate (i.e., Number of activities detected with high confidence) increases with the number of homes.

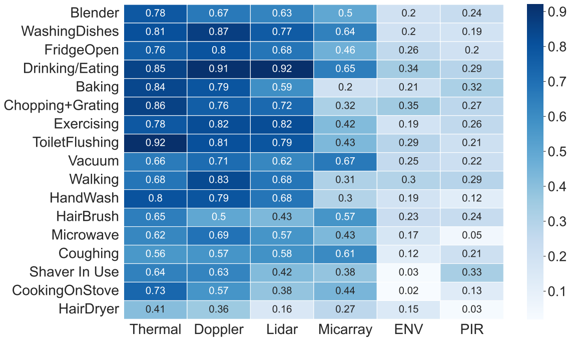
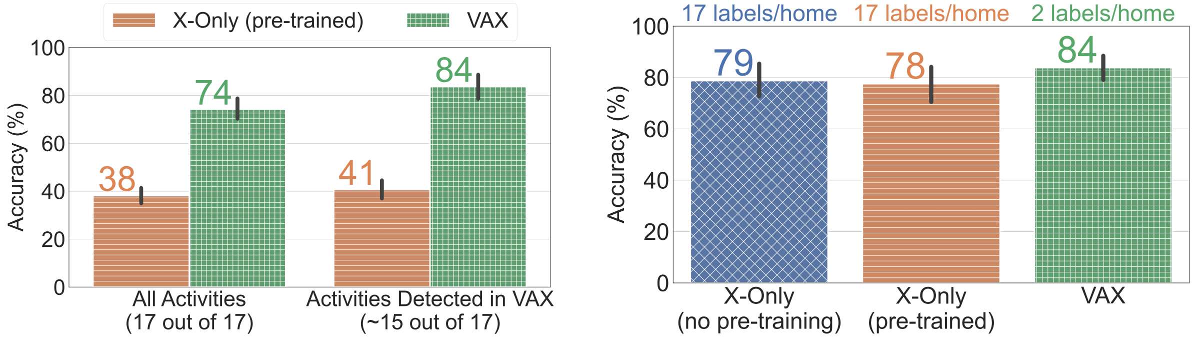


Fig. 15. Privacy sensitive pipeline: F1-Score at activity level from individual sensors. Doppler, Thermal, and Lidar show the highest f1-score across most activities. Some activities (like Coughing, HairDryer, etc.) do not show a good f1-score across any sensing modality, even if accurate labels from the A/V pipeline show a lack of separability across all sensing modalities.



(a) No User Input.

(b) With User Input.

Fig. 16. Comparing VAX and the two baseline approaches. Without any user input (left), VAX performs considerably better than baseline approach (i.e., models trained on privacy-sensitive data from reference homes) on activities detected in VAX (84% *vs.* 41%) as well as across all activities (74% *vs.* 38%). When user inputs are provided (right), VAX achieves marginally better performance with considerably less labeling effort from the user (2-3 labels/home *vs.* 17 labels per home) when compared to baseline approaches (i.e., models trained on privacy-sensitive data using input from users (X-Only (No pre-training)) and pre-trained models fine-tuned using input from users (X-Only (pre-trained))).

marginal improvement for X-Only (pre-trained) from 38% to 41%. This shows that the performance of VAX is primarily limited by the ability of the A/V pipeline to detect (at least one) sample from activities.

b) User input is provided (Figure 16(b)): For the X-Only (no pre-training) case, we provided one label per activity, i.e., 17 labels/home to train models. For the X-Only (pre-trained), we start with the pre-trained models using data from the reference homes and then fine-tune the models with one label per activity in the test home, i.e., 17 labels/home. For VAX, we only provided one sample for the activities undetected by the A/V pipeline (2 labels/home on average). Even with considerably less labeling effort for VAX (2 labels/home *vs.* 17 labels/home), VAX provides better performance (accuracy of 84%) than both baseline approaches (accuracy of 79% and 78%). We also observe that VAX's accuracy on all activities with user input on activities undetected with the A/V pipeline is similar to VAX's accuracy without any user input on activities detected with the A/V pipeline, thus showing that VAX's performance is limited only by the ability of off-the-shelf, pre-trained A/V models to separate chosen activities.

6 DISCUSSION

This section discusses the limitations of VAX and opportunities for future enhancements.

6.1 VAX hardware: modular design to disaggregate sensors

Our results indicate a limitation of our current VAX prototype. To simplify our deployment task, we integrated all the sensors, including the USB camera/microphone and privacy-sensitive sensors, and the Intel NUC compute node as a single monolithic rig. This rig could then be mounted on a tripod and carried around into different locations in the same home, across homes, and placed in a single location. However, that also meant a single vantage point for all the sensors in each room. Hence, we could not get any labels for some activities in some locations due to occlusion (e.g., the person's back was to the A/V sensor, and furniture was in the way) or just poor angles. This is important since the A/V models are otherwise mostly agnostic to the vantage point, and if they are unable to provide labels, we are unable to train models for the privacy-sensitive sensors at all. In the future, making VAX modular so that at least the Audio/Video sensors can be placed at different locations could be beneficial to detecting all activities. Similarly, some privacy-sensitive sensors are directional (e.g., the Thermal sensor, the PIR movement sensor) with a field of view, while others are more sensitive depending on the location (e.g., the IMU to measure vibrations). Thus potentially having more than one of the privacy-sensitive sensors in a room at different locations could help boost accuracy. Furthermore, we can opportunistically place multiple sensors of the same sensing modality in different environments and train them together with a portable A/V sensing solution (i.e., using your phone to capture A/V labels) to extend activity recognition beyond a single location to the entire house with a single VAX system.

6.2 Real world in-the-wild study

Our current results, which show an average accuracy of 84% without any user input on a smaller set of activities and with around two user labels per home (for harder-to-detect activities) across all activities, showcase the challenge with accurate HAR. Notably, while we were able to achieve this level of accuracy by evaluating VAX in a diverse set of 10 homes, our data collection itself lasted for a few hours in each home where we asked the participant to do these activities while we collected Audio and Video data and the ground truth for each activity instance start/stop times. The data collection itself was a significant endeavor for our team. However, this data collection setup is still not what an in-the-wild deployment of a system such as VAX would look like. We imagine a scenario where the VAX rig is deployed in a person's home for a particular duration (a day or two) while the participant goes about doing activities totally unconstrained. During this time, the models for the privacy-sensitive sensors would be trained and refined, and on the third day, the system would start to predict activities, and the Audio/Video sensors could be unplugged. We leave the exploration of this totally unconstrained in-the-wild deployment to future work.

6.3 Exploring privacy sensitive modalities

Our evaluation (see Figure 15) shows that Doppler, Thermal, and Lidar sensors can capture a wide range of activities with a high accuracy. However, these sensors need a direct field of view for detecting activities, and installing several of these sensors across all home environments might incur a heavy cost to users. One approach to reducing cost is to opportunistically install sensors based on the type of activities detected in a given environment, i.e., for a kitchen setting, installing a Thermal and a Doppler sensor might be sufficient as most activities have a distinctive thermal signature (FridgeOpen, Baking, etc.), or movement patterns (CookingOnStove, Chopping+Grating, etc.). Similarly, installing a Doppler sensor might be sufficient for a bathroom setting as almost all activities have different movement patterns (HairDryer, ShaverInUse, and WashingHands), limited location variability due to short spaces, and low variability in the thermal signature. Another approach is to explore new privacy-sensitive modalities beyond what we used in our paper. For example, IMUs in mobile and wearable devices (smart watches, smart garments, etc.) can capture rich signal variability for various activities involving body movement patterns. Some recent works also used down-sampled audio signals (16KHz to 1KHz) to suppress privacy-sensitive information (i.e., speech) and show that it can still reliably detect a set of audio-based activities [71], whereas other recent works have built custom hardware platforms (i.e., person detection sensor [100]) which uses a camera under the hood but hides the complexity of the ML implementation inside the hardware module, and only exposing privacy-sensitive information. Incorporating these privacy-sensitive modalities with VAX could provide more opportunities for reducing costs and detecting a richer set of activities.

6.4 Applications of VAX beyond activity recognition

The ideas presented in the paper to utilize A/V models to bootstrap privacy-sensitive sensing modalities can be extended beyond human activity recognition. One interesting application area could be robotics, where A/V sensors can train other sensing modalities to provide additional support in noisy conditions when A/V data is unreliable, i.e., navigating self-driving cars safely through bad weather conditions. Another interesting area could be attributing activities to individuals in a multi-person setting (i.e., more than one person using the kitchen) by correlating information from different users' personal devices (i.e., mobile phones and smart devices) with variability in data captured by VAX sensors.

7 CONCLUSION

In this paper, we present VAX, and end to end system that utilizes a set of off-the-shelf audio and video ML models to provide activity labels, for in-situ training of various privacy-sensitive sensors. VAX's A/V pipeline combines the output of these models and provides a non-linear mapping to a consistent set of activities. VAX also proposes a method to train a variety of privacy-sensitive sensors with noisy labels from the A/V pipeline. We deployed and evaluated VAX across participants in 10 homes, performing 17 different activities. Our evaluations show that our A/V pipeline can detect 15 out of 17 activities with no human supervision, Further, we show that with just one user input provided for the undetected activities (on average 2 out of 17) our privacy-preserving sensors can detect all 17 activities with a 85% accuracy. Ultimately, VAX's hybrid approach provides a compelling starting point for bootstrapping even more accurate HAR models in the future.

8 ACKNOWLEDGMENTS

This work was partially supported by NSF Award (SaTC-1801472, CSR-1526237) and the CMU's CyLab Security and Privacy Institute. We gratefully acknowledge a gift by JP Morgan Chase to support research on smart buildings at Carnegie Mellon University. We would like to thank Sudershan Bhoovaraghavan, Han Zhang, and Thomas Pope for initial prototype development and feedback provided through multiple revisions of the VAX hardware. We would like to acknowledge Sudershan Bhoovaraghavan, Han Zhang, Haozhe Zhou, Vimal Mollyn, and Aigerim Massabayeva for their feedback on the paper and system design throughout. We also thank our anonymous reviewers for their constructive feedback on our paper.

REFERENCES

[1] Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. 2019. More than Smart Speakers: Security and Privacy Perceptions of Smart Home Personal Assistants. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. USENIX Association, Santa Clara, CA, 451–466. https://www.usenix.org/conference/soups2019/presentation/abdi

[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8M: A Large-Scale Video Classification Benchmark. arXiv:1609.08675 [cs.CV]

[3] Matheus Gabriel Acorsi, Leandro Maria Gimenez, and Maurício Martello. 2020. Assessing the performance of a low-cost thermal camera in proximal and aerial conditions. *Remote Sensing* 12, 21 (2020), 3591.

[4] Antonio A Aguileta, Ramon F Brena, Oscar Mayora, Erik Molino-Minero-Re, and Luis A Trejo. 2019. Multi-sensor fusion for activity recognition—A survey. *Sensors* 19, 17 (2019), 3808.

[5] Karan Ahuja, Yue Jiang, Mayank Goel, and Chris Harrison. 2021. Vid2Doppler: Synthesizing Doppler Radar Data from Videos for Training Privacy-Preserving Activity Recognition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 292, 10 pages. https://doi.org/10.1145/3411764.3445138

[6] Reed Albergotti. 2019. How Nest, designed to keep intruders out of people's homes, effectively allowed hackers to get in, researchers claim. https://www.washingtonpost.com/technology/2019/04/23/how-nest-designed-keep-intruders-out-peoples-homes-effectively-allowed-hackers-get/?noredirect=on.

[7] India Ashok. 2016. Hackers leave Finnish residents cold after DDoS attack knocks out heating systems. https://www.ibtimes.co.uk/hackers-leave-finnish-residents-cold-after-ddos-attack-knocks-out-heating-systems-1590639.

[8] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. SoundNet: Learning Sound Representations from Unlabeled Video. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 892–900.

[9] Bharathan Balaji, Jason Koh, Nadir Weibel, and Yuvraj Agarwal. 2016. Genie: A Longitudinal Study Comparing Physical and Software Thermostats in Office Buildings. In *Proc. of the 2016 ACM Internat. Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) *(UbiComp '16)*. ACM, New York, NY, USA, 1200–1211. https://doi.org/10.1145/2971648.2971719

[10] Alex Beltran, Varick L. Erickson, and Alberto E. Cerpa. 2013. ThermoSense: Occupancy Thermal Based Sensing for HVAC Control. In *Proc. of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings* (Roma, Italy) *(BuildSys'13)*. ACM, New York, NY, USA, 1–8. https://doi.org/10.1145/2528282.2528301

[11] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding? arXiv:2102.05095 [cs.CV]

[12] Sejal Bhalla, Mayank Goel, and Rushil Khurana. 2021. IMU2Doppler: Cross-Modal Domain Adaptation for Doppler-based Activity Recognition Using IMU Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–20.

[13] Sudershan Boovaraghavan, Chen Chen, Anurag Maravi, Mike Czapik, Yang Zhang, Chris Harrison, and Yuvraj Agarwal. 2023. Mites: Design and Deployment of a General-Purpose Sensing Infrastructure for Buildings. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 1, Article 2 (mar 2023), 32 pages. https://doi.org/10.1145/3580865

[14] Bosch. 2022. Cross Domain Development Kit | XDK. https://www.bosch-connectivity.com/media/downloads/xdk/xdk_node_110_combined_datasheet.pdf.

[15] Hong Cai, Belal Korany, Chitra R Karanam, and Yasamin Mostofi. 2020. Teaching rf to sense without rf training measurements. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–22.

[16] Kelly E. Caine, Arthur D. Fisk, and Wendy A. Rogers. 2006. Benefits and Privacy Concerns of a Home Equipped with a Visual Sensing System: A Perspective from Older Adults. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 2 (2006), 180–184. https://doi.org/10.1177/154193120605000203 arXiv:https://doi.org/10.1177/154193120605000203

[17] Timothy I Cannings, Yingying Fan, and Richard J Samworth. 2020. Classification with imperfect training labels. *Biometrika* 107, 2 (2020), 311–330.

[18] Song Cao and Ram Nevatia. 2016. Exploring deep learning based solutions in fine grained activity recognition in the wild. In *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, Cancun, Mexico, 384–389. https://doi.org/10.1109/ICPR.2016.7899664

[19] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, Honolulu, HI, USA, 4724–4733. https://doi.org/10.1109/CVPR.2017.502

[20] Youngjae Chang, Akhil Mathur, Anton Isopoussu, Junehwa Song, and Fahim Kawsar. 2020. A systematic study of unsupervised domain adaptation for robust human-activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–30.

[21] Youngjae Chang, Akhil Mathur, Anton Isopoussu, Junehwa Song, and Fahim Kawsar. 2020. A systematic study of unsupervised domain adaptation for robust human-activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*

4, 1 (2020), 1–30.

[22] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[23] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. arXiv:1906.07155 [cs.CV]

[24] Qingchao Chen, Bo Tan, Kevin Chetty, and Karl Woodbridge. 2016. Activity recognition based on micro-Doppler signature with in-home Wi-Fi. In *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, Munich, Germany, 1–6. https://doi.org/10.1109/HealthCom.2016.7749457

[25] Wenqiang Chen, Shupei Lin, Elizabeth Thompson, and John Stankovic. 2021. Sensecollect: We need efficient ways to collect on-body sensor-based human activity data! *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–27.

[26] Shohreh Deldari, Hao Xue, Aaqib Saeed, Jiayuan He, Daniel V. Smith, and Flora D. Salim. 2022. Beyond Just Vision: A Review on Self-Supervised Representation Learning on Multimodal and Temporal Data. arXiv:2206.02353 [cs.LG]

[27] Shohreh Deldari, Hao Xue, Aaqib Saeed, Daniel V Smith, and Flora D Salim. 2022. COCOA: Cross Modality Contrastive Learning for Sensor Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–28.

[28] Florenc Demrozi, Graziano Pravadelli, Azra Bihorac, and Parisa Rashidi. 2020. Human Activity Recognition Using Inertial, Physiological and Environmental Sensors: A Comprehensive Survey. *IEEE Access* 8 (2020), 210816–210836. https://doi.org/10.1109/ACCESS.2020.3037715

[29] Konstantinos Drossos, Stylianos I. Mimilakis, Shayan Gharib, Yanxiong Li, and Tuomas Virtanen. 2020. Sound Event Detection with Depthwise Separable and Dilated Convolutions. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IJCNN, Glasgow, UK, 1–7. https://doi.org/10.1109/IJCNN48605.2020.9207532

[30] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin. 2022. PYSKL: Towards Good Practices for Skeleton Action Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 7351–7354. https://doi.org/10.1145/3503161.3548546

[31] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. 2022. Revisiting Skeleton-based Action Recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New Orleans, LA, USA, 2959–2968. https://doi.org/10.1109/CVPR52688.2022.00298

[32] Pardis Emami-Naeini, Janarth Dheenadhayalan, Yuvraj Agarwal, and Lorrie Faith Cranor. 2021. Which Privacy and Security Attributes Most Impact Consumers' Risk Perception and Willingness to Purchase IoT Devices?. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, San Francisco, CA, USA, 519–536. https://doi.org/10.1109/SP40001.2021.00112

[33] Pardis Emami-Naeini, Henry Dixon, Yuvraj Agarwal, and Lorrie Faith Cranor. 2019. Exploring How Privacy and Security Factor into IoT Device Purchase Behavior. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300764

[34] Baris Erol, Sevgi Z. Gurbuz, and Moeness G. Amin. 2019. GAN-based Synthetic Radar Micro-Doppler Augmentations for Improved Human Activity Recognition. In *2019 IEEE Radar Conference (RadarConf)*. IEEE, Boston, MA, USA, 1–5. https://doi.org/10.1109/RADAR.2019.8835589

[35] Christoph Feichtenhofer. 2020. X3D: Expanding Architectures for Efficient Video Recognition. arXiv:2004.04730 [cs.CV]

[36] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. SlowFast Networks for Video Recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Seoul, Korea (South), 6201–6210. https://doi.org/10.1109/ICCV.2019.00630

[37] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, New Orleans, LA, USA, 776–780. https://doi.org/10.1109/ICASSP.2017.7952261

[38] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. 2019. Large-Scale Weakly-Supervised Pre-Training for Video Action Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA, 12038–12047. https://doi.org/10.1109/CVPR.2019.01232

[39] Emily Green. 2018. Hacker terrorizes family by hijacking baby monitor. https://nordvpn.com/blog/baby-monitor-iot-hacking/.

[40] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. 2018. AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, UT, USA, 6047–6056. https://doi.org/10.1109/CVPR.2018.00633

[41] Harish Haresamudram, Irfan Essa, and Thomas Plötz. 2023. Investigating Enhancements to Contrastive Predictive Coding for Human Activity Recognition. In *2023 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, Atlanta, GA, USA, 232–241. https://doi.org/10.1109/PERCOM56429.2023.10099197

[42] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Boston, MA, USA, 961–970. https://doi.org/10.1109/CVPR.2015.7298698

[43] Zawar Hussain, Quan Z. Sheng, and Wei Emma Zhang. 2020. A review and categorization of techniques on device-free human activity recognition. *Journal of Network and Computer Applications* 167 (oct 2020), 102738. https://doi.org/10.1016/j.jnca.2020.102738

[44] Texas Instruments. 2017. Awr1642 single-chip 77-and 79-ghz fmcw radar sensor. , 60 pages.

[45] Texas Instruments. 2018. Dca1000evm data capture card. *Retrieved May* 17 (2018), 2022.

[46] S. Iwasawa, K. Ebihara, J. Ohya, and S. Morishima. 1998. Real-time human posture estimation using monocular thermal images. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, Nara, Japan, 492–497. https://doi.org/10.1109/AFGR.1998.670996

[47] Yash Jain, Chi Ian Tang, Chulhong Min, Fahim Kawsar, and Akhil Mathur. 2022. ColloSSL: Collaborative Self-Supervised Learning for Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1, Article 17 (mar 2022), 28 pages. https://doi.org/10.1145/3517246

[48] Haojian Jin, Boyuan Guo, Rituparna Roychoudhury, Yaxing Yao, Swarun Kumar, Yuvraj Agarwal, and Jason I. Hong. 2022. Exploring the Needs of Users for Supporting Privacy-Protective Behaviors in Smart Homes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 449, 19 pages. https://doi.org/10.1145/3491102.3517602

[49] Charmi Jobanputra, Jatna Bavishi, and Nishant Doshi. 2019. Human activity recognition: A survey. *Procedia Computer Science* 155 (2019), 698–703.

[50] G. R. Kanagachidambaresan. 2021. *Sensors and SBCs for Smart City Infrastructure*. Springer International Publishing, Cham, 47–75. https://doi.org/10.1007/978-3-030-72957-8_3

[51] Shian-Ru Ke, Hoang Le Uyen Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, and Kyoung-Ho Choi. 2013. A review on video-based human activity recognition. *Computers* 2, 2 (2013), 88–131.

[52] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. 2011. HMDB: A large video database for human motion recognition. In *2011 International Conference on Computer Vision*. IEEE, Barcelona, Spain, 2556–2563. https://doi.org/10.1109/ICCV.2011.6126543

[53] Hyeokhyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D Abowd, Nicholas D Lane, and Thomas Ploetz. 2020. IMUTube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–29.

[54] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-Play Acoustic Activity Recognition. In *Proc. of the 31st Annual ACM Symposium on UIST* (Berlin, Germany) *(UIST '18)*. ACM, New York, NY, USA, 213–224. https://doi.org/10.1145/3242587.3242609

[55] Gierad Laput and Chris Harrison. 2019. SurfaceSight: A New Spin on Touch, User, and Object Sensing for IoT Experiences. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300559

[56] Gierad Laput, Yang Zhang, and Chris Harrison. 2017. Synthetic Sensors: Towards General-Purpose Sensing. In *Proc. of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, New York, NY, USA, 3986–3999. https://doi.org/10.1145/3025453.3025773

[57] Oscar D Lara and Miguel A Labrador. 2012. A survey on human activity recognition using wearable sensors. *IEEE communications surveys & tutorials* 15, 3 (2012), 1192–1209.

[58] Heju Li, Xin He, Xukai Chen, Yinyin Fang, and Qun Fang. 2019. Wi-motion: A robust human activity recognition using WiFi signals. *IEEE Access* 7 (2019), 153287–153299.

[59] Xinyu Li, Yuan He, and Xiaojun Jing. 2019. A survey of deep learning-based human activity recognition in radar. *Remote Sensing* 11, 9 (2019), 1068.

[60] Dawei Liang, Guihong Li, Rebecca Adaimi, Radu Marculescu, and Edison Thomaz. 2022. AudioIMU: Enhancing Inertial Sensing-Based Activity Recognition with Acoustic Models. In *Proceedings of the 2022 ACM International Symposium on Wearable Computers* (Cambridge, United Kingdom) *(ISWC '22)*. Association for Computing Machinery, New York, NY, USA, 44–48. https://doi.org/10.1145/3544794.3558471

[61] Ji Lin, Chuang Gan, and Song Han. 2019. TSM: Temporal Shift Module for Efficient Video Understanding. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Seoul, Korea (South), 7082–7092. https://doi.org/10.1109/ICCV.2019.00718

[62] Guocheng Liu, Caixia Zhang, Qingyang Xu, Ruoshi Cheng, Yong Song, Xianfeng Yuan, and Jie Sun. 2020. I3d-shufflenet based human action Recognition. *Algorithms* 13, 11 (2020), 301.

[63] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. 2020. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 10 (2020), 2684–2701.

[64] Sicong Liu, Junzhao Du, Anshumali Shrivastava, and Lin Zhong. 2019. Privacy Adversarial Network. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (dec 2019), 1–18. https://doi.org/10.1145/3369816

[65] Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, and Tong Lu. 2021. TAM: Temporal Adaptive Module for Video Recognition. arXiv:2005.06803 [cs.CV]

[66] Ginés Hidalgo Martınez. 2019. *OpenPose: Whole-body pose estimation.* Ph. D. Dissertation. Master's Thesis, Carnegie Mellon University.

[67] Shinya Misaki, Keisuke Umakoshi, Tomokazu Matsui, Hyuckjin Choi, Manato Fujimoto, and Keiichi Yasumoto. 2021. Non-Contact In-Home Activity Recognition System Utilizing Doppler Sensors. In *Adjunct Proceedings of the 2021 International Conference on Distributed Computing and Networking* (Nara, Japan) *(ICDCN '21)*. Association for Computing Machinery, New York, NY, USA, 169–174. https://doi.org/10.1145/3427477.3429463

[68] Mites.io. 2020. Mites.io: a full-stack ubiquitous sensing platform. https://mites.io/.

[69] MMAction2. 2020. OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark. https://github.com/open-mmlab/mmaction2.

[70] MMPose. 2020. OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose.

[71] Vimal Mollyn, Karan Ahuja, Dhruv Verma, Chris Harrison, and Mayank Goel. 2022. SAMoSA: Sensing Activities with Motion and Subsampled Audio. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–19.

[72] Muhammad Muaaz, Ali Chelli, Ahmed Abdelmonem Abdelgawwad, Andreu Català Mallofré, and Matthias Pätzold. 2020. WiWeHAR: Multimodal human activity recognition using Wi-Fi and wearable sensing modalities. *IEEE access* 8 (2020), 164453–164470.

[73] Sebastian Münzner, Philip Schmidt, Attila Reiss, Michael Hanselmann, Rainer Stiefelhagen, and Robert Dürichen. 2017. CNN-Based Sensor Fusion Techniques for Multimodal Human Activity Recognition. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers* (Maui, Hawaii) *(ISWC '17)*. Association for Computing Machinery, New York, NY, USA, 158–165. https://doi.org/10.1145/3123021.3123046

[74] Curtis Northcutt, Lu Jiang, and Isaac Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research* 70 (2021), 1373–1411.

[75] Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* 16, 1 (2016). https://doi.org/10.3390/s16010115

[76] Shijia Pan, Mario Berges, Juleen Rodakowski, Pei Zhang, and Hae Young Noh. 2019. Fine-Grained Recognition of Activities of Daily Living through Structural Vibration and Electrical Sensing. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation* (New York, NY, USA) *(BuildSys '19)*. Association for Computing Machinery, New York, NY, USA, 149–158. https://doi.org/10.1145/3360322.3360851

[77] Preksha Pareek and Ankit Thakkar. 2021. A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review* 54, 3 (2021), 2259–2322.

[78] Liangying Peng, Ling Chen, Zhenan Ye, and Yi Zhang. 2018. AROMA: A Deep Multi-Task Learning Based Simple and Complex Human Activity Recognition Method Using Wearable Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 2, Article 74 (jul 2018), 16 pages. https://doi.org/10.1145/3214277

[79] Joseph Phelps, Glen Nowak, and Elizabeth Ferrell. 2000. Privacy Concerns and Consumer Willingness to Provide Personal Information. *Journal of Public Policy & Marketing* 19, 1 (2000), 27–41. http://www.jstor.org/stable/30000485

[80] Prasoon Patidar, Mayank Goel, Yuvraj Agarwal. 2023. VAX: Open-source repository for the VAX system. https://github.com/synergylabs/vax.

[81] Riccardo Presotto, Gabriele Civitarese, and Claudio Bettini. 2022. Federated Clustering and Semi-Supervised learning: A new partnership for personalized Human Activity Recognition. *Pervasive and Mobile Computing* 88 (2022), 101726.

[82] Valentin Radu and Maximilian Henne. 2019. Vision2sensor: Knowledge transfer across sensing modalities for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–21.

[83] Bhiksha Raj, Kaustubh Kalgaonkar, Chris Harrison, and Paul Dietz. 2012. Ultrasonic Doppler Sensing in HCI. *IEEE Pervasive Computing* 11, 2 (2012), 24–29. https://doi.org/10.1109/MPRV.2012.17

[84] Sreenivasan Ramasamy Ramamurthy and Nirmalya Roy. 2018. Recent trends in machine learning for human activity recognition—A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, 4 (2018), e1254.

[85] Suneth Ranasinghe, Fadi Al Machot, and Heinrich C Mayr. 2016. A review on applications of activity recognition systems with regard to performance and evaluation. *International Journal of Distributed Sensor Networks* 12, 8 (2016), 1550147716665520. https://doi.org/10.1177/1550147716665520 arXiv:https://doi.org/10.1177/1550147716665520

[86] Lipsarani Sahoo, Nazmus Sakib Miazi, Mohamed Shehab, Florian Alt, and Yomna Abdelrahman. 2022. You Know Too Much: Investigating Users' Perceptions and Privacy Concerns Towards Thermal Imaging. In *Privacy Symposium 2022*, Stefan Schiffner, Sebastien Ziegler, and Adrian Quesada Rodriguez (Eds.). Springer International Publishing, Cham, 207–229.

[87] Alex Schiffer. 2017. How a fish tank helped hack a casino. https://www.washingtonpost.com/news/innovations/wp/2017/07/21/how-a-fish-tank-helped-hack-a-casino/?noredirect=on.

[88] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. arXiv:1604.02808 [cs.CV]

[89] Hao Shao, Shengju Qian, and Yu Liu. 2020. Temporal Interlacing Network. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 07 (Apr. 2020), 11966–11973. https://doi.org/10.1609/aaai.v34i07.6872

[90] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA, 12018–12027. https://doi.org/10.1109/CVPR.2019.01230

[91] Akash Deep Singh, Sandeep Singh Sandha, Luis Garcia, and Mani Srivastava. 2019. RadHAR: Human Activity Recognition from Point Clouds Generated through a Millimeter-Wave Radar. In *Proceedings of the 3rd ACM Workshop on Millimeter-Wave Networks and Sensing Systems* (Los Cabos, Mexico) *(mmNets'19)*. Association for Computing Machinery, New York, NY, USA, 51–56. https://doi.org/10.1145/3349624.3356768

[92] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. arXiv:1212.0402 [cs.CV]

[93] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. 2018. Actor-Centric Relation Network. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 335–351.

[94] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA, 5686–5696. https://doi.org/10.1109/CVPR.2019.00584

[95] Vishnu Priya Thotakura and Purnachand Nalluri. 2022. Convolutional 3D in Activity Recognition -A Review. In *2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP)*. IEEE, Vijayawada, India, 1–6. https://doi.org/10.1109/AISP53593.2022.9760638

[96] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, UT, USA, 6450–6459. https://doi.org/10.1109/CVPR.2018.00675

[97] Kimberly T. Tran, Lewis D. Griffin, Kevin Chetty, and Shelly Vishwakarma. 2020. Transfer Learning from Audio Deep Learning Models for Micro-Doppler Activity Recognition. In *2020 IEEE International Radar Conference (RADAR)*. IEEE, Washington, DC, USA, 584–589. https://doi.org/10.1109/RADAR42522.2020.9114643

[98] Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiaris. 2015. A review of human activity recognition methods. *Frontiers in Robotics and AI* 2 (2015), 28.

[99] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 20–36.

[100] Pete Warden, Matthew Stewart, Brian Plancher, Colby Banbury, Shvetank Prakash, Emma Chen, Zain Asgar, Sachin Katti, and Vijay Janapa Reddi. 2022. Machine Learning Sensors. https://doi.org/10.48550/ARXIV.2206.03266

[101] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. 2019. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Long Beach, CA, USA, 284–293.

[102] Tong Wu, Murtadha Aldeer, Tahiya Chowdhury, Amber Haynes, Fateme Nikseresht, Mahsa Pahlavikhah Varnosfaderani, Jiechao Gao, Arsalan Heydarian, Brad Campbell, and Jorge Ortiz. 2021. The Smart Building Privacy Challenge. In *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation* (Coimbra, Portugal) *(BuildSys '21)*. Association for Computing Machinery, New York, NY, USA, 238–239. https://doi.org/10.1145/3486611.3492234

[103] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18)*. AAAI Press, New Orleans, Louisiana, USA, Article 912, 9 pages.

[104] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. 2020. Temporal Pyramid Network for Action Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 588–597. https://doi.org/10.1109/CVPR42600.2020.00067

[105] Deju Yang, Liangli Ma, and Fei Liao. 2019. An Intelligent Voice Interaction System Based on Raspberry Pi. In *2019 11th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Vol. 1. IEEE, Hangzhou, China, 237–240. https://doi.org/10.1109/IHMSC.2019.00062

[106] Yang Yang, Chunping Hou, Yue Lang, Dai Guan, Danyang Huang, and Jinchen Xu. 2019. Open-set human activity recognition based on micro-Doppler signatures. *Pattern Recognition* 85 (2019), 60–69.

[107] Zhaoyuan Yang, Yang Zhao, and Weizhong Yan. 2020. Adversarial Vulnerability in Doppler-based Human Activity Recognition. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Glasgow, UK, 1–7. https://doi.org/10.1109/IJCNN48605.2020.9207686

[108] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. 2018. Temporal Relational Reasoning in Videos. arXiv:1711.08496 [cs.CV]